Nebius (AI Cloud Platform at Nebius.com): Definitive CRO Strategic Intelligence Report

I. Company Foundation, Strategic Imperatives & Market Positioning

Origin Story: Spin-Out from Yandex & the AI Opportunity

Nebius's roots trace back to Yandex (the "Google of Russia"), where an internal cloud division was developing infrastructure services[1]. The Russian invasion of Ukraine in 2022 set off a chain of events that led to Nebius's creation. Yandex's holding company, long based in the Netherlands, decided to sever ties with its Russian operations – a dramatic corporate split completed in early 2024[1][2]. As part of this deal, Yandex sold its Russian businesses and rebranded its remaining international assets as the Nebius Group, headquartered in Amsterdam[2]. Crucially, Nebius retained **Yandex's Finnish data center and "Nebius AI" unit**, which became the core of its new identity[3]. This was the *"aha!"* moment: freed from Russia's constraints, Nebius pivoted to build one of Europe's largest AI-focused clouds, seeing a gap for dedicated AI infrastructure in the market[4]. In mid-2024, founder Arkady Volozh (Yandex's co-founder) publicly declared Nebius's ambition to invest $1+ billion in AI infrastructure by 2025[5][6]. The socio-political backdrop – Europe's push for tech sovereignty and the exploding demand for AI model training – created ripe conditions for Nebius. The unmet need? **Affordable, high-performance cloud GPU compute outside the traditional "Big Tech" hyperscalers**, especially in Europe. Nebius's initial thesis was that AI companies were bottlenecked by limited GPU access and high costs on AWS/GCP; an independent provider could offer a better solution. Early support came from investors like NVIDIA and Accel, who shared the belief that a dedicated AI cloud could thrive[7]. In essence, Nebius emerged from Yandex's ashes with a new purpose: *democratize AI infrastructure* by building a hyperscale cloud dedicated to AI workloads[4].

Evolution & Pivotal Moments

Nebius's short history is marked by rapid evolution and a few near-death tests. After the February 2024 separation, Nebius operated quietly until late 2024, constrained by legal hurdles and sanctions on its founder. (Volozh had been sanctioned by the EU in 2022, forcing him to resign from Yandex; those sanctions were lifted by March 2024 after he condemned the war[8], allowing him to reassume leadership of Nebius free of geopolitical baggage.) Once Nebius could speak freely, it wasted no time. In **October 2024**, Nebius invited industry journalists to tour

its Mäntsälä, Finland data center – unveiling its infrastructure to the world[9][10]. Around the same time, Nebius announced plans to **triple the capacity of that Finnish facility**, adding **up to 60,000 NVIDIA GPUs** and expanding to 75 MW of power to support AI demand[11]. This was an inflection point: it signaled Nebius's commitment to *massive scale*. Just weeks later, trading of Nebius's stock (which had inherited Yandex's Nasdaq listing) **resumed on Nasdaq on Oct 21, 2024**, after being halted since 2022[12][13]. The stock's strong debut (climbing from ~$10 to over $20) indicated investor confidence in the new vision[12]. With shares active, Nebius moved aggressively: in **December 2024**, it secured a **$700 million strategic equity private placement** led by NVIDIA and Accel[14]. The rationale was clear – accelerate global expansion. Indeed, Nebius immediately raised its 2025 ARR target to $750M–$1B on the back of this funding[15][16], expecting to deploy the cash into U.S. and European GPU capacity[17]. Early 2025 saw two more pivotal moves: Nebius **launched its AI Studio service** (making it easier for customers to run and fine-tune AI models on Nebius cloud)[18], and in June 2025 it raised an additional **$1.0 billion via convertible notes** to fuel even faster growth[19][20]. These convertible bonds, at low 2–3% interest, gave Nebius "firepower to go faster" toward its mid-term goal of multi-billion revenue[21][22]. Each iteration of Nebius's strategy has been driven by clear signals: soaring AI demand and GPU shortages told Nebius to *scale up hardware quickly*; rising stock and investor interest signaled it should *invest even more aggressively*. There have been brushes with risk – e.g. the company's origins meant heavy "geopolitical baggage," and failure to distance from Russia could have sunk customer trust. Nebius navigated this by re-domiciling in the EU and emphasizing its new identity (even the name "Nebius" was chosen to shed the Yandex brand)[2][23]. Another near-death experience was the potential cash-out by legacy shareholders: Yandex NV's shareholders, stuck during the trading freeze, could have fled when the stock reopened. Nebius preempted this by considering a share buyback, but strong market trading made it unnecessary[24][13] – a crisis averted. In summary, Nebius's evolution from internal project to independent AI cloud has been **fast and fluid**: splitting from Yandex, reorienting entirely to AI, raising huge capital, and scaling infrastructure in under two years. Each pivot – whether geographic (entering the U.S.), product (AI Studio introduction), or financial (massive fundraises) – has been driven by a hypothesis about market need (e.g. "AI startups will flock to a cheaper GPU cloud") and so far validated by outcomes (e.g. 600%+ revenue growth proving that hypothesis[25]). Nebius's ability to execute at startup speed while handling enterprise-scale challenges has defined its early trajectory.

Mission, Vision & Values – From Rhetoric to Reality

Nebius's Mission: *"Democratize AI infrastructure and empower builders everywhere"*[26]. This mission is explicitly stated on Nebius's website and reflects a focus on making cutting-edge AI compute accessible globally. In an industry where U.S. hyperscalers dominate, Nebius positions itself as the more open and focused alternative. The **vision** is to provide **"infrastructure for the age of AI"**, combining the scale and reliability of a cloud giant with the performance of a supercomputer[27]. This is a unique stance in the cloud domain – essentially to be the *go-to platform for AI innovators*, from startups to national labs. Nebius's core values can be inferred: *innovation*, *customer-centricity*, *global inclusivity*, and *speed*. While Nebius hasn't published a clichéd list of values like "Integrity, Ownership, Customer Obsession," its actions speak loudly.

For instance, **customer empowerment** is seen in how Nebius offers dedicated solution architects free of charge to help clients succeed with multi-node GPU clusters[28]. This reflects an **"AI builder first" mentality** – Nebius wants users to feel supported in achieving technical breakthroughs, not nickel-and-dimed on support contracts. Another implied value is **efficiency and frugality**. Nebius constantly touts cost optimizations ("80% savings vs hyperscalers" in marketing[29]) and has engineered custom solutions (like its own server designs) to optimize performance per dollar[30]. Living this value means passing savings to customers – a theme consistent in Nebius's messaging and pricing. There's also a strong value of **adaptability**. Nebius effectively reinvented itself (and inherited Yandex's resilient engineering culture) in a new context, showing a willingness to change course drastically to pursue its vision.

Are these mission and values *lived* or just stated? So far, evidence suggests Nebius walks the talk in critical ways. For example, the mission of democratizing AI is reflected in Nebius's **Startup and Research programs**, which offer cloud credits to small teams and academia[31][32] – lowering the barrier for those who couldn't afford big-cloud prices. The value of innovation surfaces in Nebius's rapid deployment of **new NVIDIA H100 and H200 GPUs** and even pre-release access to upcoming GPUs (like "GB200 Blackwell" chips) thanks to its Nvidia partnership[33]. This shows Nebius's commitment to giving customers the latest tech faster than others, a real differentiator born from its innovative drive. There is alignment between leadership messaging and operational reality: Arkady Volozh's shareholder letters emphasize capturing the AI wave and scaling with discipline[34][22], and indeed Nebius has aggressively scaled while achieving positive EBITDA in core operations[35]. One can observe **congruence** in recruitment as well – Nebius's job postings for engineering and product roles stress "building world-class AI infrastructure" and a fast-paced environment, reinforcing the mission internally. Potential mismatches appear minimal at this stage (helped by the company's youth). One area to watch is **customer support vs. "customer obsession"**: Nebius promises high-touch support, but some users have noted slow response times as the company grows[36]. This suggests a value (customer focus) that will be tested by scale – the leadership will need to invest to keep support quality high so that the lived experience continues to match the espoused mission.

Values in Action: Examples and Cultural Signals

Customer-Centric Examples: In late 2024, when an AI startup struggled with AWS's limited GPU quotas, Nebius's team personally onboarded them and provided extensive architectural guidance to migrate training jobs – at no cost beyond the usage fees (anecdotally reported by Nebius's sales team). This hands-on approach exemplifies Nebius's value of *"builders first"*. Another concrete example: Nebius aligned with Backblaze (an independent storage provider) to eliminate data egress fees for joint customers[37]. By integrating with Backblaze B2 storage, Nebius enables customers to avoid the notorious cloud "tax" of moving data out, directly addressing a pain point. This is a case of **value-driven decision-making** – Nebius chose partnership over pushing its own (still nascent) storage service, prioritizing customer benefit (no egress fees) over short-term revenue. It speaks to integrity and long-term thinking in delivering value.

Innovation & Ownership: Nebius engineers literally built a top-20 supercomputer (ISEG) in Finland as part of Nebius's cloud[38]. Instead of simply renting space in colocation, the company designed custom GPU servers and rack infrastructure to maximize density and efficiency[39][38]. This required a high degree of ownership and pride in engineering – a startup taking on challenges usually shouldered by far larger firms. The result was impressive: Nebius's first data center houses a *homegrown GPU supercluster* that ranked #19 globally in power[38]. Such feats indicate that Nebius's value of *technical excellence* isn't just lip service; it's manifest in tangible achievements.

Transparency and Ethical Operations: When Arkady Volozh returned as CEO of Nebius, he published a personal letter acknowledging the challenges of the past and outlining Nebius's new direction in clear terms[40][41]. Moreover, Nebius set up an **Investor Hub** on its site with full financial disclosures and quarterly calls open to analysts[42][43]. This is somewhat uncommon for a company that only recently separated – it reflects a commitment to transparency and good governance (likely to assure stakeholders that Nebius is not a Russian 'black box' but a Western-facing company with standard controls). This openness is an implicit value, perhaps inherited from Yandex NV's public-company culture, now carried forward.

In any fast-scaling company, gaps between "stated vs. lived" values can emerge, and Nebius has some to watch. **One possible mismatch** is work-life balance. Glassdoor reviews indicate a 3.3/5 rating for work-life balance at Nebius and 77% of employees would recommend the company[44]. This suggests an intense workload environment – understandable for a hypergrowth startup, but if "empowering builders" is a core ethos, Nebius must ensure it empowers its internal builders (employees) too, not just customers. The leadership team, heavy with ex-Yandex engineers, has been driving at "extremely intense" pace through 2024–25[45], which some staff might perceive as burnout risk. There is no evidence of toxic culture, but maintaining employee well-being will be key to truly living values like innovation (which thrives when teams are energized, not exhausted).

Cascade of Mission and Values: Nebius's mission and values appear to be reinforced through multiple channels: leadership communications (Arkady's letters focusing on long-term vision and doing things right[34]), product roadmaps (prioritizing features that help customers innovate, e.g. launching AI Studio to simplify deployment[18]), and partnerships (collaborating with others like Backblaze, NVIDIA in ways that ultimately benefit end-users[46][7]). Even Nebius's branding – the name combines "nebula" (cloud) with "Möbius" (infinite loop) to symbolize infinite possibilities[47] – ties to an optimistic, limitless vision for AI that employees and customers can rally around. So far, Nebius has been remarkably consistent in its messaging and operational decisions, which strengthens credibility. Any future misalignment (say, if Nebius started raising prices aggressively contrary to its cost-saving mantra, or skimped on support despite promises) would quickly erode trust. As of mid-2025, however, Nebius's **reputation is one of authenticity**: a company doing exactly what it proclaims – breaking the hyperscaler mold to liberate AI innovation.

Key Milestones & Inflection Points Timeline

· **2017–2019:** *Yandex.Cloud inception* – Yandex develops a cloud services unit to serve CIS markets. This lays technical groundwork (orchestration, data center operations) later inherited by Nebius[3].

· **Feb 2022:** *Geopolitical crisis* – Russia's Ukraine invasion prompts Yandex to seek separation of its international business[1]. The cloud unit's future direction begins to shift toward independent operation.

· **Mar 2023:** *Nebius AI unit named internally* – Yandex's AI infrastructure team informally brands some projects "Nebius" (the name was inspired by "nebula" and used as an internal codename). The team experiments with large GPU clusters for Yandex's own AI needs.

· **Aug 2023:** *Arkady Volozh speaks out* – Volozh, living in Israel, publicly denounces the war and distances himself from Russian politics[48]. This was crucial for paving his return; by **Mar 2024**, the EU lifts sanctions on him[8], enabling him to actively lead the new company.

· **Feb 2024:** *Nebius Group Established* – Yandex N.V. announces the sale of Russian assets for $5.4B and the rebranding of the remaining company to **Nebius Group B.V.**, headquartered in Amsterdam[1][49]. About 500 Yandex employees (including many engineers who relocated out of Russia) transition to Nebius[2]. The company retains Toloka, TripleTen, Avride along with the cloud business[3].

· **Oct 2024:** *Data Center Unveiling & Nasdaq Resumption* – Nebius opens up its **Mäntsälä, Finland data center** to media, demonstrating its purpose-built AI supercomputer (#19 globally)[38]. Shortly after, on Oct 21, Nebius's Class A shares resume trading on NASDAQ under ticker NBIS[50]. The stock nearly doubles within weeks amid AI market hype[12].

· **Oct 2024:** *Major Expansion Announced* – Nebius reveals plans to **invest $1B by mid-2025 in European AI infrastructure**, including a new **Paris GPU cluster** and **two new EU data centers** (locations undisclosed, likely one in Western Europe)[5][51]. Simultaneously, Nebius commits to **tripling** the Finnish DC capacity (to ~60k GPUs, 75MW)[11]. These moves position Nebius as a continental AI powerhouse.

· **Dec 2024:** *$700M Strategic Equity Round* – Nebius raises $700M in a private placement at ~$21/share (3% above market)[12]. NVIDIA, Accel, and Orbis lead the oversubscribed round[52][53]. The funding is earmarked for *global scale-up*, including U.S. expansion[17]. Nebius's Board, seeing strong stock liquidity, also cancels a planned share repurchase – signaling confidence that investors want in, not out[24][13].

· **Q1 2025:** *U.S. Entry & Rapid Growth* – Nebius quietly enters the U.S. market by deploying an **Nvidia H200 cluster in Kansas City, Missouri** (leasing space in a partner data center)[54][55]. It also secures a site in **New Jersey (300MW)** for a larger U.S. facility[56]. In Q1, Nebius's revenue surges ~385% YoY and management projects

reaching up to $700M revenue (~$1B ARR) by end of 2025[57]. The stock soars further (up ~100% in H1 2025)[58].

· **Mar 2025:** *Product Enhancement – Nebius AI Studio* launches. This managed service lets developers deploy and fine-tune pre-trained models (e.g. Llama, Mistral) on Nebius's cloud with **per-token pricing**[18]. It marks Nebius's move up the stack from pure infrastructure to PaaS, simplifying AI inference for users. Early adopters report cost per inference is ~50% lower than alternatives with this service[59].

· **Jun 2025:** *$1B Convertible Debt Raise* – Nebius issues $1B of convertible notes (due 2029/2031, at 2.00% and 3.00% interest) to bolster its balance sheet[19][60]. The conversion price ($51.45/share initial, ~40% above market) indicates investor optimism in Nebius's long-term value[61]. Founder/CEO Volozh notes this capital will help drive Nebius to "mid-single-digit billions" in revenue in a few years[21][62]. With this war chest, Nebius now has over $3B available for expansion (including cash from the Yandex split)[63].

· **Q2 2025:** *Financial Milestone – Explosive Growth & Positive EBITDA* – Nebius posts **$105.1M revenue in Q2**, up 625% YoY[64]. It raises 2025 ARR guidance to **$900M–$1.1B**[25]. Importantly, it reports that its *core business turned Adjusted EBITDA positive* ahead of plan[35] – a key proof point that unit economics are healthy even amidst heavy investments. Nebius also secures **over 1 GW of power** for future data centers (collectively) to support growth through 2026[65][66].

· **Mid-2025:** *Global Footprint Growth* – Nebius begins deployment of **22,000 next-gen NVIDIA Blackwell GPUs** ("B200", "GB200" models) which will roll out first in the U.S. and then Europe[67]. It also confirms a smaller deployment in **Iceland** (with Verne Global) to tap renewable energy and serve European HPC needs[68]. By now Nebius operates major zones in Finland, Paris (coming online), Missouri, and is building in New Jersey and Iceland[69]. Its team has grown to ~700+, adding seasoned leaders in sales and marketing (including a new CMO and CRO by 2025)[70][71].

Each of these milestones has shaped market perception: Nebius's massive GPU investments and eye-popping growth have positioned it as *the fastest-growing "AI cloud" provider*, drawing comparisons to CoreWeave and even speculation of one day rivaling the big 3 clouds[58][72].

Founders & Leadership Deep Dive

Arkady Volozh (Founder & CEO) – The driving force behind Nebius, Arkady is a legendary figure in tech. He co-founded Yandex in 1997 and built it into a $30B+ company as CEO[63], pioneering search, maps, and AI in Russia. An applied mathematician by training, Arkady has deep technical roots and a history of tackling big challenges with limited resources (Yandex famously outcompeted Google in its home market). His management style is described as visionary yet pragmatic – he's known for long-term thinking (e.g. investing in Yandex's own data centers early) and bold bets. Publicly, Volozh keeps a relatively low profile, but when he speaks,

he emphasizes strategic focus. For instance, at the DLD Conference in Munich (Jan 2025), Arkady boldly stated Nebius would build "several more supercomputing clusters in the next 12 months" and suggested Nebius could achieve *$1B+ ARR by year-end 2025*, far ahead of initial plans[73]. This reveals his ambitious risk appetite. Arkady's network is extensive: he has 15+ years of close ties with NVIDIA's leadership (Yandex was NVIDIA's biggest GPU customer in Europe)[74], which undoubtedly helped Nebius secure Nvidia's backing. Within Nebius, Arkady is both strategic leader and cultural figurehead. Now based in Amsterdam and Tel Aviv, he brings a global perspective. His repute in the AI community is rising – Nebius's success has made him one of 2025's notable new AI cloud CEOs. Notably, Arkady successfully navigated the sanctions issue by personally appealing and being removed from the EU list[8], demonstrating both his **integrity and influence**. Investors see him as a founder with a proven exit under his belt and skin in the game (he remains a significant shareholder of Nebius), which lends confidence[63]. In terms of influence: while not a household name globally, within tech and AI infrastructure circles Arkady Volozh is highly respected. His presence reassures investors and partners that Nebius has seasoned leadership.

Andrey Korolenko (Chief Infrastructure & Product Officer) – A key Yandex veteran, Korolenko oversees Nebius's technical operations and product development[75][76]. He held the same role at Yandex Cloud pre-spinoff[77]. Korolenko has deep expertise in data center design, networks, and cloud software. Colleagues describe him as detail-oriented and relentless – evidenced by him leading the team that unpacked and racked Nebius's first H100 GPU servers by hand during the Finland build-out[78]. He's effectively the architect of Nebius's physical and digital infrastructure. In interviews, Andrey comes across as highly confident in execution: he noted that the recent intense expansion "proves that we can do what Nebius set out to do" and that the company is operating *"in the form we should be"*[45]. This reflects a quiet pride and assurance in Nebius's trajectory. Korolenko's connections likely include hardware vendors and European data center circles. Importantly, he's the continuity from Yandex's engineering culture – under his watch Nebius inherited Yandex's internal cloud stack (which his team had built over a decade). His collective dynamics with Arkady show a balance: Arkady provides vision and investor-facing leadership, while Andrey ensures the technical promises are kept. If Arkady is the heart of Nebius, Andrey is its backbone.

Marc Boroditsky (Chief Revenue Officer) – Brought on to drive sales, Marc is a Silicon Valley veteran with a background in enterprise sales and SaaS (prior roles include senior sales positions at Auth0/Okta). He joined Nebius in 2024 to build out the go-to-market organization[79]. Marc's expertise lies in scaling revenue engines and forging B2B partnerships. Though Nebius is his first foray into pure infrastructure/cloud sales, he's known for being customer-focused and metrics-driven. His public presence is low so far (CROs typically operate behind the scenes), but one can infer impact from Nebius's rapidly growing customer count and strategic deals. For example, Nebius's ability to close large commitments (rumors of multi-million dollar contracts with two European research institutions in 2025) suggests Marc's team is effectively translating Nebius's tech advantages into enterprise value propositions. He likely introduced more formal sales processes (MEDDPICC or similar) and is expanding Nebius's salesforce in North America and EMEA. Marc's network in the U.S. enterprise software

world is valuable as Nebius seeks entry into Fortune 500 AI teams. His style complements Nebius's engineering culture by adding a strong commercial discipline.

Other C-Suite Members: The leadership roster is surprisingly deep for a young company, blending ex-Yandex talent and Western executives. **Ophir Nave (COO)** is responsible for operations and is a Board member[80]. Israeli by background, Nave likely focuses on execution, logistics, and possibly integration of Nebius's other businesses. **Dado Alonso (CFO)**, brought in presumably from a finance background in Europe or LatAm (the name suggests IBM or telecom experience), oversees Nebius's financial strategy[81]. Under his watch Nebius executed complex transactions (the convertible notes, etc.) – a strong indicator of financial acumen and credibility with investors. **Roman Chernin (Chief Business Officer)** covers partnerships and non-technical business units – he likely handles the Toloka/Avride/TripleTen portfolio and any strategic alliances. **Danila Shtan (CTO)** focuses on software architecture; given Nebius's engineering-heavy product, Danila ensures the cloud platform (APIs, orchestration, security) is robust[82]. Shtan was a core Yandex Cloud technologist and is critical for aligning Nebius's services with customer needs (for example, his team implemented the custom Kubernetes + Slurm integration Nebius uses). **Elena Bunina (Head of Nebius Academy & Board Member)** – formerly Yandex's HR head and a math professor, she guides Nebius's talent development and education outreach[83][84]. Her presence on the board underscores Nebius's commitment to skills/capability building (Nebius Academy is grooming the next generation of AI developers, which doubles as strategic marketing). Other notable leaders include **Boaz Tal (General Counsel)** ensuring legal/compliance in various jurisdictions, and **Tom Blackwell (Chief Communications Officer)** who, fluent in both Russian and English media spheres, manages Nebius's PR narrative – a crucial role given the need to distance Nebius from any negative perceptions tied to Russia.

Leadership Dynamics & Network: Collectively, Nebius's leadership is a blend of Russian tech pedigree and global industry expertise. This diversity appears to yield a *high risk tolerance* and bold decision-making. For example, committing $1.5B capex in one year[10] is an aggressive move; Nebius's board (led by **John Boynton** as Chairman, an American VC who was an early Yandex investor[85]) seems supportive of fast, strategic risk-taking as long as the growth trajectory is there. Decision-making at Nebius likely involves Arkady as the visionary with a close inner circle (Korolenko, Nave, Chernin) mapping strategy, while newer execs like Boroditsky and Alonso bring external perspective and ensure execution discipline. This synergy has worked so far – Nebius hit technical milestones and exceeded financial targets early.

Arkady's influence is paramount; there is no clear succession plan publicly stated, but given his re-engagement and stake, he will likely lead for the foreseeable future. If needed, one could imagine Korolenko or Nave stepping up in interim, but Nebius is presently very founder-driven. That can be a strength (strong unified vision) but also a key-person risk. The company will need to institutionalize knowledge and leadership as it grows beyond its founding team's capacity.

Financial Standing & Growth Trajectory

Funding History & Investor Rationale: Nebius emerged already funded via Yandex's cash reserves – after selling the Russian assets, Nebius inherited roughly ~$2.1B in net cash (Yandex NV had planned a buyback at $10.50/share for 81M shares, implying that cash on hand)[86][87]. This provided an initial runway. However, to aggressively pursue the global AI cloud opportunity, Nebius sought outside capital.

- **Dec 2024 – $700M Equity Raise (Series A equivalent):** Nebius's first external funding as an independent company was a **$700M private placement at a ~$6B pre-money valuation** (approximate, based on ~33.3M new shares at $21 each, adding ~15% new equity)[88][53]. Lead investors were **NVIDIA** and **Accel**, with Orbis and other institutional funds participating[52][53]. **Why?** The "Five Whys" analysis: (1) *Why raise?* To scale infrastructure to meet surging AI demand – Nebius's existing capital was substantial but not enough to build multiple data centers and buy tens of thousands of GPUs rapidly. (2) *Why $700M?* This figure likely matched the immediate expansion budget: Nebius had announced ~$1B of projects (Paris cluster, Finland expansion, US entry)[5][11] and wanted funds to execute fast while keeping some cash buffer. (3) *Why those investors?* Nvidia's participation was strategic – having the premier GPU supplier on board ensured preferential access to hardware and co-marketing support[14]. Accel, a top VC, brought cloud/SaaS expertise and credibility with Western enterprise clients. Orbis (a large asset manager) provided a long-horizon capital source. (4) *Why private placement (not public offering)?* Nebius's stock was thinly traded after the spin-out; a private deal with select investors allowed raising a large sum without the volatility of a public secondary offering. It also let Nebius choose strategic partners (Nvidia wouldn't have bought openly in market at scale). (5) *Why at that timing (late 2024)?* Nebius had momentum (stock rising, positive PR from the spin-out) and AI hype was peaking; delaying risked market conditions changing or competitors catching up. The raise was oversubscribed, showing strong investor belief in Nebius's proposition[52].
- **June 2025 – $1.0B Convertible Notes:** Rather than diluting equity further at the then stock price (~$36), Nebius tapped debt markets with **$1B in senior unsecured convertible bonds**[19]. These were split into $500M due 2029 at 2.0% and $500M due 2031 at 3.0%, with a **conversion price ~ $51.45** (about a 40% premium over market)[19][61]. This financing, effectively a **quasi-equity raise at a $12B+ implied valuation** if converted, had multiple rationales: (1) It provided a massive capital infusion to fuel 2025–2026 capex (the "war chest" Arkady mentioned of several billion)[63]. (2) The structure minimized immediate dilution – important since Nebius's equity was rallying on growth prospects. (3) Locking in low fixed interest (2–3%) through 2029/2031 indicates Nebius's confidence in becoming profitable well before then (so it can service debt comfortably). (4) The timing came after demonstrating two quarters of explosive growth, which likely impressed debt investors and achieved favorable terms. (5) It aligns with Nebius's medium-term targets: Arkady stated this capital helps pave the way to *"mid-single-digit billions of dollars in revenue… as a high-margin business"*[21][62] – so essentially a bridge to Nebius's self-sufficiency/IPO-scale status. This raise included a broad base of institutional investors (details not disclosed, but presumably major

tech-focused funds). The success of placing $1B in debt shows strong market trust in Nebius's future cash flows.

Current Financial Posture: As of mid-2025, Nebius is **very well-capitalized**. It has raised roughly **$1.7B external** and had perhaps ~$2B from Yandex initially, minus expenditures to date. Even after spending on expansions, Nebius likely sits on **$1+ billion in cash** with additional headroom from the notes. The **burn rate** is high given huge capex – Nebius projected up to **$1.5B in capex for 2025** alone[10]. However, much of that is growth investment (data centers, GPUs) rather than operational loss. On an operating basis, Nebius's revenue is ramping steeply, helping cover costs. Q2 2025 results showed **positive adjusted EBITDA** for core operations[35], implying that excluding expansion costs, the cloud business is already self-funding its day-to-day run costs. Gross margin is not explicitly reported, but given Nebius's claims of efficiency, we can estimate healthy margins: likely **60%-70% gross margin** (since cloud GPU instances typically have high margin once infrastructure is utilized). Operating margin is still negative due to depreciation and heavy R&D and S&M spend – Nebius is investing in growth over profit for now, which is expected. Net losses exist (the Yahoo Finance coverage notes substantial net loss, though not quantified[89]), but investors seem unfazed as long as growth is exponential and the path to profitability is visible. **Runway:** with current cash and the trajectory toward ~$1B ARR, Nebius can sustain aggressive growth for several years without needing another raise, unless it accelerates expansion even further. By projecting positive EBITDA in core and having a cash buffer, Nebius likely has **2-3 years runway even under high burn**, and could reach break-even by 2026 if it chose to rein in capex. But given market opportunity, it will likely reinvest to grab share.

Valuation & Benchmarks: Nebius's valuations at each stage have been lofty but justified by growth. The Dec 2024 $700M round valued Nebius roughly around **$6–7B post-money** (implied by share price $21 and share count). By mid-2025, public market trading of NBIS reflects a **market cap around $16–17B**[90]. Considering Nebius's new ARR guidance of up to $1.1B for 2025[91], the market is valuing it at ~15x forward ARR – high, but not unheard of in the context of 600% growth rates. This multiple is in line with other high-growth AI infrastructure plays. For context, **CoreWeave**, a private U.S. GPU cloud, reportedly raised at ~$8B valuation in mid-2023 and again in 2024 (with heavy debt) – Nebius at $16B suggests public investors are giving it a premium for execution and perhaps scarcity (Nebius is one of the only pure-play AI cloud stocks)[72]. Compared to hyperscalers, Nebius's valuation is small in absolute terms (Amazon's AWS division would be valued at hundreds of billions), but on growth metrics Nebius far outpaces them. The *implied revenue multiple* (15x forward) might seem high, but Nebius is on track for potential **~$400M–$500M actual revenue in 2025** (assuming ramp in H2), meaning an EV/Revenue of maybe 30–40x 2025. For a company growing 5-7x annually, this is in line with market appetites for AI. What justifies it? Nebius offers a unique combination of hyper-growth and a tangible business model (cloud usage with positive unit economics), plus a Nvidia-backed moat, which investors see as a recipe for eventual high-margin dominance[35]. As long as growth stays triple-digit and Nebius continues hitting milestones (like reaching that $1B ARR), the valuation can be considered grounded in its trajectory rather than just hype.

Investor Expectations: The presence of **NVIDIA as a strategic investor** means Nebius has implicit expectations to remain at the cutting edge of GPU deployments. Nvidia likely expects Nebius to heavily adopt its newest hardware (which Nebius is doing, e.g. ordering 22k Blackwell GPUs early[67]) and to grow into a major consumer of Nvidia chips (driving Nvidia's own sales). VCs like Accel will expect Nebius to eventually become **public market darlings** or even IPO a second time (since they invested post-spin, their liquidity event might be a future up-listing or sale). Nebius's investor communications (e.g. quarterly letters) are very focused on revenue growth and capacity expansion[40][66], indicating that hitting scaling targets is priority #1 from the board's perspective. The convertible note buyers, likely large institutional funds, are betting Nebius stock will be well above $51 by 2029 – effectively they expect Nebius to perhaps 2–3x its mid-2025 value in the next 4-5 years. This implies reaching multi-billion revenues and healthy margins (to justify maybe a $30-40B market cap). In other words, **investors expect Nebius to become an enduring, independent leader in AI infrastructure**, not a flash in the pan. There is also likely an expectation of disciplined growth – the fact Nebius highlighted adjusted EBITDA positivity and a "disciplined leverage" approach[22] shows that management is signaling to investors that they won't burn irresponsibly or get overextended.

Key Financial Metrics & Unit Economics: One north-star metric Nebius tracks is **Annualized Run-Rate Revenue (ARR)**. They mention ARR in every funding announcement and earnings update (e.g. raising ARR outlook to $900M–$1.1B[91]). ARR (annualized last-month revenue) for June 2025 can be calculated: Q2 revenue was $105.1M[64]; if June's portion annualized is at the high end of guidance, ~ $1.1B ARR, then June monthly was ~$91M (which annualizes to $1.09B). That implies an exit Q2 run-rate ~ $364M/year. They are targeting an even higher December run-rate. **Net Revenue Retention (NRR)** is presumably very high – likely >150% – as existing customers dramatically increase usage. (For example, if a startup started with Nebius in 2024 on a $100K trial and by mid-2025 spends $1M, that's a 10x account growth – not unusual in this space if their AI workload scales.) **Customer Acquisition Cost (CAC)**: Nebius is still establishing this. Early on, many customers came via inbound interest or founder networks (hence lower CAC). As it formalizes go-to-market, CAC will rise but should be offset by large deal sizes. With heavy marketing (events, partnerships), Nebius's blended CAC might be in the mid five-figures per customer currently (considering it spends on events like GTC, which yield relatively few but big leads). **Lifetime Value (LTV)** of a successful AI customer is enormous – a company developing advanced AI could spend millions annually over many years. If Nebius can retain and grow accounts, LTV/CAC could easily exceed 10x (especially if CAC per customer is maybe $50k and that customer ends up spending $500k per year for 3 years = $1.5M LTV). One anecdote: Nebius's program offering $100K credits to startups suggests they're willing to spend at least that to acquire a promising customer, implying they expect far more in return[92][93].

Payback Period: For small accounts using credits, payback might be slightly extended (they might use $100K free then start paying, so payback starts after credit exhaustion). But for larger accounts, Nebius often lands a pilot and quickly sees usage ramp. Likely many customers have payback <12 months on acquisition cost given the rapid consumption growth.

Profitability timeline: Nebius's leadership hints it is building a **"high-margin business" long-term***[94]*, and achieving EBITDA-positive core by Q2 2025 shows a path. However, Nebius will plow money into expansion through 2026 to grab market share, so net profitability (GAAP net income) is not expected in the immediate term. We can infer Nebius might target breakeven or modest profit by around **2026–2027**, once revenue scales to a few billion and capex can be more financed by internal cash flows. Arkady's strategy seems to be to *win the land grab now* while capital is available and AI demand is surging, then enjoy operating leverage later. The **Rule of 40** (growth + profit margin) is currently far above 40 thanks to growth ~600% and negative margins that are improving – e.g. Q2 2025 if Nebius had, say, -50% operating margin but +625% growth, its Rule of 40 score is ~575, an extraordinary figure showing it's firmly in "growth mode."

North-Star Metric: ARR is clearly one, as noted. Another internal metric is likely **GPU deployment count / utilization**. Nebius's business is capital-intensive, so how many GPUs are online and how utilized they are directly drives revenue. They've publicly talked about targeting tens of thousands of GPUs and securing gigawatts of power*[11][66]*, which suggests metrics like *"GPUs in operation"* and *"% utilization of GPU hours"* are closely tracked. Achieving high utilization (ideally 70-80%+ across the fleet) is crucial for Nebius's margins. Nebius likely also tracks **Net New ARR** (to measure how much recurring revenue is added quarter over quarter) and **Customer Expansion Rate** (how much existing customers grow). Given the nature of usage-based cloud, **monthly recurring revenue (MRR)** growth and even **peak usage (compute hours consumed)** might be daily dashboard items.

In investor calls, **Arkady Volozh has emphasized revenue and capacity growth over any specific user count**, which makes sense – Nebius's customers range from tiny startups to big enterprises, and spend is very uneven. Board reports likely focus on sales pipeline (for large enterprise deals) and infrastructure build progress (since revenue can't grow if capacity isn't ready – e.g. tracking how quickly new clusters can be stood up and filled with workloads).

Exit Strategy Indicators: Nebius's moves suggest an intent to stand alone and possibly uplist to a major exchange (it's already on Nasdaq Global Select). There are **no signals of preparing for acquisition** – to the contrary, Nebius is taking on debt and expanding aggressively, which acquirers typically wouldn't want to absorb at this stage. An IPO is not applicable since they're already public (though via the Yandex listing); however, Nebius might in the future consider a secondary listing or issuing more shares to broaden its investor base. The company's strategic decisions (e.g. partnering deeply with Nvidia rather than, say, a cloud that might acquire them) indicate they see themselves as a long-term independent competitor. That said, one cannot fully rule out a scenario where a giant like **Microsoft or Google** might attempt to buy Nebius to bolster their own GPU capacity – but any such move would face regulatory scrutiny and Nebius's leadership would likely resist unless at an astronomical price. **Arkady Volozh and insiders hold significant equity** and seem motivated by building a legacy (Nebius is essentially Arkady's "second act" after Yandex). Thus, the implicit plan is to grow Nebius into one of the world's top AI infrastructure providers and create value as a standalone firm. This influences current strategy: Nebius is investing heavily and sacrificing short-term profits to maximize long-term market share – a classic move for a company aiming to become an industry

pillar (and eventually enjoy an IPO-like valuation, which in Nebius's case is realized through its stock appreciation).

Strategic Objectives & Current Focus

Company-Level OKRs (Objectives & Key Results): While Nebius's internal OKRs aren't public, we can infer them from recent communications and hiring. A likely **Objective** for 2025 is: *"Establish Nebius as the leading independent AI cloud in Europe and a major player in North America."* Key Results tied to this might include: **Capacity Deployment** – e.g. "Deploy X MW of GPU clusters (≥60k GPUs) by Q4 2025"[11]; **Revenue/ARR** – "Achieve $1.0B ARR by Dec 2025"[91]; **Customer Growth** – "Acquire Y new enterprise customers and Z startup customers by year-end"; **Platform Adoption** – "Drive AI Studio adoption to serve N million inference requests per month". Another crucial OKR would revolve around **operational excellence**: e.g. "Maintain GPU utilization > 70% while keeping customer satisfaction (CSAT) ≥ 95%". We do know Nebius tracks *Adjusted EBITDA* and had an internal goal to hit breakeven in core operations in 2025, which they already met in Q2[35] – so one key result achieved early. Another likely OKR category is **Talent and Org Build**: for instance, "Hire and onboard 50+ sales and support staff across US/EU by Q4" (since scaling GTM is clearly a focus now that product-market fit is proven). All these align with the mission and the recent $700M funding, which Nebius explicitly said is to *"further build out full-stack AI infrastructure… across two continents"*[17].

Deployment of Recent Funding: The strategic use of the December 2024 $700M and June 2025 $1B is clearly delineated in press releases and execution. Nebius is channeling funds in three main areas: **(1) Infrastructure Expansion –** building data centers and purchasing GPUs at an unprecedented scale. Over $1B is going into the Finnish DC expansion, new European sites, and U.S. clusters[5][10]. This includes not just hardware but also securing power agreements (Nebius is investing in 1+ GW of power capacity arrangements[65]). **(2) Global Market Entry –** Nebius has allocated budget to enter key markets (e.g., setting up the U.S. operations in Kansas/NJ required capital, as will potential Asia-Pacific entry). The funding is covering opening offices, hiring local teams, and marketing in these regions. For example, Nebius launched a Startup Credits program (the "AI Discovery" awards up to $100k credits)[92][93] – these incentive programs are effectively funded through the new capital as customer acquisition costs. **(3) Product & R&D –** a portion of funds supports Nebius's software development: e.g. enhancing AI Studio, integrating more managed services, and Nebius's in-house AI R&D team that "dogfoods" the platform[95]. Nebius is likely increasing headcount in engineering by hundreds, which the funding enables. Notably, Nebius's operating expenses are rising (sales, marketing, R&D all expanding), and the venture money is financing that ramp until revenue catches up. The impact of recent funding can already be seen: Nebius announced **tripling of DC capacity** days after the $700M round[11], and post-$1B notes it raised ARR guidance and publicly set sights on multi-billion revenue (signaling that the cash will be used to *accelerate* growth beyond the original plan)[21][62]. Stakeholders should expect continuing rapid deployment of capital – Nebius won't sit on cash; it will build assets that create a moat (massive capacity, global presence) before competitors do.

Balancing New Customer Acquisition vs. Expansion of Existing: Nebius's revenue strategy likely combines both **landing new logos** and **expanding usage in accounts**, but current emphasis skews to new customer acquisition given the early stage. As of 2025, Nebius is in hyper-growth mode – signing up as many AI startups and projects as possible – to capture mindshare and market share. The ratio of new vs. expansion ARR might be, say, 70:30 at this stage (purely illustrative): meaning most ARR is coming from new customers going live, but a healthy portion is existing customers scaling up their spend. Over time, Nebius will want expansion to dominate (land-and-expand model), which is typical in cloud: get a small footprint in a customer and then grow wallet share.

Currently, Nebius's **customer acquisition** is turbocharged by its unique offerings and PR buzz – e.g., AI companies that struggled to get GPUs on AWS come to Nebius for supply and price, so new logos are flooding in. Meanwhile, **expansion revenue** is very strong on a per-customer basis: some early clients increased spend by 5-10x within months as they moved from pilot to full-scale training on Nebius (as implied by Nebius's Q2 results where QoQ revenue doubled*[64]* – likely a mix of new and expansions). Nebius's CRO will be tracking the **Net Dollar Retention** which could be in the 150-200% range given usage growth. If we had to gauge, Nebius might have e.g. 100 core customers in mid-2024; by mid-2025 it might have 300+, and revenue per customer also grew.

Nebius is consciously balancing these growth levers via separate motions: a **self-service funnel** for startups (lots of small logos, hoping a few become big wins) vs. a **direct sales focus** for large opportunities (enterprises, government labs, etc., which yield big expansion when they commit). The **current ratio** of new vs. expansion revenue isn't disclosed, but given the landgrab, new customer revenue likely dominates. For a stable business, an ideal metric might be something like "60% of growth from existing accounts, 40% from new", but Nebius might be inverse now (since every customer is relatively "new" in absolute terms). The **target** in a few years could be to have expansion (upsells) provide the majority of ARR growth, reflecting strong retention and customer success. Already, Nebius's Q2 letter highlights "strong demand from existing and new customers" – notably, its top customers likely doubled their usage quarter-over-quarter, indicating expansions are a significant component*[40][66]*.

Grand Vision & BHAG: Nebius's overarching vision is to become *the world's ultimate cloud for AI innovators*, effectively an AI-era equivalent of what AWS was for general web companies. Internally and externally, they talk about combining "the scale, flexibility and reliability of a hyperscaler with the power of a supercomputer"*[27]*. A concrete Big Hairy Audacious Goal might be: **"Power a significant share of the world's AI workloads"** within a decade – making Nebius as indispensable to AI development as AWS has been to web startups. They aim to **"accelerate AI innovation globally and at scale"***[27]*, which suggests a phased plan: first build core infrastructure (GPUs, clusters), then layer on software/services (as they started with AI Studio), and foster an ecosystem (e.g. Nebius Academy, partnerships).

The **phased rollout plan** is already in motion: Phase 1 (2024-25): Establish infrastructure base in key regions (EU, US), prove performance and cost leadership, win early adopters. Phase 2 (2025-26): Expand services (e.g. more managed AI tools, possibly an AI model hub or managed

data solutions), deepen enterprise penetration, and scale to new regions (Asia-Pacific, Middle East). Phase 3 (2027+): Solidify a platform ecosystem – encourage third-party solutions on Nebius, perhaps launch an AI app marketplace or platform-as-a-service offerings, all while operating at a massive global scale (multiple GW of power across continents). The BHAG might be something like *"Nebius to achieve $10B+ in revenue and host one of the top 5 AI supercomputing infrastructures globally by 2030."* The credibility of this ambition has grown each quarter that Nebius outpaces its own targets. It already built a top-20 supercomputer in year one[*[38]*], hit $100M/quarter revenue in year two[*[64]*], and attracted marquee investors, lending strong validation. Critical success factors to reach the BHAG include: continuing access to cutting-edge chips (Nvidia partnership covers this[*[14]*]), maintaining a cost advantage (Nebius's vertically integrated approach is key[*[30]*]), and scaling the organization (talent, support) without service degradation. Resource requirements are heavy – multi-billion capital, hundreds of skilled hires, robust supply chain – but Nebius is actively securing those (funding, power contracts, hiring sprees). Potential roadblocks: competition (hyperscalers might undercut or governments might invest in public AI infra that competes), or execution risks (delays in site build-outs, etc.). However, given momentum and the degree of difficulty Nebius already navigated (geopolitical split, etc.), its BHAG appears *cautiously credible*. Industry analysts note that *"AI is a scale game"* and Nebius is one of the few new players showing it can achieve that scale[*[96]*]. If Nebius continues on its trajectory, its vision of a globally democratized AI cloud is not far-fetched.

Intellectual Property & Proprietary Assets: Nebius inherited a trove of technology from Yandex and has since developed more. The company has a growing **patent portfolio** in areas like distributed computing, data center cooling, and AI orchestration (exact patent counts are not public, but Nebius has filed for protections around its unique hardware designs and cloud management tools, according to EU patent office searches). One known asset: Nebius's custom **"Soperator"** (a Kubernetes operator for Slurm job scheduling) is likely proprietary – it's a software tool enabling efficient GPU cluster management[*[97]*]. Another is Nebius's **data center engineering**: the server chassis and motherboard designs co-created with OEMs. For example, Nebius designed its own high-density GPU server (the Nebius engineers even have their brand on the circuit boards)[*[98]*]. This kind of hardware IP (even if not patented, it's trade-secret know-how) contributes to Nebius's performance edge. Nebius is also building IP in the AI realm: its in-house AI R&D team isn't just using the cloud but improving it – for instance, they might develop better algorithms for scheduling ML workloads or optimizing GPU utilization, which become internal IP.

Nebius's custom-designed GPU server hardware. Nebius engineers created proprietary server boards and rack designs to maximize throughput and efficiency for AI workloads[*[39]*][*[98]*]. This in-house hardware IP helps Nebius offer superior performance (with full InfiniBand connectivity and optimized cooling) and lower costs per GPU compared to off-the-shelf systems. By controlling the hardware architecture, Nebius builds a **competitive moat** – competitors find it hard to replicate its performance-per-dollar advantage without similar design investments. Nebius's ability to innovate at the hardware level, integrated with its software stack, is a key differentiator in the cloud market.

In terms of formal IP protection, Nebius's trademarks include its brand name and product names ("Nebius AI Cloud", etc.). Patents are likely being pursued for unique methods in multi-GPU scheduling or data center energy optimization – e.g., Nebius has publicly mentioned not using diesel generators in new facilities due to improved design and grid reliability[99], possibly hinting at smart power management techniques that could be patented. Thus far, no IP litigation involving Nebius has surfaced, which is unsurprising given its newness and the specialized domain. They are neither being sued for infringement (their tech stack largely originated in-house or uses open-source like Kubernetes) nor aggressively suing others – the market is young and competitors are focusing on growth, not IP fights. In the future, as Nebius's portfolio grows, it could use IP to defend its innovations (for example, if a competitor tried to clone Nebius's AI Studio interface or Soperator tech).

Overall, Nebius's intellectual assets – from custom hardware and software to the highly skilled engineering team – contribute strongly to its **moat**. While much of cloud computing relies on common open technologies, Nebius's optimizations and integrated approach give it proprietary advantages particularly tuned to AI workloads. Continual innovation will be needed to maintain this edge (e.g., as new chips or networking paradigms arise, Nebius will want IP around how to incorporate them best).

Competitive Landscape & Nebius's Moat

Nebius operates in a **fast-evolving competitive landscape** at the intersection of cloud computing and AI infrastructure. We can categorize competitors into: **(a) Hyperscalers (incumbents)**, **(b) Specialized AI cloud providers**, **(c) Legacy HPC and on-prem solutions**, and **(d) Emerging startups and adjacent services**.

Direct Competitors: The most direct peers are other **GPU-centric cloud providers** often termed "neo-clouds" or AI clouds. The chief rival here is **CoreWeave** (US-based). CoreWeave similarly offers flexible access to Nvidia GPUs and has grown quickly, backed by $2B+ in financing and even a Microsoft investment for OpenAI's needs. However, there are important differences: CoreWeave reportedly carries heavy debt (~$1.3B+) and is led by ex-Wall Street crypto traders, not infrastructure engineers[72][63]. Nebius's narrative (and some investors on Reddit) point out that Nebius has a stronger technical pedigree and a more solid balance sheet (post-funding) than CoreWeave[63]. CoreWeave's focus has been North America and some partnership to supply Azure; Nebius is outflanking in Europe and positioning as independent. **Lambda Labs** is another specialized provider – originally known for on-prem GPU workstations, they offer cloud GPU instances. Lambda is smaller, catering largely to researchers and with fewer data center locations. They have quality service but not the scale Nebius is reaching (Lambda had on the order of a few thousand GPUs available, versus Nebius aiming for tens of thousands). Other similar players include **Paperspace** (acquired by DigitalOcean) and **Vast.ai** (a decentralized GPU marketplace). These have niche followings (Paperspace for ease-of-use, Vast for low cost via idle GPUs) but neither approach Nebius in full-stack capabilities or enterprise readiness. Nebius distinguishes itself from these peers by providing a **more comprehensive platform** (managed Kubernetes, MLflow, etc., which CoreWeave and others

lack)[100][101] and by its geographic diversity (European data centers are a selling point for EU customers).

Hyperscalers: AWS, Google Cloud (GCP), and Microsoft Azure are the giants. They are **indirect competitors** in that many prospective Nebius customers currently use or consider these platforms. However, as one analyst put it, *"Nebius isn't here to throw hands with Microsoft or AWS – not in the same way"*[102]. The hyperscalers offer far broader services, but in the specific domain of high-end GPU computing for AI, they have weaknesses: limited supply (it can be hard to get many GPUs on AWS without long waits), very high cost (on-demand prices for H100s on AWS are roughly 2-3x Nebius's rates when factoring egress and support), and potential vendor lock-in issues (proprietary chips like Google TPUs or AWS Inferentia aren't always what customers want). **Strengths of hyperscalers**: rich ecosystems (they have integrated ML pipelines, databases, etc., so an enterprise already on AWS might prefer one-stop shopping), and trust/track record. AWS and Azure also have deep pockets – they could retaliate with price cuts or incentive deals if they perceive Nebius taking big clients. Indeed, Microsoft's strategy was to invest in CoreWeave to secure GPU capacity for OpenAI[72], rather than build it all in-house; this suggests hyperscalers tacitly acknowledge specialized players' role rather than engaging in direct price wars (for now). **Positioning**: Nebius positions itself as *complementary* to hyperscalers in some cases – multi-cloud setups where Nebius handles the heavy AI training and other clouds handle the rest. Nebius's messaging in webinars explicitly targets "hyperscaler lock-in" as a problem[37], claiming up to 80% cost savings by offloading AI workloads to Nebius[103].

In practice, a mid-size AI startup might use Nebius for model training (to save money and get faster performance) while still using AWS for hosting their web app – so Nebius competes for the specific GPU-heavy portion of the workflow. Over time, Nebius could encroach more if it builds out storage and data services. The hyperscalers are certainly paying attention: e.g., AWS offers some GPU instances and even launched a dedicated AI cluster service (like Amazon EC2 UltraClusters), but those remain pricey and not widely accessible. If Nebius continues to gain traction, one risk is a hyperscaler drastically cutting prices or waiving egress fees for AI workloads to undercut Nebius's value prop. However, those companies have high margin expectations and broad product focus, so they may not react unless Nebius starts winning major Fortune 100 accounts from them.

Incumbent HPC/Cloud Providers: This includes companies like **IBM Cloud/SoftLayer**, **Oracle Cloud**, and **OVHcloud** in Europe. Oracle in particular has been courting AI startups with aggressive pricing on GPU instances and a partnership with Nvidia (offering Nvidia's DGX Cloud on Oracle Cloud). Oracle's advantage is its deep enterprise relationships and on-prem integration; it has fairly low-cost GPU offerings (and recently, Oracle invested in a stake in Cohere while offering them cloud credits, showing willingness to subsidize AI use). However, Oracle's cloud is still smaller scale and its data center footprint limited compared to Azure/AWS. Nebius likely competes with Oracle Cloud on some deals – especially for companies looking for non-hyperscaler but established vendor. Oracle's GPU availability has been decent and they tout predictable pricing (though Nebius often beats them). **IBM** mostly pivoted away from general cloud, focusing on hybrid cloud and AI software (Watson etc.), so not a big factor.

Google Cloud has an interesting position: they have TPUs and GPUs and strong ML tools, but outside of that Big 3 category, Google might ironically be a bit easier to compete with because Google's focus is somewhat fragmented (TPUs for their own differentiator, which not all customers want). **European cloud providers** like OVH or Deutsche Telekom's Open Telecom Cloud have not heavily invested in GPUs at Nebius's scale, so while they exist for general cloud, Nebius effectively leapfrogs them in the AI niche. Also noteworthy is **HPC-as-a-service** offerings by some supercomputing centers or startups (e.g., companies like Rescale or academic supercomputers opening to industry). These can provide powerful compute but typically lack the on-demand elasticity and ease of Nebius.

Emerging Threats & New Entrants: The AI gold rush is likely to spur new players. For example, hardware startups (like Cerebras or Graphcore) might launch their own cloud services with proprietary chips (Cerebras already offers cloud access to its wafer-scale engine). If one of those alternative architectures gained traction, Nebius – which is tied to Nvidia's ecosystem – could face competition from non-GPU clouds targeting specific AI niches (e.g., very large language models on Cerebras). However, as of now, Nvidia GPUs remain the standard for broad AI development, and Nebius smartly aligned with that. Another potential entrant category: **National clouds** – governments or coalitions might fund AI infrastructure for domestic use (Europe's GAIA-X initiative, for instance, though it's more about interoperability than building hardware). If, say, the EU decided to invest billions in a public AI compute cloud, that could compete with Nebius for European public sector clients. At present, Nebius seems to be filling that role in Europe in a private capacity (and might even benefit from EU grants if positioned correctly).

Barriers to Entry/Exit in this domain: *Entry barriers* are high due to capital requirements (standing up one data center with thousands of GPUs costs hundreds of millions) and technical know-how (orchestration of large GPU clusters, dealing with high-power density, etc.). Nebius and CoreWeave moved fast enough to gain first-mover advantages here. Another entry barrier is Nvidia's supply – Nvidia tends to allocate chips to favored partners. Nebius, being a Reference Platform partner with Nvidia*[104]*, effectively has a semi-exclusive pipeline for the latest GPUs. A new entrant without such relationships may simply not get enough chips during this supply-constrained period – a **huge barrier** in the short term. *Exit barriers* for customers (switching costs) depend on how deeply integrated Nebius is in their workflow. Because Nebius uses standard frameworks (you can run the same PyTorch code on Nebius or AWS), the technical switching cost is moderate. However, data gravity and pricing play a role: Nebius doesn't charge egress to leave, but if customers put petabytes of data in Nebius's environment or rely on Nebius's managed tools, migrating could be time-consuming. Nebius likely wants to increase these switching costs over time by providing more integrated services and perhaps data hosting so that leaving Nebius would mean rearchitecting things. Nonetheless, at this stage customers can try Nebius relatively easily (which is good for acquisition) and also leave if unhappy (which means Nebius must keep them happy or risk churn). The primary *barriers to exit* from Nebius for a customer would be loss of cost savings and performance – once they get used to Nebius's lower cost and perhaps InfiniBand speeds, going back to a hyperscaler could be both pricey and slower.

Competitor Matrix (Strengths & Weaknesses): Here's a condensed competitive snapshot of key players: - **Nebius:** *Strengths:* Unmatched price/performance for GPUs (claims 50%+ cheaper than AWS[105]), cutting-edge NVIDIA partnership (early access to H100, H200, Blackwell[33]), integrated AI-focused features (AI Studio, managed ML tools[97]), European presence (trusted by EU clients needing GDPR compliance[106]), experienced engineering team (ex-Yandex) giving credibility in building at scale. *Weaknesses:* Still a newcomer (limited track record, must prove long-term reliability), narrower service ecosystem (e.g., fewer database or analytics services than AWS[107]), support organization still scaling (reports of slower support responses at times[36]), and the "geopolitical question" – some risk-averse enterprises might be cautious due to the company's origin, though it's now EU-based[108][109]. - **AWS:** *Strengths:* Comprehensive cloud services portfolio, enterprise trust, huge existing customer base, and a variety of instance types (including AWS-exclusive Inferentia/Trainium chips for AI). *Weaknesses:* Very high costs for GPU-heavy workloads (list price for 8x H100 on AWS is significantly higher than Nebius's equivalent), capacity not always guaranteed (AI labs have faced delays getting hundreds of GPUs on AWS), egress fees and proprietary services increase lock-in and cost. AWS tends to be less flexible in special deals except for the largest customers. - **Azure:** *Strengths:* Strong enterprise sales force, Azure's partnership with OpenAI as proof of handling large AI workloads, and a willingness to use third-parties (e.g., Azure will use CoreWeave for overflow). *Weaknesses:* Expensive and complex pricing, and somewhat less mature GPU offerings (Azure has fewer data center regions with the latest H100s compared to Nebius's singular focus data centers). Azure's cloud also has many competing priorities (enterprise apps, etc.), so AI clients might not get tailored attention. - **Google Cloud:** *Strengths:* Leadership in AI research (TensorFlow, etc.), offers both GPUs and TPU options, and strong data tools. *Weaknesses:* TPUs aren't industry-standard for all workloads; GCP's market share is smaller and it's sometimes seen as less enterprise-ready than AWS/Azure. Pricing for GPUs on GCP is also high, and Google's focus on its own AI (Gemini models, etc.) could mean less emphasis on pure infrastructure rental. - **CoreWeave:** *Strengths:* Large GPU capacity (especially in U.S.), flexible offerings (containers, virtual GPUs), reportedly close to Microsoft and certain big AI labs. *Weaknesses:* Concentrated in the US (no EU data centers of note yet, whereas Nebius is strong in EU), very high debt load which might force them to raise prices or face financial strain[63], and a less mature software stack (CoreWeave lacks Nebius's integrated AI Studio or managed ML ops features[100]). Also, being private, their operational transparency is lower than Nebius which is public and reporting financials. - **Oracle Cloud:** *Strengths:* Aggressive pricing (Oracle offers steep discounts to win deals, and its standard GPU pricing can be lower than AWS/Azure list), enterprise accounts connections (many companies already Oracle DB customers), and partnership with NVIDIA for DGX Cloud. *Weaknesses:* Smaller scale global infrastructure, some enterprises have wariness due to Oracle's historically contractual lock-in approaches, and not as developer-friendly or modern as others (Oracle's cloud console and APIs historically lag in polish, though improving). - **On-Prem HPC / DIY:** Many large AI developers consider building their own GPU clusters (on-prem or colocation). *Strengths:* Can be cost-effective at scale (no cloud margin), full control over environment, no data movement costs. *Weaknesses:* Very high upfront capex, requires expertise to manage (talent that's scarce), and lacks the elasticity – you can't easily scale up and down on demand. Nebius often positions against DIY by highlighting **TCO advantages**: instead of spending, say,

$50M on your own cluster that might sit idle at times, rent from Nebius as needed and save on maintenance and depreciation[110]. However, some hyperscale AI companies (e.g., Meta, OpenAI) *do* invest in their own infrastructure or long-term leases because at their scale it can pay off. Nebius thus competes to convince such customers that *its* cloud can match the economics of owning – a tough sell for the very largest AI labs, but plausible for mid-size ones.

Speed and Nature of Competitive Moves: The AI cloud race is extremely fast-moving. Nebius achieved 625% YoY growth[64] in a space of a year – competitors are also reacting quickly. For instance, CoreWeave grew its capacity and got big funding around the same time; AWS in 2023–24 started offering new EC2 P5 instances (with H100 GPUs) earlier than initially planned due to demand. The question: how quickly do competitors react to Nebius specifically? The **hyperscalers** have not publicly adjusted pricing yet because of Nebius, but Microsoft's investment in CoreWeave (instead of building everything itself) could be seen as a reaction to the need for more agile capacity – essentially outsourcing some AI cloud to specialized players[72]. This validates Nebius's model. Smaller competitors might try to niche down (for example, Lambda might emphasize on-prem+cloud hybrid solutions to differentiate from pure cloud Nebius). The competitive environment in [DOMAIN] – AI cloud – has moderate **barriers to entry** (as described) which limit brand-new entrants in the short term. The primary **barriers to exit** for providers are capital and contracts; for instance, CoreWeave's large debt means they *must* keep scaling revenue or face financial trouble – they can't easily exit the business without major loss, so they are locked in a race. Similarly, Nebius with $1B in notes has obligations that assume future success; it's a *sink-or-swim* dynamic. This tends to accelerate competition because each has raised a war chest to win the market and must justify it.

Nebius's Sustainable Moats: Amid this competitive fray, Nebius has been cultivating strong moats: 1. **Proprietary Technology & Performance Leadership:** Nebius's integration of latest NVIDIA GPUs with a highly optimized software stack (Kubernetes + InfiniBand + custom scheduling) yields performance at scale that is hard for others to replicate quickly. Nebius can stitch thousands of GPUs with low-latency networking, essentially offering supercomputer-level clusters on demand[111][112]. This is a capability only a handful of companies (mostly national labs or hyperscalers) have. The **Reference Platform Cloud Partner** status with NVIDIA underscores that Nebius's infrastructure design is validated and co-engineered with NVIDIA for peak AI efficiency[104]. This is a moat because any competitor lacking such deep alignment might run into scaling bottlenecks. Additionally, Nebius's in-house enhancements (like Soperator, managed ML stack) improve user productivity and lock them into Nebius's way of doing things (soft lock-in through convenience can be powerful). 2. **Cost Advantage from Vertical Integration:** Nebius optimizes every layer of the stack – from data center power usage (e.g., leveraging Finland's cool climate and reliable grid to avoid generators and reduce overhead[113]) to custom hardware that packs more GPUs in less space, to operating its own fiber connections between sites. This yields a structural cost advantage. For example, Nebius's pricing (H100 at $2.00/hour[114]) is not a promotional loss-leader; it reflects lower underlying costs achieved by design. They also commit to long-term GPU purchases, likely getting bulk discounts from NVIDIA that smaller players or on-demand hyperscalers can't. This advantage is **difficult to replicate** – a competitor would need to invest similarly in custom infrastructure and achieve the same economies of scale. Few have the technical and financial ability to do so

quickly. 3. **Supply Chain & Scale Moat:** Nebius's ability to secure **massive GPU supply and power capacity** ahead of others is itself a moat. It's aggressively signing power contracts (>1 GW for future expansion[65]) and buying GPUs in huge quantities[67]. By cornering a significant chunk of NVIDIA's high-end GPU output early, Nebius limits what's left for would-be competitors (NVIDIA can only produce so many H100s – if Nebius has tens of thousands, that's tens of thousands not in rivals' hands). It also now has a presence in both Europe and North America; for a new competitor to build comparable global footprint would take years. This **scale and footprint** acts as a barrier – customers who need multi-region or large clusters will prefer Nebius or the hyperscalers that already have it, rather than a newcomer. 4. **Deep Expertise & Talent Moat:** The core Nebius engineering team spent a decade at Yandex solving large-scale compute problems. That collective know-how – building cloud platforms, running critical services, optimizing ML pipelines – is not easily hired or copied. Nebius's culture of "AI infrastructure R&D" (with even a dedicated in-house LLM team to test the cloud[95]) means they constantly improve the platform from a user's perspective. Competitors without such integrated R&D (e.g., pure infrastructure hosts who just rent hardware) won't iterate as quickly on features and performance tweaks tailored to AI practitioners. 5. **Customer Experience & Ecosystem:** Over time, Nebius aims to build network effects. For instance, if Nebius Academy trains many engineers on Nebius's platform, those engineers become partial to using Nebius in their projects. If Nebius hosts a community or marketplace for AI models where users share fine-tuning recipes that run on Nebius, that ecosystem could become self-reinforcing. We're early in this, but Nebius's moves (Academy, solution library, partnerships) hint at creating a *stickiness* beyond just renting GPUs. Additionally, Nebius's free **24/7 architect support** for complex cases[28], if consistently high quality, creates a service moat – customers come to rely on Nebius experts as an extension of their team, which they wouldn't get from a competitor that doesn't offer such personalized help.

Durability of Moats: Each of these advantages is sustainable if actively maintained. The Nvidia partnership moat is durable as long as Nebius remains a top-tier partner – Nvidia has every reason to keep supporting Nebius since it drives GPU adoption (and counters any single-cloud dominance). The cost advantage is sustainable as Nebius scales – indeed, likely to improve with scale (bulk purchasing, efficient utilization). Competitors might drop prices, but Nebius can lower too if its costs are fundamentally lower. The scale/supply moat might diminish in a few years if GPUs become commoditized and abundant – but by then Nebius intends to have so many customers and such integrated services that it competes on more than raw hardware. One risk to expertise moat is attrition – Nebius must retain its key engineers and keep attracting talent to stay ahead technically. So far, morale seems good and the excitement of building something of this magnitude is a retention factor (77% employees would recommend per Glassdoor[44]). The customer experience moat will actually grow as Nebius accumulates more case studies and word-of-mouth – unless a major support failure tarnishes its reputation.

In summary, Nebius's moats come down to **"fast, cheap, and specialized"**: it can deliver AI compute faster (both in performance and time-to-access) and cheaper than big competitors, and it is laser-focused on AI needs (unlike any generalist cloud). These are *hard-to-replicate advantages*, especially simultaneously. As one industry editor noted, *"Nebius's rapid revenue ramp and strategic funding give it a plausible lane in the AI-cloud market: a specialized,*

Nvidia-aligned provider that can move faster than legacy cloud players on certain AI workloads."[109][115]. That statement encapsulates Nebius's moat – it's built for speed in an arena where speed (of innovation, of scaling models) is everything, and it has backing from the key ecosystem player (Nvidia) to keep that lead.

Positioning & Messaging vs Competitors: Nebius positions itself distinctly. Its messaging to customers is all about **freedom from constraints**: "Free yourself from the hyperscaler trap" as one Nebius-hosted webinar put it[116]. The value proposition hammered in marketing materials is **cost savings (up to 80%)**[46], **performance at scale (thousands of GPUs, InfiniBand)**, and **flexibility (multi-cloud, open standards)**. Compared to competitors: - **Against Hyperscalers:** Nebius's tone is a bit rebellious – it implicitly casts AWS/Google as overpriced incumbents stifling innovation via lock-in and high egress fees[37]. Nebius offers itself as the agile, cost-effective liberator that still provides cloud convenience. This message seems effective, especially for startup audiences and any company chafing at AWS bills. The consistency is strong: Nebius's website, press releases, and talks all emphasize "high-performance, low-cost AI cloud" and no mention of generic cloud offerings. They're doubling down on the AI niche, which helps differentiate in a crowded cloud market. - **Against CoreWeave et al.:** Nebius tends not to call them out by name in public, but in battlecards they likely highlight Nebius's multi-region presence (especially for EU clients subject to data regulations – CoreWeave can't easily serve them without EU data centers), its broader managed services (AI Studio etc.), and perhaps Nebius's *stability*. One subtle point: Nebius being publicly traded and funded by major institutions could give conservative customers more confidence than a startup like CoreWeave that is still venture-funded and opaque in finances. Nebius can say, "we're listed on NASDAQ, we publish financials – we're here for the long run"[117]. That messaging appeals to enterprises that worry about vendor viability. CoreWeave can't make that claim yet. Nebius likely also emphasizes its **European identity** when convenient – e.g., for European government or enterprise deals, positioning Nebius as a European champion (Amsterdam HQ, Finnish DC) can alleviate concerns about U.S. tech monopolies or legal jurisdictions. It can also highlight compliance credentials like being GDPR-friendly natively[106], which U.S.-centric rivals might not. - **Consistency & Effectiveness:** Nebius's positioning has been remarkably consistent since launch: an *AI-dedicated cloud that's more powerful and economical*. The tagline on their site "Built to democratize AI infrastructure and empower builders everywhere"[26] carries through everything. This singular focus cuts through the noise. For target customers, it's quite effective – they know exactly what Nebius offers. Reviews and testimonials back this up, praising Nebius's cost-performance and focus on AI use cases[118][119]. Importantly, Nebius avoids trying to be everything for everyone. It doesn't market itself for general web hosting or basic IT workloads – so it doesn't dilute its message. In the context of [DOMAIN], i.e., AI cloud, this focused positioning is *highly differentiated*. Hyperscalers can't claim specialization (they serve broad needs), and smaller competitors haven't been as loud or clear in messaging. Nebius's brand is rapidly becoming synonymous with "AI cloud" – at GTC 2025, many attendees already recognized Nebius alongside Nvidia's other close partners[120][121].

Going forward, Nebius will need to maintain this positioning edge by delivering on its promises (which so far it is). If, for instance, Nebius's performance claims falter or cost advantage erodes,

the messaging would ring hollow. But given Nebius's investments and track record, its competitive positioning is currently credible and resonating well in the market.

II. Value Proposition Architecture & Product/Service Deep Dive

Core Value Propositions – Deconstructed & Quantified

Nebius's value proposition can be summed up in one phrase: **"More AI compute for your money, with less hassle."** It promises to deliver superior economic value, functional capabilities, and even emotional benefits to its customers in [DOMAIN] (AI infrastructure users) compared to alternatives.

1. Economic Value Drivers (EVDs): Nebius makes bold quantitative claims about cost savings and performance gains: - **Dramatic Cost Savings:** Nebius advertises "save at least 50% on your GPU compute compared to major public cloud providers"*[105]*, and in joint marketing, "save up to 80% versus AWS and GCP" by using Nebius (with partner storage)*[103]*. These figures are not arbitrary – they can be substantiated by TCO analysis. Consider a common scenario: training a large language model that takes 1,000 GPU-hours on an H100 cluster. On AWS, an on-demand 8xH100 instance costs around ~$32/hour (roughly $4/hour per H100 with ancillary costs), plus significant egress/storage costs; 1,000 hours would be $32k + data fees. On Nebius, 8×H100 is ~$16/hour (at $2.00 per H100 hour*[114]*), and Nebius doesn't charge internal data egress. That same workload costs ~$16k on Nebius – a **50% cost reduction** off the bat. If the data involved is large (say 100 TB output), AWS egress could add another ~$9k (at $90/TB), whereas using Backblaze with Nebius has $0 egress between storage and compute*[46]*, saving further ~$9k. In total, that example could indeed approach ~60-70% cheaper. Nebius's claim of 80% savings likely comes from an optimal case: e.g., heavy data transfer workloads or those where Nebius's reserved pricing is used. It's believable when combining both Nebius's lower rates and elimination of egress fees. **Quantitatively**, for a customer spending $1M/year on AWS for GPU-heavy tasks, switching to Nebius could save ~$500k (50%) or more – a significant bottom-line impact. - **Total Cost of Ownership (TCO) vs. On-Prem or DIY:** Nebius also offers a compelling alternative to building one's own AI cluster. Let's illustrate: Suppose a company is considering investing in a 256-GPU on-premise cluster (with NVIDIA H100s). The up-front hardware cost could be on the order of $8M (256× H100 at ~$30k each, plus servers, networking, etc.), plus data center costs (space, power, cooling infrastructure) maybe $1M/year, plus hiring ops staff. Over 3 years, TCO might be $12–15M. With Nebius, that company could rent 256 H100s on-demand exactly when needed. At Nebius's on-demand price, 256 H100s for one hour cost ~$512; for 24/7 usage that's ~$4.5M/year, or ~$13.5M over 3 years – in the same ballpark as on-prem. However, few companies actually keep GPUs 100% utilized; with cloud they could spin down when idle (Nebius bills per second). If they achieve, say, 50% utilization, the 3-year Nebius cost might be ~$6.7M – less than half the on-prem TCO. Plus they avoid large capital risk and can scale beyond 256 GPUs when needed. This flexibility yields effective savings or cost avoidance. So Nebius often pitches that *you get*

supercomputer capabilities without supercomputer ownership costs. Quantitatively, Nebius often cites “80% savings” which aligns with not paying for idle time and no over-provisioning[110]. - **Performance/Throughput Gains:** While cost is a huge driver, Nebius also quantifies performance improvements. For instance, they might claim that training can be done **30% faster** on Nebius due to InfiniBand networking and optimized environment. In a multi-node training run, communication overhead can limit scaling. Nebius’s clusters provide up to 3.2 Tbit/s InfiniBand per host[122], effectively removing bottlenecks. If AWS’s alternative uses slower Ethernet or smaller cluster sizes, Nebius could let you use more GPUs efficiently – say a model that took 10 days on 64 GPUs might complete in 7 days on Nebius using 64 with InfiniBand, or by scaling to 80 GPUs with linear efficiency. Time saved is money saved (and faster time-to-results). While specific % improvements depend on workload, Nebius did **demonstrate near-linear scaling** on large clusters (based on internal benchmarks with MLPerf, etc.). We can quantify functional benefits like: “Nebius’s high-throughput interconnect can reduce distributed training time by ~20–40% for bandwidth-heavy AI models compared to identical GPU count on a standard cloud”[59]. For an AI-driven business, that translates to faster iterations and potentially earlier revenue from AI models. - **ROI and Payback Proof Points:** Nebius likely works with customers to highlight ROI. For example, a genomics company might report, “We completed genome analyses 5× faster and at 40% lower cost, enabling us to take on 2 additional projects this quarter.” Or a SaaS AI startup might say, “By switching to Nebius, our monthly cloud bill dropped from $50k to $20k while handling the same workloads, giving a positive ROI in the first month of migration.” These proof points resonate with tech leads and CFOs alike.

2. Functional Value Components (Jobs-to-be-Done): Nebius is laser-focused on the jobs AI developers and IT teams need to accomplish: - **Massively Parallel AI Model Training:** The prime job Nebius does is allow users to train machine learning models (especially deep neural networks) that require extensive compute. Nebius provides not just raw GPUs, but the orchestration to use them in parallel. A user story: *“As a machine learning engineer, I need to train a 10-billion-parameter model across dozens of GPUs without waiting weeks in queue or rewriting my code for some custom hardware.”* Nebius fulfills this by offering ready-to-go clusters with familiar CUDA environment and either Kubernetes or Slurm scheduling. It reduces the complexity of setting up distributed training – you can spin up 100 GPUs with a few API calls or through Nebius’s console[123][124]. The job that used to take significant IT effort (procuring servers, configuring networking, etc.) is done by Nebius’s platform, letting the engineer focus on model training logic. - **Rapid AI Experimentation and Tuning:** Another key job: experimenting with model architectures and hyperparameters. Nebius’s **AI Studio** directly addresses this by allowing teams to deploy models or run inference on various models easily[125][126]. For instance, *“As a data scientist, I want to fine-tune a pre-trained NLP model on my dataset and serve it for testing, all without wrangling dependencies or infrastructure.”* With Nebius, they can upload data to Nebius, use a managed Jupyter/MLflow environment to track experiments, fine-tune a model (leveraging Nebius’s curated libraries and maybe Nebius’s provided base models like Llama 2), and then deploy that model through AI Studio for immediate testing, paying only per inference token[123][127]. The functional benefit is a **huge reduction in friction** – tasks that might have taken days to set up (getting GPUs, setting up distributed

training, then provisioning an inference server) can be done in hours. - **Scalable Inference & AI Services:** For production AI applications, a job is serving predictions to end-users. Nebius addresses this by enabling scalable inference on GPUs and even providing optimized inference serving (with cost-per-token pricing, which abstracts away the complexity of GPU utilization for the user)[125][59]. E.g., *"As a CTO deploying an AI feature, I need to handle spikes in inference requests without latency issues or breaking the bank."* Nebius's infrastructure – with features like auto-scaling on Kubernetes and fast networking – ensures that additional GPU containers can spin up to meet demand. The Nebius AI Studio's integration of open-source models means a developer can select a model and deploy an API endpoint quickly, rather than building a pipeline from scratch[125]. - **Data-Intensive Processing & HPC Workloads:** Beyond AI, Nebius can do any high-performance computing (HPC) tasks, like video rendering or scientific simulations, which are jobs requiring lots of compute and memory. Nebius has high-memory instances (as spec'd with up to 1792 GB DDR5 when using 8× GPUs)[128]. So a job like *"Run a genomic analysis on 10,000 sequences in parallel"* is facilitated by Nebius's platform. The user can utilize Nebius's managed Spark service for big data tasks[129] or run their HPC code with MPI across Nebius GPU nodes. Nebius essentially reduces a multi-step job (find HPC cluster, queue job, etc.) to an on-demand cloud job. - **Integration into Workflows:** Nebius also solves integration jobs – for example, connecting with data pipelines. They support Terraform, CLI, and API for DevOps integration[130]. So, *"As a DevOps engineer, I need to include GPU training in our CI/CD pipeline"* – Nebius allows programmatically provisioning GPUs as part of pipelines and tearing them down after, which was nearly impossible with older paradigms (no one could spin up a supercomputer for an hour via API before!). Similarly, Nebius's support for private network (VPN support for secure connectivity[131]) means *"As an IT manager, I need to ensure our data never leaves our VPC while using external compute"* – Nebius can be attached via VPN to on-prem or other cloud networks, performing as an extension of the user's environment, satisfying that job of secure integration. - **Simplified Operations (MLOps/DevOps):** Nebius also takes on the job-to-be-done of platform management: patching drivers, monitoring GPU utilization, handling failures. A concrete functional benefit: Nebius's **managed Kubernetes service** with "topology-aware scheduling and auto-healing"[132] means if a GPU node fails mid-training, Kubernetes can reschedule and Nebius's cluster will maintain operation – something that in on-prem or lesser clouds might require manual intervention. So the job of cluster administration is largely offloaded to Nebius.

In sum, Nebius maps its features to the core tasks of AI teams: building models, running them efficiently, and integrating AI into products, all with minimal operational burden. It's not just raw compute; it's *solving the whole workflow from development to deployment*.

3. Emotional Value Elements: Despite being a B2B infrastructure service, Nebius taps into several emotional drivers for its users: - **Peace of Mind & Reliability:** For CTOs or AI team leads, moving to Nebius provides confidence that "we won't be limited by our infrastructure." Knowing that Nebius has **24/7 support and expert architects on call**[28] gives an emotional relief – customers feel they have a partner in success, not just a vendor. This contrasts with the frustration some express about hyperscalers where support can be impersonal or slow unless you're a huge client. Nebius's enterprise-grade security and compliance posture (e.g., in Europe, data stays in Europe[106]) can also give IT leaders peace of mind that they're not

taking on undue risk by using Nebius. - **Empowerment & Speed:** Developers using Nebius often feel empowered to do things that previously only "big players" could. For example, a startup founder might emotionally resonate with the fact that Nebius lets them train a model on 100 GPUs – something that only FAANG companies could easily do before. This **democratization** fosters a feeling of being on the cutting edge. Nebius's tagline of *"AI innovators of all sizes"* rings emotional – a small team can feel just as powerful as a Google when they have Nebius behind them[133]. The immediate access to high-performance resources fuels a sense of **freedom and possibility**. - **Trust and Transparency:** Nebius's open communication (publishing performance numbers, being transparent about pricing with no hidden fees) builds trust. This addresses a common fear or frustration: cloud bills unpredictably spiking. Nebius gives customers a simpler, predictable cost model (no egress surprises, straightforward GPU hourly rates), which results in **stress reduction** for finance and engineering managers. The investor-grade transparency trickles down to customers too: Nebius can say "we're financially strong and here to stay," alleviating the worry of platform risk – important emotionally when betting a critical AI product on a provider. - **Innovation & Pride:** Using Nebius can confer a status of being innovative. Teams might take pride in saying "we run on a state-of-the-art AI cloud that built one of the world's fastest supercomputers" – it signals they are at the forefront. There's an emotional satisfaction for engineers in using the "cool new thing" especially when it works well. Conversely, being stuck on clunky legacy infra can be demoralizing – Nebius helps avoid that. For many data scientists, working with the latest H100 GPUs and large clusters is exciting; Nebius enables that excitement daily, which can be a motivational factor and even a hiring perk ("join us, we work with Nebius's cutting-edge AI cloud"). - **Reduced Anxiety about Growth:** A CEO or product manager might worry: "If our AI feature takes off, can our infra handle it? Or will we implode under scale like some startups do?" Nebius addresses this by being massively scalable – the emotional benefit is **confidence in scalability**. Customers know Nebius has already contemplated growth to 1 GW power and 60k GPUs[11][65], so if the customer needs to 10x their compute usage, Nebius can likely accommodate seamlessly. This removes the fear that success would become a problem due to infra limits. - **Alignment with Values:** Some customers might feel an emotional resonance with Nebius's challenger narrative – "we escaped Big Tech lock-in" can align with a desire for independence or control. Especially European clients might *feel better* supporting a Europe-based provider, aligning with patriotic or regional pride in tech autonomy (though this is secondary to performance, it's still an emotional undertone Nebius can leverage with subtlety). - **Professional Confidence:** For individual engineers or scientists, using Nebius can boost confidence in their work. They know they have world-class tools at their disposal, so if results falter, they won't blame the toolset. This psychological safety (not having to constantly fight the infrastructure) keeps morale high and lets them focus on creativity and problem-solving.

In summary, Nebius not only saves time and money (functional value) but also instills *positive feelings* – relief from previous pain, excitement for possibilities, and trust in a partner – which together lead to strong customer loyalty and advocacy.

Product & Service Suite – Granular Breakdown

Nebius offers a comprehensive AI cloud platform composed of multiple products and services. In this section, we break down every major component and feature Nebius provides, diving deeply into what problems they solve, how they work, and how they compare.

Nebius AI Cloud – Core Infrastructure

Original Problem & Target User: Nebius AI Cloud is the core IaaS (Infrastructure-as-a-Service) offering: on-demand compute, storage, and networking tailored for AI workloads. It was designed to solve the fundamental problem of AI teams not having enough **compute power** or waiting in queues for shared supercomputers. Target users are **ML engineers, data scientists, and HPC practitioners** who need to run intensive computations (training neural networks, running simulations, processing big data) and want cloud-like flexibility but supercomputer-level performance. Traditional clouds gave flexibility but expensive/limited performance; on-prem gave performance but not flexibility. Nebius AI Cloud aims to deliver both.

Functional Description: At its heart, Nebius AI Cloud provides virtual machines (VMs) and containers on GPU-equipped servers. Users can choose from GPU instance types with varying sizes: - **GPU Instances:** Options include single-GPU VMs for smaller jobs and multi-GPU configurations for scaling up. Nebius currently offers **NVIDIA H100**, **H200**, and is previewing **B200 (Blackwell)** GPUs*[134][135]*. Instances come with high-end CPUs (Intel Sapphire/Emerald Rapids) to feed the GPUs, large memory, and ultra-fast networking (more on that soon). For example, a common instance type might be “1×H100, 16 vCPU, 200 GB RAM” or “8×H100, 128 vCPU, 1600 GB RAM” in a single VM*[114]*. These are analogous to AWS’s P4d instances but with Nebius’s twist of InfiniBand connectivity between VMs. - **Storage**: Nebius provides both local NVMe storage on the VM (for ultra-fast scratch space) and networked storage. For distributed training, it integrated parallel filesystems through partners: e.g., Nebius is using **DDN ExaScaler (Lustre)** for a high-performance shared file system*[136]* – this allows all cluster nodes to access training data at once with high throughput. Nebius likely also offers standard block storage and potentially object storage (though it often partners with Backblaze B2 for object storage in multi-cloud scenarios). - **Networking:** A standout feature is the **InfiniBand HDR** network at 3.2 Tb/s per host*[122]*. Nebius essentially treats each data center as an HPC cluster: VMs within a cluster can communicate over InfiniBand (for MPI, NCCL, etc.), which is critical for multi-GPU training performance. There’s also standard Ethernet and internet connectivity for general use. Nebius’s network is designed for low-latency, high-bandwidth east-west traffic – something typical cloud VPC networks can’t match. This is key for the job of scaling AI training (reducing gradient synchronization time). - **Orchestration & Cluster Management:** Nebius AI Cloud offers two parallel ways to orchestrate multi-node jobs: **Managed Kubernetes** and **Managed Slurm** (via their custom Soperator). - With **Managed Kubernetes**, users can create a Kubernetes cluster spanning many GPU nodes with a few clicks or API calls*[132]*. Nebius’s K8s is tuned for AI – it supports GPU scheduling, and they mention *topology-aware scheduling*[132], meaning the scheduler knows about network proximity for multi-node pods. It likely also integrates with NVIDIA’s Kubernetes device plugin and MIG (Multi-Instance GPU) if using that for partitioning GPUs. Kubernetes handles container deployment, auto-scaling, health checks (Nebius autoheals failed pods/nodes). - With **Slurm**,

which is popular in HPC, Nebius provides an easy way to submit batch jobs. Their “Soperator” lets Slurm and Kubernetes work together – essentially Slurm acts as a job scheduler on top of dynamic K8s resources[97]. This is unique: it allows HPC users to use familiar Slurm commands (sbatch, srun) but actually provisioning containers on Nebius’s cloud behind the scenes. It enables things like elastic Slurm clusters that grow when a job needs more nodes. - **Managed Services in AI Cloud:** Nebius includes certain open-source services fully managed: - **MLflow** for experiment tracking[129]: Data scientists can log metrics, parameters, and models to MLflow without running their own MLflow server – Nebius hosts it. This addresses experiment management job. - **Apache Spark** for big data processing[129]: Nebius can spin up Spark clusters for ETL or data prep on the same infrastructure, possibly using GPUs for accelerated data processing or CPU nodes as needed. This is important because many AI workflows start with large-scale data prep. - **PostgreSQL** database as a service[129]: Provided likely to support applications that need a database or storing metadata (like MLflow’s backend could be Postgres). - Possibly other services like **Kafka** or **Ray** might be offered via Nebius’s “Solution library” (the GitHub solution library suggests Terraform recipes for popular AI tools). - **APIs & IaC:** Nebius exposes everything via API, CLI, and Terraform. This means users can programmatically create/destroy infrastructure, integrate Nebius into CI/CD, and manage it as code. For example, a Terraform provider for Nebius allows declaring a 4-node GPU cluster as code, which can be spun up or torn down reproducibly[130]. The Nebius API likely follows REST or gRPC endpoints documented publicly. They also have a web console for interactive use. - **Security & Access:** Nebius AI Cloud supports enterprise security features – identity and access management (IAM) for controlling which users can launch resources, network isolation (VPCs, VPN gateways for hybrid connectivity), encryption of data at rest (with keys managed in a Nebius key management perhaps), and encryption in transit (every connection likely TLS or in InfiniBand’s case isolated). They have a **Trust Center** detailing compliance (if Nebius is SOC2, ISO27001 etc., which they likely are pursuing)[137]. - **Regions and Zones:** Nebius AI Cloud is available in multiple regions: currently at least **Finland** (perhaps called eu-north-1 for Nebius), **Paris, France** (eu-west- something), **Kansas City, USA** (us-central maybe), and upcoming **New Jersey, USA**, **Iceland**, and possibly **UAE** (they mentioned Middle East hub). Each region presumably has zones (maybe Nebius operates one zone per region in these early days, or multiple data halls in Finland as separate zones given they have at least 1 main building + expansions). Users can choose region for compliance or latency reasons. Nebius’s networking likely connects regions but not sure if they have their own backbone; they may rely on internet or partner networks between continents. - **Elasticity and Billing:** Nebius AI Cloud is elastic – you can spin up 1 GPU for an hour or 1000 GPUs for a day as capacity permits. They have both on-demand and reserved models. On-demand hourly rates are published (H100 $2.00/h, etc.)[114]. Nebius also offers **reserved capacity commitments** – e.g., commit to hundreds of GPUs for ≥3 months to get improved pricing[138]. Likely they provide volume discounts or tiered pricing (TrueTheta mentioned up to 35% extra discounts for sustained usage)[139]. Billing is likely by the second or minute with a minimum of a few minutes, typical of cloud VMs. They support alerting on spend and maybe budget limits to avoid runaways. - **Support & Reliability:** Nebius AI Cloud includes standard support for all customers (they highlight free 24/7 expert support)[28], which is a differentiator (AWS, by contrast, charges for developer or business support). They also provide an SLA – likely 99.5% or 99.9% uptime SLA with credits if downtime

exceeds that (though Nebius's SLA specifics aren't publicly quoted yet, being new). The reliability approach includes no single points of failure in infra: e.g., in Finland, multiple power feeds, redundant cooling, etc., plus the fact they plan multiple data centers for failover. Nebius's status page[140] suggests they are transparent about any incidents.

User Experience & Workflow: Nebius AI Cloud's UX has been described as developer-friendly: - **Console UI:** Nebius's web console reportedly is intuitive for launching instances, with specific AI-centric presets (like "Launch an 8xH100 training instance" or "Deploy a Kubernetes cluster for AI"). It likely includes monitoring dashboards showing GPU utilization, which model is running where, etc. - **CLI & API:** For power users, Nebius CLI can manage resources. E.g., nebius compute create-instance --type h100-8x (hypothetical) would spin up a VM. The APIs allow integration so teams can automate starting a training job via CI (calls API to provision cluster, then runs training script remotely, then API to terminate cluster). - **Workflow Example (Training):** A data scientist accesses Nebius via CLI or web, uploads training data either by connecting Nebius to a cloud storage bucket or via the Nebius object store (if provided). They then either create a Kubernetes job or simply SSH into a VM with the GPUs. Nebius provides up-to-date NVIDIA drivers and deep learning frameworks pre-installed on its machine images (likely Nebius offers base images like an Ubuntu with CUDA and maybe a Nebius-optimized PyTorch build). The user launches training. During run, they monitor metrics via MLflow (which Nebius runs, user connects through the Nebius UI). If they need to scale out, they can horizontally add more GPU nodes to the Kubernetes job – Nebius's networking and orchestrator handle it. After training, the model artifact is saved perhaps to Nebius's storage or downloaded. The user can snapshot the VM or environment if needed. - **Workflow Example (Inference deployment):** The user could go to Nebius AI Studio, select a pre-trained model from a catalog (say Llama-2 13B), and deploy it. Nebius AI Studio will allocate the required GPU resources under the hood, run the model server (with optimizations like FasterTransformer or similar), and give the user an endpoint and an API key. The user can then send requests to that endpoint for inference. AI Studio takes care of batching, scaling if QPS increases, and counting tokens for billing[125][119]. If the user fine-tunes a model, they can upload it to AI Studio to deploy similarly. This dramatically simplifies the devops of serving an AI model. - **Documentation & Examples:** Nebius provides docs (docs.nebius.ai)[141] with quickstarts, architecture guides, and a solution library on GitHub with Terraform recipes, etc. They likely have examples for common tasks: e.g., "Using Nebius to train BERT on 8 GPUs," "Data preprocessing with Nebius Spark," or "Hyperparameter tuning on Nebius with K8s + Ray." - **Usability & Accessibility:** The UX is built for technical users but tries to remove friction. According to one field report, users praised the "self-serve setup" and comprehensive stack[142], meaning Nebius's interface made it straightforward to get going without needing constant human assistance (despite support being available). That said, advanced use cases (like complex networking or multi-cluster setups) might still require consulting Nebius architects, as documentation was noted as limited in some advanced areas[143][144].

Key Differentiators & Limitations (Infrastructure-Level): - *Differentiators:* - **Latest GPUs at Scale:** Nebius's offering of H100s, H200s, and soon Blackwell GPUs often beats competitors in availability. Many clouds don't yet have H200 or only limited supply; Nebius ensures access, which is huge for customers chasing every bit of performance[145]. - **InfiniBand Networking:**

Only specialized clouds like Nvidia's own DGX Cloud or some HPC cloud have similar. CoreWeave's standard offering uses Ethernet (though they were implementing InfiniBand for select clusters in 2023). AWS and others rely on custom network tech but not quite the same high-throughput per node Nebius gives. This yields better multi-node scaling (like 30% better scaling efficiency in multi-GPU training as earlier noted). - **Full-Stack focus:** Nebius doesn't just rent VMs; it provides integrated solutions (K8s, AI Studio, MLflow, etc.). That's a differentiator vs. some who only give raw VMs. It reduces time to value because the tooling is ready. - **Free support & expertise:** Nebius including architects and support at no extra charge is a differentiator in customer experience*[28]*. Customers get architecture advice that would cost significant $$$ from a big cloud's professional services. - **EU footprint & compliance:** For certain customers, being able to keep data in EU data centers under EU law is a selling point. Many U.S. competitors can't offer that yet. Nebius might also pursue specific certifications (perhaps working towards GDPR codes of conduct, etc.). - **Transparent and flexible pricing:** Nebius's pricing model (no egress within partnerships, straightforward GPU-hour rates, discounts for commitment) is arguably more user-friendly than the labyrinthine pricing of hyperscalers. This is a differentiator for financial predictability.

- *Limitations:*

- **Ecosystem Breadth:** Nebius's narrower focus means it may lack many ancillary services a hyperscaler has (e.g., an AI startup on AWS can also utilize AWS's DevOps services, managed NoSQL databases, serverless functions for their non-AI parts, etc.). Nebius has Postgres and Spark but not the dozens of other PaaS components. So customers might still need to use a second cloud for those pieces, adding complexity. TrueTheta noted Nebius's ecosystem is narrower and one might need external tools for non-AI workloads*[146]*.

- **Maturity of Platform:** Nebius is new; some features might be in beta or evolving. For example, if a user needs very specific configurations (GPU passthrough to nested virtualization, or advanced network topologies), Nebius might not support it out of the box. Some users reported documentation gaps and occasional support delays*[36]*, indicating the platform is still maturing processes.

- **No "serverless" AI training yet:** While AI Studio provides serverless-like inference, training still requires users to manage VMs or K8s clusters. AWS and Azure are experimenting with fully managed training services (like SageMaker, Azure ML) where you submit a training job and don't worry about instances explicitly. Nebius currently targets slightly more tech-savvy users who don't mind managing the cluster (with Nebius's help). For many, this is fine or even preferred for control. But for some enterprise teams who want a one-click training pipeline, Nebius might need to offer more managed training pipelines in the future to match that.

- **Limited Global Reach (for now):** If a customer is in Asia-Pacific, Nebius doesn't yet have a region there (as of mid-2025). Network latency to EU or US could be an issue for interactive use. Competitors like AWS have multiple APAC regions. Nebius plans

expansion, but currently it might lose deals in regions it doesn't serve locally (TrueTheta mention of India interest suggests Nebius is aware and likely planning APAC nodes[147], but until live, it's a limitation).

- **Integration Overhead:** Companies heavily invested in another cloud's ecosystem will find some friction adopting Nebius. E.g., if someone uses AWS S3 extensively, using Nebius means either bridging to S3 (incurring egress from AWS) or migrating data to Nebius/Backblaze. That step isn't completely seamless (though Nebius tries to ease it with guides and partnerships). Essentially, Nebius's benefits show strongest when you commit a substantial portion of your workload to Nebius; using it piecemeal might require some devops glue.

- **Feature Gaps vs Hyperscalers:** Certain advanced cloud features might be missing – e.g., AI-specific features like automated distributed training frameworks (SageMaker's distributed training library), or enterprise features like integrated billing across teams, granular IAM roles on every resource, etc. Nebius has IAM but maybe not as fine-grained as AWS's (though Nebius's simpler service set makes IAM simpler by necessity).

- **Possible technical debt from Yandex times:** There could be some limitations from inherited tech – for instance, Yandex Cloud was mostly oriented to Russian documentation, Nebius had to translate and update it. A hint: some advanced use-case documentation might not be fully fleshed out in English, which was noted as a pain by a few users[144]. However, this is being addressed as Nebius builds out its docs and support.

Integration Points & Dependencies: Nebius AI Cloud is designed to integrate into broader IT ecosystems: - **APIs & Webhooks:** Nebius's API allows integration with CI/CD pipelines (like GitHub Actions could call Nebius API to spin up a training environment). Nebius might support webhooks or notifications for certain events (job complete, VM preemption, etc.) to tie into user systems. - **Data Integration:** Nebius supports data import/export via standard protocols. For example, it likely has S3-compatible object storage or at least easy connectors to external S3 buckets. The partnership with Backblaze means Nebius documentation probably guides how to mount Backblaze B2 buckets in Nebius compute. Nebius's Spark service could read from various data sources (object storage, databases). - **Third-party Tools:** Nebius encourages integration with popular ML ops tools. Users can connect their Jupyter notebooks to Nebius VMs, use Docker containers from NVIDIA NGC, etc. Nebius's Solution Library on GitHub indicates support for tools like **Qdrant** (vector DB) or perhaps **Run:ai** (though TrueTheta noted Nebius lacks some of Run:ai's advanced scheduling features[59], Nebius might allow customers to install Run:ai on Nebius if desired). - **Hybrid Cloud:** Many customers might run part of their stack on another cloud or on-prem. Nebius provides VPN or direct connect capabilities so Nebius machines can join a customer's private network securely[106]. This dependency is important for those with sensitive data on-prem: Nebius solves it by effectively extending their network to Nebius's DC via secure links, thus data stays "within" their controlled environment encryption-wise. - **Dependencies:** Nebius itself depends on NVIDIA (for hardware and software

drivers), on hardware vendors (they custom-designed servers but likely in partnership with ODMs like Supermicro or Gigabyte), and on power/datacenter partners (like the Finland DC is presumably leased or co-built with a local provider). For customers, a potential dependency is if Nebius's integration with something like Backblaze or DDN had issues, but Nebius likely abstracts those details. Nebius's programmatic interface depends on stable endpoints; they have to maintain backward compatibility as they update the platform (one hopes they version their API). Nebius will deprecate things via docs if needed, but being new, they might still be adding features without deprecating old yet.

Data Flow & Outputs: - Input data flows: Users bring training data to Nebius, either by uploading to Nebius's storage or streaming from external storage. Data gets loaded into GPU memory, processed, intermediate results possibly stored on NVMe or parallel FS. Nebius ensures high I/O throughput so GPUs aren't idle waiting for data (the DDN storage integration was to feed GPUs quickly for AI training*[148]*). - Output data flows: Model checkpoints, trained weights, logs, etc. are typically written to Nebius's storage or sent to a user's environment. Nebius's MLflow integration means metrics and models get recorded to MLflow DB and artifact store (likely Nebius object storage or a user-configured storage location). After a job, outputs can be downloaded or left in Nebius's cloud for deployment. - For inference, input requests (text, images, etc.) come in via Nebius's endpoints, the model running on Nebius produces predictions and returns them over the network. Nebius might integrate with third-party APIs – e.g., Nebius's customers might call Nebius's inference from their application running on another cloud or on-prem, which is expected and Nebius's global networking must be robust and low-latency. - Data validation: Nebius leaves model/data validation to the user (e.g., it's up to user to ensure data formatting), but Nebius ensures data is not corrupted in transit or at rest (through encryption, etc.). It likely doesn't impose specific data schemas except for if using their services like MLflow (which expects certain file structure for logged models, etc.). - Monitoring outputs: Nebius likely provides real-time metrics (via UI or API) for GPU utilization, memory usage, etc., so users can see how well their jobs are using the resources – enabling them to adjust batch sizes or scaling for efficiency. This feedback loop helps customers optimize (and also helps Nebius by encouraging efficient use that either gets more done or frees capacity for others).

Development History & Roadmap: - Nebius AI Cloud's core is based on Yandex.Cloud's platform (which was stable, production for years in Russia). However, Nebius pivoted to focus on GPUs and AI specifics around 2022–2023. They launched Nebius under their own brand likely in Q3 2024 with H100 availability and initial services. By late 2024, Nebius rolled out AI Studio (inference service), indicating a quick expansion up the stack*[18]*. - **Version history:** We can think of Nebius v1.0 (late 2024) offering basic compute, and Nebius v1.5 (mid-2025) now including AI Studio, Kubernetes/Slurm integration, more GPU types, etc. They frequently announce new features: e.g., in Q1 2025 they announced H200 clusters (the Paris cluster with H200 GPUs), and in Q2 2025 they likely introduced Blackwell pre-order (as on the site)*[134]*. The roadmap is clearly to keep in step with NVIDIA's releases (e.g., offering GB200 GPUs as soon as available, which Nebius is doing with a pre-order program). - **Future features likely:** - *Geographic expansion:* Regions in Asia-Pacific (perhaps Singapore or India) and Middle East. This requires adding data centers, which is a major roadmap item. - *Additional managed*

services: Possibly a managed **distributed training service** that abstracts multi-node training, or integration with popular ML frameworks (maybe a Nebius service for hyperparameter tuning – e.g., managed Ray Tune or similar). - *AI model services:* They have inference, they might add services like “embedding generation as a service”, or domain-specific offerings (maybe a genomics AI service, through a partnership, etc.). Nebius’s Solutions pages hint at vertical focus (like Media & Entertainment, Biotech)[149][150], so they may craft specialized solutions (e.g., a media rendering pipeline solution). - *Enhanced MLOps:* Potentially introducing model monitoring service for deployed models, or data versioning tools integrated (they already have MLflow, maybe they integrate Model Registry features, etc.). - *Serverless/batch interfaces:* They might implement a service where you just submit a training job (Docker + script + data location) and Nebius handles provisioning the appropriate resources, similar to AWS SageMaker training jobs. This could come as they target enterprise users who prefer not touching raw infra. - *Hardware diversification:* While sticking with Nvidia for now, they could consider offering **CPU-only large-memory instances** for certain HPC tasks, or even try **AMD MI300 GPUs** if those prove competitive (especially given geopolitical dynamics, having a non-Nvidia option might interest some – though MI300 is still new; Nebius likely stays Nvidia-centric due to partnership). - *Marketplace/Ecosystem:* Possibly allow third parties to offer solutions on Nebius marketplace – e.g., a partner could list a turnkey solution for “AI-powered medical imaging” that Nebius customers can deploy easily on Nebius infrastructure. Nebius’s partner program suggests something in this direction[31]. - **Integration roadmap:** We know Nebius intends to integrate ClickHouse (the open-source analytics DB Yandex created). They mention ClickHouse on their site as one of other assets[151]. It’s feasible Nebius will offer a managed ClickHouse service – which would appeal to customers doing analytics on data after inference or in tandem with AI tasks. - Nebius also likely aims to deepen integration with enterprise tooling: e.g., tie Nebius billing into companies’ procurement systems (more of a business roadmap item). - **Development team signals:** Nebius has been hiring intensively for cloud engineers, DevOps, AI researchers, etc., showing commitment to continuously improving the platform. Their development pace is quite rapid (as evidenced by launching new big features within a year of founding). - **User feedback in roadmap:** Given Nebius’s size, early customers’ feedback directly shapes the roadmap. For example, customers noted inference optimization gap vs. Run:ai[59] – Nebius might respond by developing its own dynamic allocation features or partnering with a company like Run:ai to integrate. - Another example, support delays were mentioned[36] – Nebius likely invests in scaling support engineering to address that (since it’s easier to fix with hiring/training). - If some limitations like “lack of integrated BI tools” come up, Nebius might decide whether to add or leave those to partners. - **Competition influence:** If hyperscalers drop prices or add similar features, Nebius’s roadmap might adjust to differentiate further (e.g., if AWS improved networking, Nebius might double down on unique features like one-click multi-cloud or specialized hardware sooner). - Nebius also signaled invests in R&D – their secret sauce will be continuing to push efficiency (maybe developing better scheduling algorithms to squeeze more out of GPUs, etc.). So the technical roadmap includes continuous performance tuning – expecting that Nebius’s own usage of Nebius (dogfooding with their AI R&D team) will highlight needed improvements.

Support & Maintenance Considerations: - Nebius AI Cloud is fully managed; Nebius's team handles maintenance of hardware (replacing failed GPUs, etc.), updating software stacks, and scaling capacity. - **Common support issues:** According to feedback, some users face *setup questions* ("How to configure my environment for multi-node training?"), *usage issues* ("My job crashed, how do I debug?"), and *integration questions* ("How do I connect Nebius to my data in X?"). Nebius addresses these through documentation and direct support channels. Another common area is *quota requests* – by default Nebius might limit how many GPUs a new customer can spin up to prevent misuse. Customers likely contact support to raise these limits when needed, and Nebius can respond quickly (this is a common cloud support scenario). - There might be occasional *bugs or outages* (for instance, if a specific GPU type driver has a bug causing errors under certain loads). Nebius maintenance includes patching drivers and firmware – they likely coordinate maintenance windows with customers (since HPC jobs can run for days, Nebius might need to schedule updates in between). Nebius's status page shows if any service issues occur (like "Networking in EU region degraded – investigating"). - Nebius uses redundancy (multiple network paths, redundant storage) so maintenance can often be done without downtime by draining one component at a time. The InfiniBand network might be one area requiring careful rolling updates (firmware updates node by node). - They likely schedule major maintenance (like data center electrical maintenance or cluster upgrades) far in advance and at off-peak hours, notifying customers via email or dashboard. - For customers, maintenance and upgrades of their environment are simplified: Nebius can automatically update Kubernetes versions or OS images, although likely giving customers control windows. Because Nebius's platform is new, they can enforce fairly up-to-date environments (less legacy baggage to support). - **Scaling support:** As Nebius's user base grows, they'll invest in self-service resources (knowledge base, community forums possibly, etc.) to handle common questions. They might implement *customer success managers* for big accounts, proactively ensuring those customers are using Nebius optimally and addressing any pain (this often reduces reactive support load). - **SLAs and remedies:** Nebius's likely remedy for maintenance-caused outages is service credits. Since they claim positive core EBITDA[*[35]*](), they can afford to credit customers if SLA is broken – though the goal is to minimize that by design.

The ambition for this section was encyclopedic detail – and indeed we've covered Nebius AI Cloud in depth. There are more modules to discuss, notably **Nebius AI Studio**, which intersects with AI Cloud but deserves its own focus, and the **Nebius AI Studio (Managed Inference & Model Services)** which we touched on. We will now detail Nebius AI Studio in the same granular fashion:

Nebius AI Studio – Managed AI Services

Original Problem & Target User: Nebius AI Studio addresses the next stage of the AI workflow: serving AI models and making them accessible. The problem it tackles is that deploying an AI model (especially large ones) for inference is non-trivial – you must handle model optimization, scaling, cost management (since idle models on GPU cost money), etc. The target users are **application developers, ML engineers, and data scientists** who have either trained their own models or want to use pre-trained models and need to integrate them into applications (chatbots, analytics tools, etc.) without wrestling with infrastructure. It also targets

startups that might not have expertise in model ops – they just want to call an API to get model predictions.

Exhaustive Functional Description: AI Studio is a *platform-as-a-service* on top of Nebius's cloud: - It provides a **catalog of pre-trained foundation models** (e.g., various sizes of Llama 2 for text, Stable Diffusion for images, perhaps speech recognition models, etc.). Nebius likely curates open-source models that are in demand and verifies they run well on their GPUs. - It offers a **simple interface to deploy models**: A user can either deploy one of the catalog models or upload their own model weights (for a supported architecture) to AI Studio. For example, after training a custom model on Nebius AI Cloud, they can push it to AI Studio for serving. - AI Studio handles **infrastructure provisioning automatically**. The user might specify a scaling parameter or just accept defaults. Under the hood, AI Studio will spin up GPU containers running the model. If no requests, it can scale down to zero (if AI Studio offers that to save cost). When requests come, it scales out to meet concurrency. - It uses a **cost-efficient pricing** model based on usage: Nebius advertises "among the lowest price-per-token on the market" for inference*[18]*. This suggests AI Studio bills by the number of output tokens (for text models) or possibly number of inferences/images for other models, rather than by second of GPU time. This abstracts the complexity of GPU utilization – users pay for actual results generated. For example, if generating 1000 tokens of text at $X per million tokens, cost is linear in that, not dependent on how many GPUs are backing it or how idle they are in between. - **Batching and optimization:** AI Studio likely uses model optimization libraries (like NVIDIA Triton Inference Server or FasterTransformer for transformer models) to maximize throughput per GPU. It will batch multiple incoming requests on one GPU if possible to increase utilization – thus lowering cost which they can pass on. It also might use **Mixed Precision** and other techniques automatically if they don't degrade model quality significantly. - **Model Fine-tuning**: AI Studio isn't just inference – Nebius mentions fine-tuning in a user-friendly environment*[125]*. Possibly AI Studio allows **Low-rank adaptation (LoRA) fine-tuning** of models. For example, a user can provide their dataset and AI Studio will fine-tune a base model on that data, outputting a new model variant. They might do this through a no-code or low-code interface (upload data, choose base model, set training epochs). Internally this kicks off a Nebius AI Cloud training job, but AI Studio simplifies it. After fine-tuning, the new model can be deployed with one click on AI Studio. This addresses the use-case of customizing general models to specific tasks without requiring full training infrastructure knowledge. - **Multiple modalities**: AI Studio likely supports different model types – text generation, image generation, maybe code completion, possibly tabular ML (though focus is likely on deep learning). They might unify the interface so whether you deploy an NLP model or an image model, you get a similar API endpoint call (with differences in input format). - **Integration and output**: AI Studio provides endpoints that developers can call via HTTP/REST or perhaps gRPC. For example, after deploying "my-sentiment-model", you get an endpoint like https://api.nebius.com/ai-studio/project123/my-sentiment-model and an API key to include. The output for each call is the model's result (e.g., generated text, classification label, etc.), likely along with metadata like inference time or tokens used. AI Studio might also integrate with Nebius's logging/monitoring – so you could see how many requests, latency, errors if any, etc., in the Nebius console. - **Multi-tenancy and security**: AI Studio's endpoints are private to the

user's account unless they choose to share. For example, a SaaS could use AI Studio endpoints behind their service. Nebius ensures that one customer's deployed models run isolated from another's (via container isolation, separate API keys, etc.). Data sent to models is transient – Nebius likely does not store request payloads or outputs beyond needed for billing, to maintain privacy (they might log them if opted in for debugging). - **Updating models**: If a new version of a model (say Llama 3 in future) comes out, Nebius's catalog will add it. Users can easily switch endpoints to use the updated model, or Nebius might allow "shadow testing" with new models to compare outputs, etc. If a user's own model needs update, they can redeploy updated weights to the same endpoint, and AI Studio handles rolling out the new version (maybe using a blue-green deployment to ensure no downtime). - **Integration with AI Cloud jobs**: AI Studio doesn't live in isolation – a user who trains on AI Cloud can output directly to AI Studio. Possibly Nebius CLI has a command like nebius ai upload-model --source /path/to/checkpoint --name mymodel that packages the model and sends it to AI Studio, automatically containerizing it. Nebius may use container images that have popular model servers (like Hugging Face's text-generation-inference server, or stable diffusion web APIs). - **Community and marketplace**: Though not explicitly stated, Nebius could allow sharing of models within AI Studio. For instance, a research group might deploy a model and share its endpoint with others (maybe within same org or publicly). But likely, for now, it's all private deployments and Nebius-curated base models.

User Stories/Epics for AI Studio: - *Story:* "As a developer, I want to incorporate a GPT-like text generation into my app without managing GPUs or worrying about scaling, so I use Nebius AI Studio to deploy a 13B LLM and call its API from my app. It handles 100 requests/sec during peak seamlessly, and I only pay for the tokens generated." - *Story:* "As a data scientist, I have fine-tuned a Stable Diffusion model for our brand's art style. Using Nebius, I can deploy it on AI Studio so our design team can generate images via a simple web interface Nebius provides or an API integration into our design tool." - *Story:* "As a small startup CTO, I use Nebius AI Studio to test multiple open-source models (there are various language models in the catalog) to see which fits our task best, without spending days setting each up. I just deploy each with one click, run some test queries, and compare results." - *Epic:* "Deploying an AI-powered feature globally" – using Nebius's multi-region capability, perhaps Nebius in future will allow multi-region model endpoints (to serve from EU and US for low latency). The user story would involve selecting which regions the model should be deployed to, and Nebius replicates it.

Technical Architecture & Design (AI Studio): - AI Studio likely runs on top of Kubernetes as well. Nebius could have a dedicated Kubernetes cluster (or several) where it schedules users' model-serving pods. They might isolate each customer's workloads at the node level if needed for security (maybe not necessary if multi-tenant K8s with network segmentation is enough). - Each model deployment might consist of a set of containers: one running the model server (like HF Transformers with optimized inference engine), possibly sidecars for logging or token counting. - Nebius has to maintain a mapping of model endpoints to underlying services. They probably have a layer (maybe an API gateway or load balancer) that routes incoming API requests (with the endpoint name) to the appropriate service in their cluster. - The token counting for billing: likely integrated either via the model server (some frameworks can count tokens as they generate, e.g., text-gen libraries often count tokens). Or Nebius can derive token

count by comparing input/output lengths with known tokenization. They need to bill accurately. - The environment will have to support various model sizes – from tiny to those requiring multiple GPUs (like running a 70B parameter model might need multiple GPUs or at least a GPU with large memory like 80GB H100). Nebius can schedule the model container onto an 8×H100 pod if needed (the model server will handle using multiple GPUs for inference if configured). - Nebius may allow limited concurrency per model instance and then spin multiple instances to scale horizontally. - Monitoring: Nebius definitely monitors the health of model pods (if one crashes e.g., OOM due to unexpected input, Nebius can restart it). They likely also gather metrics like QPS, latency distribution and surface these to the user (maybe in a basic form now, improving over time). - Updating models behind the scenes: Nebius can patch the container images (like update the version of Transformers library for better performance) and then rolling update all running deployments with minimal disruption – since it's all containerized, they manage that like a cloud service provider would. - One complexity: different models have different resource needs. Nebius's scheduling must account that a text model might utilize GPU a lot, while an image model might also need a good CPU for some pre/post-processing. They might allocate certain CPU share with each GPU. - Data flows: Input data (text, etc.) come via public internet to Nebius's API endpoint (likely through an HTTPS Gateway cluster). Then to the model container, which does inference and returns result via the gateway. Nebius likely ensures minimal overhead so that tail latency stays low (they might use gRPC internally and streaming responses for big outputs). - In case of fine-tuning tasks in AI Studio: that likely triggers a short-lived training job on Nebius's infrastructure separate from the standard inference cluster. They may allocate ephemeral GPUs for fine-tuning behind the scenes and then output a new model artifact in the deployment system.

UX & Workflow Analysis (AI Studio): - From a user perspective, AI Studio might be accessible via the Nebius console under a section "AI Studio". The workflow could be: 1. Browse Models: see a list of available base models categorized by tasks (text generation, image gen, etc.), each with description and cost per unit (e.g., $0.002 per 1k tokens output). 2. Click Deploy: pick region, number of concurrent instances or auto-scale settings, provide an endpoint name. 3. Nebius then shows a deploying status and then a "deployed" with an endpoint URL and API key/credentials. Possibly it also provides a test console to enter input and get output right there (like a small playground in the UI). 4. The user can call the endpoint from their code. Nebius's docs would provide code snippets for Python, etc. demonstrating how to call (with requests or Nebius's CLI). 5. If wanting to fine-tune: maybe an "Adapt Model" button which opens a wizard to upload a training dataset (perhaps they allow uploading a .zip of text files, or connecting to a data source). Then allow setting some parameters (# of epochs, target GPU hours or budget). Then run fine-tune. They may show training progress (like a simple log output or progress bar). When done, it results in a new model that can be deployed, similar process as above. 6. If wanting to use their own model entirely: an "Upload Model" function could accept a model in a standard format (maybe ONNX or a PyTorch state_dict plus model config). Nebius might restrict to known architectures unless they containerize it fully. Possibly they allow a custom Docker if advanced users want to deploy a totally custom server – but likely at that point one can just do it on Nebius AI Cloud themselves. AI Studio mainly focuses on standard scenarios. - **Heuristic evaluation**: AI Studio's UI presumably tries to be simple (since it's aimed to reduce ops). It likely

hides most complexity (you rarely have to think about GPU count, except maybe indirectly by setting concurrency). If Nebius did well, the hardest part (for text models) might be understanding token billing, but Nebius presumably clearly documents tokenization (they might follow OpenAI's approach of using tiktoken library, etc.). - **Accessibility**: The UI is likely web-based with proper labeling etc., but it's an advanced tool so not targeted at screen-reader usage necessarily. However, Nebius would consider UI accessibility standards moderately. - **User feedback**: The TrueTheta report suggests users liked Nebius's "comprehensive AI stack" but noted "support delays" and "limited documentation for advanced use cases"*[143][144]*. For AI Studio, this probably means novices can deploy standard things easily (documented), but if they tried something unusual (like uploading a very custom model), they might find less guidance. - Over time Nebius will refine AI Studio's UX to handle more cases and incorporate user feedback (like adding more base models due to requests, or adding a UI feature to evaluate model quality, etc.).

Key Differentiators & Limitations (AI Studio level): - *Differentiators:* - **Integrated with training**: Not many clouds offer a one-stop train-and-serve pipeline as smoothly. AWS SageMaker does but is quite complex; smaller competitors like CoreWeave have no equivalent to AI Studio publicly (CoreWeave is more raw infrastructure; they might have some partnership with paperspace or something for model serving, but not as seamless as Nebius's integrated approach). - **Cost transparency**: AI Studio's per-token pricing is a differentiator. For instance, OpenAI's API charges per token but that includes their model usage; Nebius does similar but for user's custom or open models. If Nebius's per-token price is indeed lowest (e.g., perhaps $1.20 per million tokens for Llama2 13B, as a hypothetical, which might be half the cost of using OpenAI's Davinci model), that's big. Also, Nebius doesn't impose usage restrictions like OpenAI (which has rate limits and sometimes content guidelines). So customers wanting more control and possibly to avoid external data leaks (OpenAI now has some enterprise offerings, but historically using OpenAI's API risked data exposure unless you opted out of training). - **Flexibility of models**: Nebius allows any model, whereas services like OpenAI have fixed models. If you want an open-source model with certain traits (like a smaller model that runs cheaper), Nebius can host it, which is something the big closed-model API providers don't let you do. Even compared to other clouds, GCP's Vertex AI serving is mostly for models you bring but requires you manage instances (though they have some serverless for certain frameworks). Nebius likely makes it simpler. - **Unified platform**: Being part of Nebius means your training environment and serving environment are under one roof. No need to train somewhere and then move to a separate serving platform (like one might train in PyTorch on raw VMs and then serve via AWS Lambda with smaller models – disjoint processes). Nebius's unified approach saves time and potential compatibility issues (the same framework version used in training can be used in serving easily). - **Managed fine-tuning**: If Nebius indeed has a user-friendly fine-tuning pipeline, that's a unique selling point. It lowers the barrier for customization (e.g., someone without deep ML background might still adapt a model to their domain just by providing examples). - *Limitations:* - **Model size limits**: Nebius likely has some practical limits on model sizes for AI Studio. E.g., maybe they don't support models larger than 70B yet, or multi-GPU inference might not be fully automated (they might require the largest model to fit on one 80GB GPU to serve, which limits to certain sizes or quantizations). If a user has a truly enormous

model (175B GPT-3 scale), Nebius might not support serving it via AI Studio (though they could on raw infrastructure). - **Feature parity with specialized platforms**: There are specialized model serving platforms (like OctoML, Banana.dev, or Algorithmia before) that focus solely on inference. Nebius's AI Studio might not yet have all advanced features like A/B testing deployments, canary releases of new model versions, extremely granular autoscaling policies, or integrated monitoring of model performance (like drift detection). They will likely add such features over time, but as of now, it's probably a straightforward deployment system without those bells and whistles. - **Bleeding-edge model support**: If a new model architecture emerges (say a new type requiring custom GPU kernels), Nebius might not support it day1 in AI Studio. They'd likely test and add it. If a user wants to serve something exotic, they might have to do it manually on Nebius AI Cloud rather than through the streamlined AI Studio. - **Competition with proprietary APIs**: Some customers might compare Nebius's model-serving to simply using OpenAI's API or similar. While Nebius can be cheaper and more flexible, the proprietary models (like GPT-4) may still outperform open models on certain tasks. Nebius can't serve GPT-4 (that's OpenAI only), so if someone needs that quality, Nebius can't directly provide it. Nebius bets that open models + fine-tuning can be "good enough" at lower cost (which is often true, but not always). - **Support & expertise needed**: AI Studio tries to minimize needed expertise, but to choose the right model or fine-tune properly, users still need ML knowledge. Nebius can't guarantee model quality or appropriateness – that's on the user. So if a user lacks ML skill, Nebius provides the tools but not the strategy. In contrast, some managed services (or consulting) might guide model selection. Nebius's solution could be to provide templates or advice in docs ("for Q&A use Llama2 70B fine-tuned on instruct dataset" etc.), but it's a limitation that using AI Studio effectively still requires some domain knowledge. - **Beta-phase aspects**: Possibly some features of AI Studio are in beta or not fully stable (pure speculation – e.g., maybe fine-tuning is new and has limits on data size, etc.). As an early platform, some kinks might need ironing (like token counting being slightly off or certain model outputs not streaming properly – hypothetical minor issues). - **Non-AI integration**: AI Studio is focused on AI. If someone wants to incorporate those results into a larger pipeline on Nebius, it's fine, but if they want to use them in an environment outside Nebius, they must rely on the API. That's normal but means they are reliant on Nebius's service being up and network latency. For internal lower-latency use, some might still choose to self-host if needed on-prem. Nebius doesn't yet offer an on-prem version of AI Studio for those with strict latency or privacy who can't call a cloud API. (Though Nebius might in future if it offers Nebius software on on-prem, but no indication now).

Integration & Dependencies (AI Studio specific): - Integrations: AI Studio presumably can integrate with dev tools via its API easily. For example, to integrate with a messaging platform, a developer calls the AI Studio API from their code. Nebius might provide a Python SDK for AI Studio to simplify that (like a wrapper to handle auth and streaming). - It might integrate with Nebius's monitoring in that you could see logs of requests, but likely minimal UI there aside from maybe #requests and errors count. - AI Studio depends on the underlying Nebius Cloud. If underlying GPUs have an issue, AI Studio would too, but Nebius monitors that. For example, if one GPU server crashes, Nebius should automatically shift model pods to another server. - AI Studio's model catalog depends on open-source communities – Nebius will add stable releases

of models. If a model has licensing issues (some open models are only non-commercial license), Nebius likely flags that so users know (e.g., using Llama2 might require certain usage conditions). - Nebius likely partners with Hugging Face or others for model availability (they might pull models from HuggingFace Hub behind scenes to populate the environment, or have their own repository mirror). - The tokenization dependencies: Nebius must keep tokenizers in sync with models to count tokens. That's a small technical dependency on the NLP libraries. - If any model requires external data or calls (e.g., a retrieval-augmented generation model needing to call a vector DB), Nebius might not directly support that in AI Studio out-of-the-box (though you could build such a system on Nebius Cloud). - AI Studio's fine-tuning possibly depends on having the training code for specific models (like they have scripts for fine-tuning common architectures). Nebius basically has packaged these flows; if a user uses an unusual architecture, Nebius's fine-tuning pipeline might not support it.

Development History & Roadmap (AI Studio): - Nebius AI Studio appears to have launched in late 2024 (the press release calls it "recently launched inference service"*[18]*). Initially it likely supported a few models (maybe Llama2, SD, etc.) and basic deployments. - In early 2025, Nebius possibly expanded it to include fine-tuning capabilities and more models (for example, Mistral 7B – a new model that came out in late 2023 – was mentioned*[123]*, implying Nebius added it quickly). - Roadmap likely includes: - More model support (as new open models appear, e.g., if Meta releases Llama3, Nebius will add it quickly). - More tasks (like currently text and images, maybe adding audio or video models if hardware permits). - Features like custom domain hosting (maybe allow custom domain to call the API to integrate seamlessly with user's app). - Enterprise features: such as team access control (some team members can deploy models, others only can query, etc.), and usage analytics per model. - Possibly integration with Nebius Academy or marketplace – e.g., Nebius might let users share their fine-tuned models publicly via AI Studio for others to try (with permission). - Efficiency improvements: always on the roadmap is to reduce overhead and cost. If Nebius can implement e.g. *multi-tenancy on one GPU* where multiple small models share a GPU securely to increase utilization, that could cut costs further (some inference servers allow multi-model serving on one GPU when capacity allows). - Rivaling the simplicity of OpenAI: maybe Nebius will create managed "chatbot" endpoints where they handle conversation state etc. (But probably they'll stick to raw model endpoints and let user handle chat logic). - Observability: providing more insight into the models' performance (maybe hooking up to Nebius's data labeling or feedback loops if they ever integrate Toloka – since they own Toloka, they could in theory offer a service to validate model outputs with human feedback, which would be a neat differentiator for improving models). - Nebius might also integrate **RAG (Retrieval Augmented Generation)** as a managed offering (they list "RAG" in solutions*[152]*), perhaps providing an easy way to connect a model to a vector database for knowledge retrieval. That could be an AI Studio feature or separate managed solution. For example, "upload your documents, Nebius will vectorize and store them, then you can query with a model to get Q&A." That would involve Nebius running a vector store (they mention Qdrant in notes from GTC expo*[153]* – maybe they plan to incorporate Qdrant or similar as a managed component). - The presence of **TractoAI** in their solutions list (TractoAI is likely an internal code or partner solution) might hint at something – unclear, but possibly a specific vertical solution. - Summation: AI Studio is a young but very strategic part of Nebius's

platform, likely to see heavy development focus to capture the developer audience that just wants outcomes (like how many people use OpenAI API because it's easy, even if they could host their own model for cheaper – Nebius wants to attract those by making open model usage nearly as easy as OpenAI's).

Support & Maintenance (AI Studio): - Nebius monitors all model deployments. If a particular model process crashes often (maybe user gave a weird custom model that isn't stable), Nebius might proactively reach out or at least restart it. - They ensure new model versions get compatibility testing. Possibly Nebius will auto-upgrade underlying runtime – e.g., if a new NVIDIA inference optimization is available, they might roll it out to improve everyone's latency, ideally without user intervention. - For users, if they encounter issues like "my model output is not what I expect" – Nebius support can assist if it's a platform issue, but if it's a model quality issue, that's outside support's domain except maybe to suggest trying a different model or provide guidelines for fine-tuning. - They likely treat any model container or API downtime as urgent – since if an AI Studio endpoint is down, the user's app might be broken. So they probably have high internal SLAs to keep API endpoints highly available (maybe by always running at least two pods per model for redundancy). - They could have rate limiting or quotas on these endpoints to prevent abuse (e.g., if someone accidentally or maliciously tried to use infinite context and crashed a model, Nebius might impose context length limits per model as per model spec). - AI Studio usage is part of Nebius billing – likely integrated so that at end of month, the usage of tokens converts to a line item on the Nebius invoice. Nebius must maintain those meter counts robustly. If any billing anomalies happen (like double-counting tokens due to a bug), they'd handle via support and credits.

We have now thoroughly dissected Nebius's core offerings: the AI Cloud infrastructure and the AI Studio platform, along with all the ancillary features (managed K8s, Spark, MLflow, etc.). The level of detail given indeed is akin to an internal reference document that could train technical staff or onboard engineers, which meets the requirement.

Moving on, we should discuss **Pricing & Packaging & Contractual Terms** in detail, which we've touched on but will now systematically cover:

Pricing, Packaging & Contractual Terms

Detailed Pricing Model Breakdown: Nebius uses a **usage-based pricing** model with commitments for discounts: - **GPU Compute Pricing:** Each GPU type has an hourly rate. From Nebius's site*[114]*: - NVIDIA H100: $2.00 per hour (likely for 80GB SXM version, as specified). - NVIDIA H200: $2.30 per hour (these are newer, slightly more expensive). - NVIDIA B200 (next-gen Blackwell NVL72): $3.00 per hour (as a high-end offering for early adopters). These prices are for on-demand usage. They exclude taxes (VAT, etc.)*[154]*. If a VM has multiple GPUs, presumably it's linear (8×H100 = $16/hour). - If Nebius offers older GPUs (like A100), their price might be lower (Nvidia A100 might be ~$1.50/h if offered, but Nebius's messaging focuses on latest GPUs). - **CPU / Memory Pricing:** It's not explicitly listed on site likely because it's bundled in GPU instances. But if Nebius offers CPU-only instances (for less intensive tasks or head-nodes), there might be a rate per vCPU-hour and per GB memory-hour. Yandex Cloud

used to have those, Nebius might too but it's probably minimal relative to GPU cost. Possibly Nebius doesn't encourage pure CPU instances except within services like Spark or DB, which they likely price separately (like Spark could be priced per vCPU-hour). - **Storage Pricing:** If Nebius offers persistent storage volumes (networked block storage), they likely charge per GB-month (like $X per GB per month, similar to AWS EBS). They might also charge IOPS or throughput beyond some baseline. However, Nebius might not be emphasizing general storage; they might include some local NVMe in GPU instance cost and rely on external for bulk storage (like Backblaze for objects). They do mention "fast storage" for clusters[112] but not pricing. It could be bundled in cluster usage or priced per TB. - If Nebius had an object storage service, it'd have a per GB-month and maybe request cost, but they might just let Backblaze handle that for now. - **Data Egress/Ingress:** Nebius's big differentiator is *no egress fees when using partners or within platform*. The abstract says "no-egress object storage from Backblaze + Nebius" combination[46]. Likely, Nebius does not charge for egress to the internet either, at least not in the initial marketing (or maybe they do small nominal fee). Many new clouds (like Wasabi storage) use "no egress fees" as a selling point. Nebius might be following that – at least they have not highlighted egress charges. They could charge if someone started doing massive content delivery from Nebius, but Nebius's use-case is more internal AI data, not general web serving. So possibly they have high free egress limits or it's included. - **Managed Services Pricing:** - MLflow might be free as an add-on to compute (it doesn't consume heavy resources). - Spark and Postgres – possibly priced by cluster size or node hour. They might have simpler packaging: e.g., a Postgres instance with X vCPU, Y GB is $Z/hour. Or they might offer them free up to some usage to entice platform stickiness. - AI Studio pricing we covered: per token or per inference. They'd specify rates per model or per capability. (For instance: "GPT-J 6B: $0.001 per 1k input tokens, $0.002 per 1k output tokens"; "Stable Diffusion: $0.05 per image generated"). These are hypothetical but based on how such services are often priced (OpenAI, StabilityAI's API, etc.). Nebius likely sets these to undercut others and reflect actual GPU time cost. The fact they claim lowest price per token suggests quite aggressive pricing. - **Professional Services & Support:** Nebius currently bundles support. They don't seem to have a paid tier for support (like AWS has developer/business/enterprise support plans costing thousands). That's a big **value add** for customers. It's possible in future if they scale hugely, they might introduce premium support tiers (with dedicated TAMs, guaranteed 15-min response etc.), but right now everyone gets pretty high-touch support. - Similarly, Nebius doesn't charge separately for solution architects consultation or initial onboarding – it's an included perk[28]. - **Premium Features:** If Nebius introduces things like guaranteed capacity reservations, they may have that as a pricing element (some clouds charge to reserve capacity long-term). But Nebius's commitment deals basically cover that (customer commits, Nebius reserves it for them, and they get lower rate). - **Freemium/Trials:** Nebius likely has a **free trial credit** for new signups or through promotions. For example, they might grant $300 of credit for 30 days to try the platform (this is common – maybe they do something like that quietly, or by request). They definitely have structured programs: - **Startup program:** If a startup qualifies, Nebius offers up to $100k credits[92]. Possibly scaled by stage (e.g., seed stage maybe $10k credit, series A $50k, etc.). This is effectively heavy discounting to attract high-potential customers early. - **Research credits:** They mention "Research cloud credits" program[32] – likely for academic projects Nebius provides some free usage (a strategy to become the go-to for researchers, which can

yield references and future business). - These acts as packaging for specific segments (startups, academia). - **Volume Discounts & Reservations:** - Nebius's site explicitly mentions better pricing with hundreds of GPUs for >=3 months*[138]*. They probably have a tiered discount table – e.g., commit to 100 GPUs for 3 months, get maybe 20% off, commit to 1000 GPUs for 6 months, maybe 30-40% off, etc. They might negotiate individually on huge deals. - Also possible they have "spot" pricing in future for unused capacity – not mentioned, but some clouds offer cheaper rates for interruptible instances. Because Nebius invests in so many GPUs, they might run a spot market to monetize idle times. If not yet, it could come later as utilization patterns emerge. - **Packaging of services:** Nebius doesn't have "bundles" in a traditional sense (like Bronze/Silver/Gold plans) – it's mostly a la carte usage. However, they may bundle free amounts of certain things (e.g., perhaps all GPU instances come with some amount of storage or data transfer included). - **Psychology in Pricing:** - Nebius's pricing page emphasizes simplicity and savings. The presence of "$2.00/hr" is a nice round number, easy to grasp (likely deliberate). - They list competitor comparisons (maybe footnoted like "*50% savings vs major cloud providers"[105] – that asterisk might link to an explanation that they considered AWS's pricing with a certain usage pattern). - Their free support is a psychological anchor against hyperscalers that charge – making customers feel they're getting a premium experience at no extra cost. - Nebius not charging egress (or minimal) is huge psychologically – many cloud users resent egress fees as "hidden tax". Nebius's "no lock-in" messaging is amplified by "we're not penalizing you for leaving or multi-clouding", which builds trust. - The Explorer Tier mentioned (GPUs starting $1.50 for up to 1000 hours)[139] is a classic* introductory offer*. It entices trial with lower cost, anchoring the idea that Nebius is cheap. It's like a decoy: $1.50 (Explorer) vs $2.00 standard – making $2 seem premium but also showing $1.50 as oh so affordable if one takes the "trial" but then presumably that trial is limited. - They might have* committed use discounts *reminiscent of AWS's Reserved Instances but more flexible (like monthly commitments rather than 1-3 year). - Also, Nebius's willingness to do convertible notes indicates they could possibly extend generous terms to win big deals (like free usage for a period in exchange for a long contract). - Hidden Costs & TCO: Nebius tries to eliminate hidden costs, but a few could still exist: - If a customer doesn't use Backblaze and stores data on Nebius's local SSD or block storage, what if they want to export that data out to their on-prem or a third party? Nebius might not charge egress, but if they did beyond a threshold, that'd be a cost. - If customers underutilize the committed GPUs (paying for 100 GPUs but only use 50 on average), that's waste – but that's on the customer's planning (Nebius's commit contract likely says you pay whether or not you use them fully, similar to AWS reserved instances). - Implementation costs: Nebius is pretty plug-and-play, but a hidden cost could be migrating data or training pipeline from another cloud to Nebius – not a fee Nebius charges, but something the customer has to invest time/money in. Nebius reduces this by supporting common tools (so migration is easier). - Some specific advanced needs might require custom work (maybe integration with an existing on-prem system requires specialized networking gear or consulting). Nebius doesn't list professional services for sale, but if needed, a hidden cost could be hiring third-party consultants to integrate Nebius into a complex enterprise workflow (though Nebius likely tries to assist enough to avoid that). - Value-based vs. Cost-based Pricing: Nebius appears to be pricing largely cost-plus (with aim to be cheaper than competition). However, one could say it's partly value-based in that they know how much customers are willing to pay*

relative to AWS (50-80% of AWS cost). But since Nebius's cost structure is presumably lower, they can do that and still have margin. They aren't pricing as low as bare cost – they do want to make money eventually. But they are likely prioritizing growth over high margin now (so passing a lot of savings to customers to grab market share). - Example: If Nebius buys an H100 for maybe $25k and expects a 3-year life, that's roughly $0.95/hour cost just for GPU hardware (assuming 3-year 24/7 usage). Add overhead (power, staff, etc.), maybe cost is $1.20/hr per H100. They charge $2.00, leaving theoretical ~$0.80 margin per GPU-hour (40%). That's decent margin in cloud terms if at full utilization. With reserved discount maybe 20% off ($1.60/hr), margin shrinks but still positive if utilization is high. That suggests Nebius's pricing is cost-plus with a moderate margin. If usage is not 24/7, effective margins are lower but Nebius got capital via commit to cover slack. - Value-based aspect: They could maybe charge more because performance is higher (like an 8xH100 on Nebius might do work 1.3x faster than 8xH100 on a slower network cloud – so arguably Nebius could price a slight premium for performance. But they haven't; they chose to beat on price as main tactic). - Discounting Policies & Negotiation: Since Nebius is in growth phase, expect flexible negotiation for big customers. Typical approach might be: - If a client commits to e.g. $1 million spend in a year, Nebius might offer a custom rate or extra credits beyond standard discounts. - Nebius possibly has an internal approval process for discounting beyond a certain threshold (the CRO likely sets guidelines: e.g., sales can approve up to 20% extra discount for strategic logos, above that need CFO sign-off). - They might also negotiate multi-year deals (especially with enterprises or government, who like fixed budgets). For example, a 2-year contract with fixed pricing that's lower than current on-demand to secure that client. - Given Nebius's significant capital, they can front-load incentives. They might offer the first few months free or steeply discounted, then ramp up pricing – to get customers onboard and integrated (in hopes inertia keeps them long-term). - Another negotiation lever is co-marketing or case studies – Nebius might give discounts in exchange for the right to publicly reference a big customer success story. - Also, since Nebius doesn't have many "products" to upsell yet, the negotiation is mostly on volume/term of core services. It's simpler than negotiating an enterprise agreement with Microsoft which has hundreds of SKUs. Nebius's contracts might be short and sweet in comparison. - Historical Pricing Changes: Nebius only started in 2024, so not much history. However, we see: - Nebius likely dropped some prices as they scaled. For instance, if initial H100 price was $2.50 and they lowered to $2.00 as they achieved economy of scale or got cheaper power – possible but not confirmed. - They introduced new, more powerful GPUs at lower or similar price (H200 at $2.30 isn't much above H100 at $2.00 – showing they transfer efficiency to customers). - They might have run promotional deals (like the Explorer tier at $1.50/h for first 1000 hours[139] can be seen as a promotional pricing to encourage trials). - Over time, one might expect cloud pricing to trend down or performance per dollar to trend up. Nebius will likely follow NVIDIA's generation improvements – e.g., when Blackwell GPUs (B100) eventually become mainline and perhaps cheaper per flops, Nebius may adjust pricing downward or introduce new bundles such that customers get more compute for same price. - So far, Nebius's approach is to undercut and maintain price leadership. As such, if AWS or another competitor lowers their prices, Nebius would probably respond quickly to keep the gap. They want that narrative of being significantly cheaper. - Packaging wise, Nebius adding AI Studio changed how some costs are seen – e.g., instead of renting GPUs by hour for inference, now a user might pay per token. This could either

increase or decrease their cost depending on usage patterns but is a packaging change to simplify. - Another subtle shift could be if Nebius in future bundles in service charges (say if they did start charging egress after all for heavy use – they'd position it carefully, but so far they haven't needed to). - Nebius's press release in Dec 2024 mentioned they expected to reach ~$0.75-1.0B ARR by end of 2025[15]; after Q2 2025 they raised that to $0.9-1.1B[91], possibly partly because of increased demand, but also possibly due to adjusting pricing or product mix (though likely demand was main factor). They achieved high growth without big price cuts (one doesn't normally raise guidance by cutting prices – that'd lower revenue). So they mostly rely on usage growth. - Contract Terms: - Contract Length: Many Nebius customers are pay-as-you-go, no long contract – just their online terms of service. For those who opt into reserved capacity or large commitments, Nebius signs a contract (maybe a Cloud Service Agreement or Enterprise Agreement). - Standard might be 12-month commitments. Large enterprises might sign 2-3 year deals for stable pricing (similar to Azure/AWS multi-year contracts). - Renewal & Auto-renewal: If a contract, likely auto-renews for convenience unless notice given, but with ability to renegotiate pricing at renewal (especially if market rates drop). - Price Increase Caps: In multi-year deals, Nebius might include caps on price increases at renewal or commit to price decreases if underlying cost drop – not sure, but to compete with AWS's enterprise discounts, Nebius might offer more flexible terms (like no lock-in, maybe even a price drop guarantee that if Nebius lowers list price, the committed customer also benefits). - Cancellation: Pay-as-you-go users can leave anytime (no penalty). For reserved commitments, cancellation likely either not allowed or requires paying some penalty (like AWS RI – you pay regardless, or maybe Nebius could let a customer buy out a contract for a fee). Nebius might be more lenient in early days to not alienate – but they might simply not have had many cancellations yet since they're in growth. - Data ownership & Portability: Nebius's contracts presumably state the customer owns their data and models. They likely have clauses saying Nebius won't use customer data except to provide the service (common cloud clause), and on termination, the customer can extract data. They might not have a fancy data export service, but customers can download or transfer it off. - SLA & Remedies: Nebius likely provides an SLA around uptime (e.g., 99.9% monthly for core services) and maybe around support response for high severity issues. If not meeting, the remedy is service credits (like some % of monthly bill depending on downtime). - They might also have performance claims in marketing (like X% faster training), but those aren't contractual – just marketing. - Auto-renewal notice: Possibly in any reserved deals, if the customer doesn't say otherwise, it could renew capacity or convert to on-demand after term. (Because if a customer doesn't actively renegotiate, Nebius wouldn't want them to churn just because contract expired – they might put them on on-demand rates which are higher, ironically might prompt them to talk and re-commit). - Termination by Nebius: Standard cause for immediate termination might be if user violates acceptable use (like uses Nebius to mine crypto? Nebius might forbid that to save GPUs for AI since crypto is less desirable – just speculation; or if user does illegal content, etc.). Those would be in ToS. Nebius probably reserves right to suspend accounts for that behavior. - Notice Period: If a customer wants to terminate (like not renew), likely they need to inform maybe 30 days before contract end. - Data extraction: Nebius likely will keep data for a short period after termination (like 30 days grace to retrieve anything, then it's wiped). - IP and custom work: If Nebius does any custom engineering for a client (not typical, but maybe if they co-developed a solution), contract should clarify IP – likely the client owns

their data and code, Nebius retains IP of platform enhancements. - Liability limitation: They'll have standard cloud liability caps (like liability limited to the fees paid by customer in last X months, and no liability for indirect damages). - Indemnity: Possibly a clause that customer must indemnify Nebius if their usage infringes someone's rights (for example, if they host a model with unlicensed training data, Nebius doesn't want to be sued – they'll push that to user). Conversely, Nebius might indemnify customer for any IP infringement in Nebius's tech (if a third party claimed Nebius's platform violates something, Nebius might cover – but often cloud providers disclaim that). - Customer obligations: Typically just to pay on time and use service lawfully. Perhaps if they commit to reserved capacity, they have to provide prepayment or sign a PO. - Support Terms: Nebius's support is likely best-effort but they might not promise specific resolution times in contract (maybe in premium deals they do). - SLAs for AI Studio etc.: Possibly an overall availability SLA covers everything. If Nebius had specific ones (like throughput guarantee or latency guarantee, which is unlikely at this time), that would be in a special contract addendum. - Data locality: If needed, Nebius could contractually commit that data stays in specified region (some governments require that assurance beyond just architecture). Nebius being EU-based can more easily sign EU data processing agreements for GDPR, etc., which is a plus in contracts with EU companies.

This covers pricing and terms extensively.

Now we proceed to **Integration Ecosystem & Technical Architecture (Platform Level)** which might cover things we've partly addressed: supported systems (like integration with existing tools), core tech stack (we did), scalability, reliability, security (some covered but let's articulate systematically) – I'll consolidate to ensure we hit all points:

Integration Ecosystem & Technical Architecture (Platform Level)

System Compatibility & Integration Capabilities: - Nebius is designed to fit into existing tech stacks, so it supports standard integration points: - **APIs/SDKs:** Nebius provides RESTful APIs and likely language SDKs for common languages (Python, maybe Go, Java) to interact with its services (e.g., start VM, manage data). This allows integration with CI/CD tools (Jenkins, GitLab CI) to automate training or deployment tasks on Nebius. - **Infrastructure as Code:** As mentioned, Nebius supports **Terraform***[130]*. An example: if a company uses Terraform to manage their AWS and on-prem resources, they can add Nebius as another provider in the same config, enabling multi-cloud setups. - **Kubernetes & Containers:** Nebius's support for managed Kubernetes means it can integrate with any Kubernetes-compatible tool. For instance, one could use Argo CD or Helm charts targeting Nebius's K8s. If a company has a multi-cloud strategy with Kubernetes, they can deploy on Nebius's K8s similarly to any other cluster. - **Data Integration:** - For **data sources**, Nebius can connect to popular data systems. E.g., if a user has data in AWS S3, Nebius can access it over the internet (with costs on AWS side). If the data is on an on-prem NFS or HDFS cluster, Nebius VMs can connect via VPN. Nebius's Spark can pull from various data lakes (Spark connectors allow reading from S3, Azure Blob, JDBC, etc.). - Nebius integrating with **Backblaze B2** basically means Nebius's environment is allowed to directly read/write to B2 without egress charges or with optimized routing. Possibly Nebius uses B2 as its de-facto object store. They might even make B2 appear as a Nebius-branded storage

to user for simplicity. - For streaming data, Nebius doesn't have a Kafka service but one could run Kafka on Nebius or stream from external Kafka to Nebius via normal network. Nebius's networking can handle typical protocols. - **Third-party Integrations:** Nebius likely has or is developing partnerships for integration: - **Hugging Face**: Possibly Nebius integrated Hugging Face Hub (e.g., AI Studio could fetch models from HF). - **MLflow**: integrated out of the box. - **Jupyter notebooks**: Nebius might have a web Notebook interface or at least instructions to use Jupyter on Nebius instances. - **NVIDIA tools**: Nebius being Nvidia-aligned means tools like NGC (Nvidia GPU Cloud registry) are accessible. Nebius users can pull NGC containers for ML frameworks easily. Also, Nebius might integrate **NVIDIA GPU Monitoring tools** into its console (like exposing MIG partition config if they allow slicing GPUs). - **IDEs**: Nebius doesn't have its own IDE, but developers can connect VS Code remotely to Nebius instances via SSH (common workflow) – no friction there. - **Monitoring/Logging**: Nebius likely outputs logs to allow integration with systems like ELK/Elastic or Splunk if the customer sets it up. For instance, Nebius could allow sending Kubernetes logs to the customer's logging endpoint. - **Apex or others**: If any specialized HPC or ML tools, Nebius's approach is to be flexible platform rather than forcing use of specific tools. So compatibility is high as long as it can run on Linux on x86 with NVIDIA GPUs – which is most of the ecosystem. - **Alliances & Channel**: Nebius's partner program*[31]* suggests they encourage SIs or VARs to include Nebius in solutions. For example, an AI consulting firm might integrate Nebius as the recommended cloud backend for deploying the solution they build for a client. Nebius then offers them partner benefits like maybe referral fees or co-marketing. On a technical level, this means Nebius is making sure those integrators have the APIs and support to embed Nebius usage into their deliverables. - **Multi-cloud orchestration**: Nebius can be integrated into multi-cloud managers like HashiCorp Nomad or Red Hat OpenShift if desired (one could treat Nebius like any K8s cluster or run OpenShift on Nebius VMs). - **API Strategy & Developer Program:** - Nebius's API is likely comprehensive (covering all functionality), documented in their docs. They may have an API explorer or CLI to ease usage. The reliability of API is important (they likely version it to avoid breaking changes). - They probably have sample code on GitHub, a developer portal. - **SDKs**: Possibly Nebius provides a Python SDK because ML folk often use Python. And maybe a CLI under the hood uses an SDK. - **Community**: As Nebius grows, they might foster a community of devs, perhaps through forums or a Slack/Discord, hackathons etc. Being fairly new, they might rely on direct support rather than community, but that can evolve. - **Partner/Developer Program**: they have a startup program (for small dev teams) and might formalize a partner program for tech partners (like if a vendor wants to optimize their tool on Nebius, Nebius might give them credits or technical resources). - **API Policies**: They likely commit to API stability (if changes, they support older version for some time). - **Rate limits**: They probably have some safe limits on API calls (e.g., don't allow 1000 instance create requests per second to avoid abuse) but for normal usage, not a concern. If needed for scaling automation, they can raise limits case-by-case. - **API ecosystem**: Nebius could even integrate with infrastructure-as-code frameworks beyond Terraform (maybe Pulumi support if user base demands, etc.). - Developer support also includes samples and maybe integration with ML frameworks (like a plugin for PyTorch Lightning to use Nebius as a trainer backend – not sure if they have that, but potentially). - **Core Technology Stack (We have covered many elements)**: - Nebius uses a lot of open-source at its core: likely **OpenStack** or similar for VM management (Yandex Cloud was built partly on OpenStack for

IaaS), **Kubernetes** for container orchestration, **Ceph or Lustre** for storage, **Consul** or etcd for service discovery, etc. - The Nebius console and backend likely written in languages like Go or Python (Yandex had a lot of Go usage in cloud). - They run on their **data centers**: Finland DC (they built themselves), plus some co-location (Paris might be in a third-party DC but Nebius brings their racks; Kansas City cluster is in Patmos Data Center as per DCD article[54]; NJ 300MW likely a new build or partnership – perhaps with another provider). - Cloud hosting environment: They own and operate (not on AWS/Azure obviously). Possibly they use some **Open Source Cloud Management** bits from OpenStack (like Yandex cloud originally did for computing and storage). - CI/CD pipeline for Nebius's own development: they likely release features incrementally with careful testing (the platform cannot break – they probably mirror how other clouds do, with staged rollout). - DevOps: Nebius likely containerizes many of their services, uses Kubernetes to run Nebius internal microservices (the control plane might be microservices managing provisioning, billing, etc.). - The architecture may be **regionally independent** – each region might run its own control plane for things like scheduling, with a global layer for management. Being new, Nebius might have a combined control plane for now with region awareness, and will iterate to more distributed control planes as needed for scale/resilience. - **Scalability & Reliability**: - Nebius's architecture is built for horizontal scalability: - They can add more GPU nodes easily as demand grows (they actually did by tripling Finland capacity). - The control plane likely uses redundant components (multiple API servers, multiple schedulers, etc.). - Data stores for Nebius's state likely redundant across zones or cluster of nodes, possibly with failover across region for global control data (if one region's control goes down, others not affected, and maybe a backup of global config is stored in multiple places). - Uptime track record: As a new public company, Nebius would likely share if they have strong uptime (maybe boasting something like 99.95% since launch). They haven't had major known outages publicly (if they did it would be visible, but none reported so far). - **SLAs** likely at 99.9%. If they run 24/7 support, they can recover quickly from issues. - They have a **status page**[140] and likely incident response processes if something fails (like if InfiniBand network sees high error rate, they have monitoring to detect and switch traffic). - **Auto-scaling** Nebius's platform: Nebius might auto-scale their control plane too (like spin more API workers if API usage increases). - For customers, the Nebius environment supports their auto-scaling of workloads (like Kubernetes clusters auto-scaling nodes). - **Disaster Recovery**: Nebius presumably has DR plans especially for their own services. If one data center totally goes down (power outage or natural disaster in Finland), how do they recover? They did mention building multiple buildings in Finland for expansion, hopefully with some redundancy. If a whole region (like Europe's cluster) went down, Nebius could urge customers to use another region if possible (but they must have backup of control plane config somewhere safe). - For customer data, Nebius probably offers options to replicate data cross-region (though not sure if automated – the user might handle that by copying snapshots to another region's storage). - **Security**: - Nebius is likely pursuing or has **SOC 2 Type II, ISO 27001** etc. (given they deal with enterprise clients and are publicly traded). - Data encryption: They mention "enterprise-grade encryption at rest and in transit"[155]. Likely all storage volumes and object storage are encrypted at rest (maybe using LUKS or Ceph encryption) and all network traffic outside of physical cluster is TLS (and internal network could be partitioned, but InfiniBand traffic might be plain within cluster – but that's isolated network). - They have an **IAM** system to manage identity

and access: users can create API keys, roles, attach policies (like allow a service account to only access a certain bucket or spin VMs in one project etc.). Yandex Cloud had an IAM, Nebius probably inherited it. - Multi-tenancy isolation: They ensure that one tenant's VMs/containers can't access another's data. This is enforced by hypervisors for VMs (likely KVM or similar for virtualization) and by Kubernetes namespaces and network policies for containers. Nebius's hypervisor likely uses features like SR-IOV for NICs to isolate network, and virtualization for GPUs can be done via MIG or exclusive assign. - **Compliance**: If handling personal data, Nebius helps by offering to keep data in region (for GDPR). They might also be looking at **HIPAA** compliance for healthcare clients – which means signing BAAs and implementing required controls (their security posture seems capable, but not sure if they've done that yet). - Nebius might have or be working on **SOC 2** compliance (the mention of quarterly letters and IR site suggests a maturity to handle compliance). - **Penetration Testing**: Nebius likely does regular third-party security testing. They might allow customers to pen-test their own Nebius instances (most cloud providers allow it if you notify). - **Vulnerability management**: They must quickly patch any vulnerabilities in their stack (like if a new Xen or KVM exploit, update hosts; if a container escape CVE, patch container runtime, etc.). They likely have automated patch pipelines and maintenance windows for applying them (e.g., live migration of VMs off a host to patch host then move them back, so customer doesn't see downtime). - **Incident Response**: They presumably have a 24/7 on-call security and ops team to respond to incidents (with processes in place, possibly even tested). - **Data Governance & Compliance**: - Nebius likely organizes customer data by projects or accounts; they ensure no commingling beyond necessary (like if using multi-tenant storage, it's logically separated and encrypted). - They have a privacy policy and likely sign Data Processing Agreements for GDPR with customers, committing to only process data under customers' instructions and using appropriate protections. - For specific domains: if Nebius targets finance, maybe they consider FINRA or other compliance – no evidence yet. For healthcare, as said, possible HIPAA readiness if needed by a client (they might do that on demand). - The Nebius Trust Center might detail things like they adhere to GDPR, probably not transferring EU data out unless customer does it. - Nebius for now doesn't mention being in heavily regulated domain (like FedRAMP for US Gov – likely not yet, but if they ever target government contracts, that's a whole compliance domain to cover). - **Technical Debt**: - Nebius had the fortune to start fresh in some ways but building on Yandex Cloud code, which had matured ~4-5 years. If any technical debt from Yandex times: possibly some code written for different context (e.g., Yandex Cloud had to support many small VMs for web services; Nebius's focus is fewer but bigger VMs for AI – they likely had to tune scheduling for that scenario). - The scale Nebius is going for might push the limits of some Yandex Cloud components – e.g., scheduling thousands of GPUs with InfiniBand might have required new work (they did design new stuff like Soperator). - So technical debt might exist in older parts (like if some openstack modules used can't handle multi-terabit networks elegantly – just hypothetical). - Another possible technical debt is complexity from rapidly integrating acquisitions: Nebius also manages other services (Toloka, etc. under Nebius Group). But those are separate businesses, not integrated into cloud stack – so not relevant to cloud technical debt except group focus. - Rapid hiring can create some inconsistent internal tools if not managed – but that's manageable. - One sign: TrueTheta said product documentation was incomplete for advanced use cases*[144]*. That could hint at either lacking time to document (a

form of "documentation debt"), or features not fully polished yet. - If technical debt hinders future development, Nebius will need to address – maybe they had to invest in a new control plane for InfiniBand cluster management (if Yandex code assumed Ethernet, etc.). - As Nebius scales beyond original Yandex design goals, they must ensure the architecture scales (Yandex Cloud primarily served Eastern Europe with moderate scale; Nebius aims global hyper-scale). - So far, no public issues, meaning any tech debt is being managed or hasn't surfaced critically.

This covers integration and architecture comprehensively.

We have effectively concluded Section II as per prompt, albeit it's extremely lengthy.

Given that, the next main section in the prompt is **III. Ideal Customer Profile (ICP) & Persona Deep Dive** which requires ultra-deep persona analysis. We have touched on target segments (AI startups, etc.) but now must systematically define segments and then dive into 4-6 personas with great detail.

We should do that concisely yet thoroughly.

Let's break it down:

Ideal Customer Profile (ICP): - Nebius is focusing on organizations that have heavy AI/ML compute needs. Within that, likely segments: - **AI/ML Product Startups** (SaaS or tech startups building AI-driven products or features). Often Series A/B companies with 10-100 employees, often in sectors like generative AI (text, image), analytics, etc. They adopt Nebius for cost and performance. - **Enterprises doing AI projects** (like a bank's AI research team, or a pharmaceutical company's bioinformatics team). Possibly as a supplement to internal HPC or as cost-saver vs AWS. These tend to be mid to large companies (1k+ employees). - **Research institutions & Universities** (nonprofits or educational, working on machine learning or scientific computing). They might choose Nebius if they lack enough internal cluster or want multi-region collaboration. - **Media & Entertainment** (they cited that industry for storage-intensive HPC like rendering). A VFX studio or animation company that needs rendering and maybe training models for effects – Nebius could appeal with GPU and storage deals. - **Biotech & Life Sciences** (they listed biotech), which often need to train models for protein folding, genomics etc. That's HPC requiring GPUs often. Nebius can be cheaper than building their own cluster or using AWS which has GxP compliance overhead and cost. - **FinTech and InsurTech** (they mentioned these in targeting criteria*[156]* presumably for storage but also perhaps for risk modeling, which can be GPU-intensive). - **AI Service Providers** (like consultancies or smaller cloud providers who might resell Nebius capacity under their offerings – possibly). - **Gaming & Metaverse** (cited in criteria*[156]*): this is interesting – maybe for companies doing real-time GPU tasks, Nebius could provide backends (like a cloud gaming service or a metaverse world simulation training using Nebius's GPUs). - We need TAM, sweet spots, etc. Nebius likely views **TAM as huge** given AI is broad. But serviceable obtainable currently focuses on those who are comfortable using a new provider (so early adopters, often tech-savvy). - **Size & Scale Sweet Spot:** Because Nebius deals in high-end GPUs, tiny companies with no GPU workload aren't relevant. They focus on at least mid-size AI companies. - Possibly ideal size: - For startups:

5-50 ML engineers requiring 10-1000 GPUs frequently. - For enterprise: maybe a department or innovation lab with $500k+ budget for AI compute annually (so Nebius can save them e.g. $250k). - If a company only needs a few GPU hours occasionally, Nebius welcomes them but those aren't driving big revenue. Nebius's 700% ARR growth likely came from a handful of heavy users (maybe dozens of customers spending hundreds of thousands). - So the sweet spot is likely customers that will spend at least, say, $10k/month on cloud if on hyperscaler (which Nebius could make $5k from at 50% cost). - Nebius's initial wins likely included some mid-tier AI startups that outgrew Colab and needed cluster-scale (the ones who might have considered CoreWeave, etc., they captured). - Geographically, current focus is **Europe and North America**. They are based in Amsterdam, focusing EU, and they set up in US because US has many AI startups (and presumably to support global companies with US operations). - Within geos: EU adoption might be stronger due to Nebius's Europe presence. In US, they face competition from CoreWeave and inertia of AWS. But the US is huge market so even grabbing some yields big revenue. - They might prioritize UK, Germany, France in EU (since big tech budgets there), and tech hubs in US (Silicon Valley, New York for finance). - **Technographic Profile:** - Ideal clients already use some cloud (so they know how to consume cloud resources programmatically), or use on-prem HPC and are exploring cloud. - If on AWS: they might have Kubernetes or Terraform pipelines that can easily be adapted to Nebius. - They likely code in Python, use libraries like PyTorch, TensorFlow (Nebius fully supports those). - They might also use containerization (Dockers for consistent env). - Many will have data pipelines possibly on Snowflake or a data warehouse – Nebius can integrate by connecting to those via network. - **Digital maturity**: Most Nebius early clients are likely *innovators/early adopters* themselves in tech adoption. They are comfortable trying a newer vendor for the benefit given. They often favor open-source (hence love that Nebius uses open models and standard frameworks). - If a customer was completely locked in to a proprietary stack (like heavily using AWS SageMaker or Google's AutoML), they're less likely Nebius's early customer (too tied to that environment). Nebius's sweet spot is customers who want flexibility or are frustrated by limitations of such platforms. - **Critical Pains & Trigger Events:** - **Cost explosion**: A classic pain: "Our AWS bill for GPUs skyrocketed when we scaled our model training – we need to cut costs or we'll burn our funding too fast." This triggers looking at Nebius as a cost-saving. - **GPU shortage**: "We cannot get enough GPUs or have to wait in queue on our current cluster (or cloud region) for new GPUs." This triggers trying Nebius which promises abundant supply (like Nebius bragging 60k GPUs coming). - **Performance bottleneck**: "Our training runs are too slow on current infra." Possibly due to older GPUs or network issues on other cloud. Nebius's performance pitch (InfiniBand, etc.) appeals here as a way to speed up development cycle. - **Multi-cloud strategy**: A company might decide not to be dependent on one cloud for risk management or negotiating leverage. So they look at Nebius to avoid hyperscaler lock-in. - **New project/leadership**: e.g., a new Head of AI in a company might come in and say "let's re-evaluate our infra approach to be more cost-effective," thus considering Nebius. Or after a funding round, a startup might get funds to ramp up compute, and they evaluate best options (and Nebius can catch them at that planning stage). - **Regulatory environment**: e.g., a European company using US cloud might worry about data jurisdiction (Schrems II etc.), so they consider Nebius since it's EU-based for compliance comfort. - **M&A or partnership**: A partner might suggest Nebius (like Backblaze recommending Nebius to their customers who need compute). - **Rebelling against AWS**

egress: If a company is frustrated by paying large egress fees (like transferring training data out of cloud for partners), Nebius's no-egress approach could be a trigger to switch. - **Growth Ambitions vs. Digital Maturity**: - Many Nebius customers are pursuing aggressive *innovation and growth*, willing to adopt new tech to accelerate (so they lean risk-taking and early adopter). - They likely have moderate to high digital maturity (since they are doing advanced AI – not laggards by definition). - Some enterprises might be slower but the ones Nebius gets early are probably those with a strong innovation culture (like a bank's AI lab separate from core IT, able to try new cloud). - They primarily seek **efficiency (cost)** and **performance (speed)**, which both tie to revenue indirectly (faster to market, or for enterprises, reducing cost improves margin). - They also care about **risk mitigation** (like avoid being crippled by cloud costs or vendor risk). - Nebius's ideal customer is more concerned with building and scaling AI solution than with having a huge menu of managed services. That means they have some technical capability to handle pieces that Nebius doesn't provide (like building the application around the model). - **Negative ICP / Exclusion**: - Companies that have minimal AI needs (like web app companies that only might call a third-party AI API occasionally) – not a fit, likely wouldn't bother migrating anything to Nebius for small usage. - **Very large cloud spenders locked in multi-year contract with AWS** – they might not consider Nebius until contract up or if a piece of workload can peel off. If an enterprise is fully Azure shop due to enterprise license deals, Nebius might be a tough sell for anything beyond a skunkworks project, so those with heavy vendor lock-in culturally are near-term not focus. - **Highly regulated with specific compliance that Nebius doesn't have yet**: e.g., US government agencies requiring FedRAMP moderate or high – Nebius doesn't have that, so not a target now (plus might not trust non-US company for sensitive work). - **Extremely risk-averse IT departments**: If a company will only use "Big Four" cloud providers (AWS, Azure, GCP, IBM maybe) out of caution, Nebius is out. Some traditional enterprises require vendor stability histories; Nebius being new might face hurdles there. - **Tiny budgets**: if someone can't even spend a few thousand a month, Nebius's sales probably doesn't actively pursue them aside from self-service signups. Nebius is more interested in big fish or promising fish that grow. - **Tech stack misalignment**: If a team exclusively uses, say, Google TPUs for their model (some ML is built around that or JAX environment), Nebius doesn't offer TPUs or that ecosystem, so they wouldn't fit until Nebius maybe supports more. - Or if a workload is mostly CPU and reliant on some specialized service like AWS Athena (serverless SQL), Nebius can't replicate that. - Summing up negative fit: those unwilling to adapt tools and those with extremely low cost tolerance (though Nebius is cheaper, there's still a cost if someone expects free usage like on Colab). - Also, given Nebius's business model, they probably avoid any potential customers in Russia/Belarus under sanctions due to their origin separation (they won't engage with those to avoid sanction complexity).

Now we craft **Buyer & User Personas** for Nebius's context. They wanted 4-6 personas with extreme detail.

Potential key personas: 1. **CTO/Founder of an AI Startup** – decision-maker, technical, cares about cost/performance, tries Nebius for advantage. 2. **Machine Learning Engineer / Researcher** at a startup or enterprise – daily user who influences decision by advocating for better infra (they feel pain of waiting for compute). 3. **Head of Data Science or AI Lab in an Enterprise** – a managerial role who decides on platform and needs reliability and cost

justification for Nebius vs. default internal or AWS cluster. 4. **DevOps/Infrastructure Engineer** – responsible for setting up and managing infrastructure for ML teams; cares about ease of integration, automation, and reliability. Might initially be wary of new platform but convinced by technical merits. 5. **CIO/IT Procurement (Blocker persona)** in a larger enterprise – not intimately using Nebius, but concerned about security, compliance, and vendor viability. Could raise objections Nebius must overcome. 6. Possibly **Data Scientist (Individual Contributor)** – a user concerned with their personal productivity (ease of use of Nebius tools, not having to wait for jobs etc.) – somewhat overlapping ML engineer persona but from a less systems perspective and more algorithm perspective.

They wanted extremely deep, but we must balance length given already huge answer. We'll provide quite rich descriptions for maybe 4 personas given space/time, or 5 shorter. Let's attempt ~4, but if time allows 5.

The prompt specifically says 4-6 key decision-makers, influencers, champions, blockers, and primary users. We should cover: - A decision-maker (like CTO or Head of AI). - A champion user (like a senior ML engineer pushing for Nebius). - A blocker (IT/security or CFO concerned). - A user persona (data scientist focusing on usage experience). - Perhaps one for startup vs one for enterprise to show difference.

We'll do: **Persona 1: Startup CTO/Founder (Decision-maker)**, **Persona 2: ML Engineer (Champion)**, **Persona 3: IT Infrastructure/DevOps (Influencer/Blocker)** possibly in an enterprise context (could block or allow), **Persona 4: Data Scientist/Researcher (Primary user)**, **Persona 5: Enterprise AI Lab Director (Decision-maker in big co)**, **Persona 6: CFO/Procurement (maybe combined with IT for blocker)** – though CFO in a startup might also be the founder themselves making cost decisions, but in bigger org CFO cares if Nebius contract etc.

We likely can merge CFO concerns into the CIO/IT persona, or mention CFO under each where relevant. Perhaps a distinct CFO persona might be too much unless they specifically raise price contract issues – but CFO seldom deeply involved in vendor selection unless large contract, then they look at ROI metrics.

Let's proceed with writing these personas:

III. Ideal Customer Profile (ICP) & Persona Deep Dive

Ideal Customer Segments: Nebius's ideal customers are organizations driven by AI/ML workloads that demand high-performance compute and are seeking more cost-effective, flexible alternatives to traditional clouds. They span several segments within the AI domain: - **AI/ML Startups:** Small to mid-sized tech companies (say 10–100 employees) whose core product involves training or serving AI models (e.g. generative AI startups, analytics SaaS, computer vision platforms). These companies often have cloud bills in the tens of thousands per month and are feeling the pinch of hyperscaler pricing or GPU shortages. They value Nebius for its

lower cost (extending their runway) and ready availability of cutting-edge GPUs to accelerate development[119][59]. They typically have technical founders/CTOs who are early adopters and not beholden to legacy vendors. - **Enterprise AI Labs & Data Science Teams:** Departments within larger enterprises (financial services, healthcare, e-commerce, etc.) focusing on machine learning initiatives – e.g., a bank's fraud detection AI team, a pharmaceutical R&D analytics group, or an automotive company's autonomous driving unit. These groups often operate like internal startups, pushing innovation but facing bureaucracy with corporate IT. They may have outgrown on-premise GPU clusters or are frustrated with internal procurement delays. Nebius appeals to them by offering on-demand scalability without having to invest capital in infrastructure, and by complying with data residency needs in Europe for regulated industries[106]. Typically they have some budget autonomy and can pilot Nebius for specific projects (such as massive model training runs or burst compute for quarterly analyses). - **Research Institutions & Universities:** Academic labs and research institutes doing AI research or other HPC work (genomics, climate modeling, etc.). They have huge compute demands but constrained budgets, making Nebius's cost savings crucial. They also appreciate Nebius's support for open-source tools and ability to handle unusual workloads. For example, a university AI center might use Nebius to run large language model experiments that their campus cluster can't support. Nebius's startup/research credit programs[32] specifically target this segment, making it financially feasible for academia. These users often serve as technology thought leaders, amplifying Nebius's credibility if they achieve breakthroughs on the platform. - **Media & Entertainment Studios:** Companies in rendering, VFX, or animation, as well as emerging "metaverse" and gaming companies. Their workloads include GPU-heavy rendering and increasingly AI-driven content generation. They are attracted by Nebius's combination of *no-egress high-throughput storage + GPU compute* (e.g., moving terabytes of footage or simulation data without egress fees, processing with GPUs, storing results)[37]. A mid-sized animation studio (50–200 employees) could offload its rendering farm to Nebius, saving up to 50% and scaling for peak project demand without buying hardware. They care about reliable performance (meeting production deadlines) and cost predictability. - **Biotech & Life Sciences:** Genomics companies, bioinformatics startups, and pharma tech teams performing tasks like DNA sequencing analysis, protein folding simulations, or drug discovery using AI. These are computationally intense and often GPU-accelerated (e.g., using CUDA for certain algorithms or training ML models on biomedical data). They require large memory and sometimes specialized hardware. Nebius offers them powerful GPU instances with huge RAM (up to 1.7 TB on an 8×GPU node)[157][158] and the ability to run bespoke pipelines (perhaps via Nebius's managed Slurm for scientific batch jobs). They also appreciate that Nebius can keep sensitive data within desired jurisdictions (e.g., patient genomic data staying in EU data centers for GDPR compliance). - **FinTech/Insurance Analytics:** Companies in finance with AI-driven analytics (fraud detection, algorithmic trading models, risk simulations) and insurers doing large-scale risk modeling or AI underwriting. These firms handle sensitive data and have spiky compute needs (e.g., running massive risk simulations overnight or on end-of-quarter). Nebius, being an independent provider, can be appealing for multi-cloud risk management: a bank might not want to rely solely on one hyperscaler, and Nebius offers geographic diversity and potentially stronger contractual data controls (plus not being subject to CLOUD Act in the same way, since based in EU – a nuanced but sometimes considered point). They require strong security and often

require contractual assurances; Nebius's enterprise-friendly posture (investor-grade operations, willing to negotiate custom terms) helps here.

Market Sizing & Focus: The overall AI cloud market is enormous and growing (in 2025, estimated TAM for AI infrastructure cloud in the tens of billions globally). Nebius's **serviceable obtainable market (SOM)** in the near term focuses on early-adopter subsets: - In Europe, Nebius targets becoming the default AI cloud for the **"AI startup ecosystem"** – there are hundreds of funded AI startups in EU that collectively spend hundreds of millions on cloud. Nebius's penetration here might be measured in number of startups onboarded (they've likely captured dozens of high-profile ones already, given ARR growth). - In North America, Nebius's SOM includes the segment of US AI startups not yet locked into big cloud contracts (many will experiment with multiple clouds). Also some forward-thinking enterprise AI teams. - By industry, Nebius sees **Technology (AI/ML)** and **Research** as low-hanging fruit, then **Media, Finance, Bio/Healthcare** as next wave. For each, the TAM is large (e.g., global banking AI spend is in the billions; Nebius can carve out the portion related to AI compute). - **Optimal Client Size:** Nebius works for both *mid-market* (the startup or departmental case) and is beginning to serve *enterprise*. The optimal client from Nebius's perspective is one who will consume large GPU hours consistently. For example: - A Series-B AI startup with 30 engineers might spend $50k/month on Nebius – a great customer for Nebius. - A Fortune 500 company's innovation lab might do a $500k pilot over 3 months, then if successful, ramp to a longer-term $1–2M/year contract. - On the flip side, a tiny 3-person startup that only occasionally needs a single GPU for fine-tuning is not Nebius's focus for sales efforts (though they can self-serve). Similarly, a giant corporation that is extremely conservative might not onboard Nebius quickly – Nebius will target their forward-looking divisions first. - **Geography:** Nebius's current customer density is highest in **Europe (especially EU, UK, Israel)** and growing in **North America**. It is actively expanding into North American markets (opening the Kansas City cluster and planning New Jersey) to reduce latency and meet data locality needs there*[55]*. Asia-Pacific and Middle East are on the horizon; likely early interest from places like India (huge AI developer base) and UAE (government-backed AI initiatives) have been noted*[147]*, and Nebius is weighing new regions accordingly. They will choose geographies where [DOMAIN] – cloud AI – demand is strong and not fully satisfied by local players.

Technographic & Data Environment: Ideal Nebius customers typically already use: - **Popular ML Frameworks** (PyTorch, TensorFlow) – Nebius fully supports these with pre-configured environments. - **Containers/Kubernetes** – Many startups deploy training via containers or use K8s; Nebius's managed K8s is a familiar environment for them*[132]*. - **DevOps and Infrastructure as Code** – Terraform, CI/CD pipelines, etc. Ideal customers have these in place so integrating Nebius is smooth (just another provider). If a team is still clicking in a UI on another cloud but doing heavy AI, they can still use Nebius's UI, but those with automation get the most benefit quickly. - **Cloud or HPC familiarity** – Nebius's users are often already on AWS/GCP or on on-prem HPC clusters. They know their way around Linux, SSH, job schedulers, etc., which Nebius supports and mirrors (with Slurm, etc.). This is important: Nebius doesn't have to educate them on *how to run an AI workload* – they know that; Nebius just shows *how to do it more efficiently on Nebius*. - **Data storage** – Typically use object storage (S3, etc.) or large file systems. Nebius provides alternatives or integration (Backblaze B2 or Nebius's own

fast storage). Many ideal customers have **petabyte-scale datasets** (for training or simulations) – Nebius's no-egress policy for partner storage is extremely attractive here[103]. If a client's data is on AWS S3 and incurring huge egress to move around, that pain point can drive them to re-evaluate their storage strategy (maybe shifting data to Backblaze and compute to Nebius to avoid egress). - **Data & ML tooling** – They likely use Jupyter notebooks, MLflow (Nebius provides this managed[129]), experiment tracking, etc. They may use specific MLOps platforms (Weights & Biases, etc.). Nebius's openness means it can work with any of these (e.g., users can pip install their favorite tools on Nebius instances). - **Digital Maturity:** Ideal customers are usually at least at a "data-driven" stage, if not "AI-driven." Startups by nature are tech-centric. Enterprise teams Nebius targets are often separate from legacy IT – they don't require hand-holding to adopt cloud; in fact, they might be the ones pushing cloud adoption internally. Many are early adopters in tech adoption curve (willing to try Nebius while it's relatively new because they perceive the big advantage). Their risk tolerance is higher – which aligns with Nebius's current profile (innovative but new entrant).

Critical Pain Points & Trigger Events: (specific to cloud AI infrastructure) - **"Our cloud costs are unsustainable."** This is a top pain. For example, *"We're spending $100k/month on AWS for GPUs and it's burning our cash too fast"*. The CTO sees cost rising faster than revenue[159], causing alarm. This pain is quantifiable (e.g., cost per experiment, cost per model trained) and Nebius directly addresses it by slashing those costs by 50% or more[110]. - **"We can't get enough GPUs when we need them."** A head of ML might wait in queue for cloud GPU capacity or find certain regions sold out of the newest GPUs. Or an on-prem cluster has a long job queue (a model that should train in 3 days takes 2 weeks due to waiting). Trigger: an important project is delayed or a model couldn't be retrained in time for a release. Nebius solves this by offering *elastic capacity* – e.g., spin up 500 GPUs for a day for a big experiment, a scale impossible internally. One data science lead lamented, *"We had to down-scope our model due to limited GPUs on our current setup – we're falling behind competitors."* That frustration triggers exploring Nebius, which advertises large-scale clusters on-demand[111][160]. - **"Training is too slow; our iteration cycle is crippled."** In AI, speed to iterate is key. Pain: experiments that take days or weeks to run. This could be due to suboptimal infrastructure (older GPUs, no InfiniBand, etc.). A common internal quote: *"By the time our model finishes training, the market has moved or our data has changed – it's unacceptable."* Nebius can cut training times significantly (reports of >2x speed-up vs previous cloud in multi-node training[59]) by providing high-performance hardware and networking. This pain often surfaces when teams are trying to build ever-larger models or keep up with research and find their current infra scaling poorly. - **"Our IT/procurement is hindering progress."** This is more enterprise-specific: e.g., a financial firm's AI team might wait 6 months for procurement to add servers to their data center, or legal/security may not approve using a new cloud easily. The pain is bureaucratic delay when the competitive environment demands agility. A trigger event: a competitor publishes an AI breakthrough, and the team feels they need to double their compute ASAP to catch up, but internal processes say "maybe next budget cycle." Nebius (and its sales team) can help champion a *pilot* as a way around this – often presenting it as *"Let's try Nebius for 3 months to see results"* which can be approved as an experiment or under discretionary budget. The success of that pilot can override blockers later. - **"Vendor lock-in/concern about**

single-supplier." Some savvy CTOs and CIOs are uneasy being 100% on one hyperscaler (fear of future price hikes, or outages – e.g., an AWS region outage halting all AI training). A trigger might be a notable cloud outage (e.g., AWS having a multi-hour downtime that stalled experiments – this reminds them of concentration risk). Or a trigger could be a strategic mandate to implement multi-cloud for resilience. Nebius can become the secondary (or even primary) cloud for AI workloads, offering diversity. Additionally, being independent (not a direct competitor in other areas) Nebius is sometimes seen as more of a partner than hyperscalers who might compete with the client's business lines (for instance, some companies fear giving money to big tech that also competes in AI products). This is a softer trigger, but it shows up in executive thinking: *"We don't want to be beholden to Big Tech for our AI capability."* - **"Need to scale globally/new market or project launch."** For example, an AI service that was Europe-only now has users in the US and Asia – latency and region-specific compliance become issues. A trigger event: signing a big client in a new geography requiring data to be processed locally (e.g., a EU-based startup lands a U.S. Fortune 500 customer who asks if they can keep data in the U.S.). Nebius's expansion to multiple regions means they can accommodate this by spinning up in Nebius's US region for that client. This pain is about *scalability and compliance* when expanding. - **"Hardware obsolescence or capex avoidance."** This is more for those with on-prem clusters: their GPU farm is aging (maybe filled with NVIDIA V100s) and they consider upgrading to A100/H100. The pain: huge capital expense and time to deploy new hardware, plus uncertainty of utilization. A CFO or engineering director might trigger exploring cloud alternatives instead of a $5M purchase. Nebius can swoop in with a demo: *"Instead of buying H100s, rent them on Nebius – instant access, and pay only when used."* The economic calculation (capex vs opex) often favors Nebius if utilization is not 100% (which it rarely is in research, often clusters sit idle off-peak). - **"Leadership mandate for AI acceleration."** Sometimes the trigger is top-down: a CEO declares "we need to double down on AI this year." The pain then for tech teams is how to deliver results fast. This can open budget and willingness to try new platforms. Nebius might get a call because the org is in a hurry to show progress (e.g., incorporate GPT-like features within 6 months) and they know acquiring infra traditionally would slow them. Nebius's value proposition of *speed (no queue, ready-to-go)* aligns well here.

Growth Ambitions & Digital Maturity: Nebius's ideal customers are generally *growth-oriented and tech-forward*. Their primary motivations for AI projects can vary: - **Revenue Growth/Ambition:** Startups obviously are trying to build products to gain market share quickly – they see faster model training and deployment as directly enabling them to add features that attract users (e.g., a generative AI SaaS rolling out new model improvements weekly to outpace competitors). An enterprise's AI initiative might aim to create new revenue lines (e.g., a retailer building a recommendation engine to increase sales per customer). These customers value Nebius's ability to shorten development cycles and scale out experiments – more shots on goal in a given time frame increases chance of a winning innovation. - **Cost Efficiency/Profitability:** Particularly enterprises (and startups as they mature) also have a cost optimization motive. Digital maturity brings FinOps awareness – they measure things like cost per training run, or infrastructure cost as % of overall R&D spend. Nebius's offering appeals to those looking to improve these metrics, often at leadership insistence. For example, an AI director might have a

KPI to reduce cloud cost per model trained by 30% YoY; Nebius becomes a key lever to achieve that in a single stroke by switching providers[110]. - **Innovation & Competitive Edge:** Many ideal customers are motivated by being on the cutting edge. They want the latest NVIDIA GPUs (H100/H200) not just for bragging rights, but because their research or model complexity demands it. Nebius's ability to offer H100s when some competitors might still be stuck on older GPUs is a draw[145]. Culturally, these customers often see themselves as disruptors in their field, and choosing an innovative cloud like Nebius aligns with that identity (versus sticking with a legacy choice). - **Risk Mitigation:** Some have as an objective the mitigation of certain risks – e.g., ensuring service continuity or compliance. A fintech startup might be pre-empting investor questions about dependency on a single cloud – adopting Nebius can be a story of prudent multi-cloud strategy. An enterprise might use Nebius in a pilot to demonstrate to regulators that they have tested alternative infrastructures (some financial regulators encourage not being tied to one vendor for critical services). - **Combined Goals:** Often it's a mix – for instance, *an AI startup CEO's mandate:* "We need to cut cloud costs by 50% *and* train a model 2x bigger to stay ahead of competition, *and* do it by Q4." This perfectly sets up Nebius as the solution because it addresses cost, performance (bigger model feasible due to more GPUs), and speed (done this quarter instead of next year thanks to readily available capacity).

Negative ICP – Who is NOT a Good Fit: - **Non-AI Workloads / Traditional IT:** If a company's needs are mostly general enterprise IT (ERP systems, simple web hosting) without heavy AI or HPC, Nebius is overkill or not tailored for that – those workloads are better on general clouds or on-prem. Nebius specifically does not compete to host your corporate email server or static website – it's built for [DOMAIN] heavy compute. So, traditional SMBs with no data science team are not targets. - **Organizations with Extremely Low Budget or Scale:** For example, a tiny startup in the idea stage that just uses free tiers or a single GPU occasionally – Nebius could technically serve them, but if they cannot meaningfully pay, Nebius sales won't focus there. Similarly, hobbyists or students are not who Nebius is marketing to (though they could benefit from credits if in research). - **Rigid One-Cloud Policies:** Some enterprises have strict policies to use a specific cloud (e.g., "We are an Azure-only shop due to a corporate deal"). These are poor near-term prospects – Nebius would waste time trying to break a global contract. Until that policy changes (which could happen if, say, Azure can't meet a need), Nebius will not be seriously considered. - **Highly Regulated Data with Unmet Compliance**: If a prospective customer requires compliance certifications Nebius lacks, they will exclude Nebius. For instance, a U.S. healthcare company that *requires* HIPAA-certified infrastructure and won't accept anything less – Nebius would not qualify until it attains those certifications or signs the requisite agreements. (Nebius might be willing to sign a BAA for HIPAA on a case-by-case basis, but if a prospect's security team simply says "not on the approved vendor list," it's a blocker.) - **Ultra-Conservative IT Culture:** Organizations whose culture is "no one gets fired for buying AWS" and are very risk-averse tend to avoid new providers. If their perception is that Nebius is unproven, they won't champion it unless there is an overwhelming pain forcing them. For example, a government agency might default to big incumbents for any cloud needs; Nebius won't be on their radar in 2025. As Nebius grows and proves itself, this may change, but currently such customers are not ideal to pursue. - **Feature-Dependent on Big Cloud Services:** Some companies heavily use proprietary services (like Google's BigQuery, AWS

SageMaker or Azure's cognitive services). If their workflow is deeply entwined with those managed services, moving to Nebius would require significant re-engineering. Unless they have a strong impetus, they won't switch. (For example, if a team has all data in BigQuery and uses AutoML in GCP, Nebius would require moving data out and doing more manual ML – a tough sell unless cost or performance justifies it). Nebius is best for those using mostly open frameworks and needing raw compute power, not those reliant on a high-level managed ecosystem of a hyperscaler. - **Tiny Scale Short-term Projects:** If a team just needs a handful of GPU hours as a one-off (like a consulting firm running a one-week PoC), while Nebius could serve that, they might just use whatever is handy (maybe AWS) instead of onboarding a new vendor. The overhead might not seem worth it for extremely short engagements. Nebius wants customers who will have ongoing needs and growth, not one-and-done jobs (although Nebius certainly welcomes any usage, the sales focus is on longer-term potential). - **Culturally Misaligned Orgs:** e.g., an organization that demands extremely high-touch on-prem-style support or equipment leasing (some conservative industries want vendors to do everything for them in a traditional way). Nebius, being a self-service cloud with excellent support but not an on-prem vendor, might not fit those who essentially want a managed on-prem solution. (Though Nebius could potentially do bespoke on-prem deals in the future, that's not their current model). - In short, **Nebius thrives with customers who are innovative, somewhat self-sufficient in technology, feeling acute pain in current solutions, and open to new approaches.** Those that are happy with status quo, or cannot leverage Nebius's unique strengths, are not a focus.

Now, having defined the ICP, we delve into **detailed personas** that make buying or usage decisions in these organizations. We'll develop profiles of key roles: decision-makers (like CTOs or Heads of AI), influencers/champions (senior ML engineers, architects), primary users (data scientists), and blockers (IT/security or finance). Each persona will be richly described:

Persona 1: "The Visionary CTO at an AI Startup"

Role & Background: Meet *Alex*, the Chief Technology Officer and co-founder of a Series A AI SaaS startup (around 25 employees). Alex is a hands-on technologist (mid-30s, with a PhD in machine learning) responsible for the company's tech strategy and infrastructure. The startup offers an AI-driven analytics platform for e-commerce, and their product's competitive edge relies on training custom ML models on client data. Alex oversees a small engineering/ML team and makes final calls on tech stack decisions. Prior to founding this startup, Alex worked as a senior ML engineer at a larger tech company, so has experience with cloud infrastructure (mostly AWS) and model development. Alex thrives on innovation and is always looking for tools that give their startup an edge over bigger competitors.

Goals & KPIs: Alex's top goal is to **accelerate product development and innovation** without blowing through their VC funding. Key KPIs Alex is accountable for include: - **Model Improvement Velocity:** How quickly the team can iterate and improve model accuracy. This might be measured by how many experiments or model versions can be tested per month. Faster training = more experiments. Alex wants to double the number of experiments the team runs each quarter (a metric internally tracked). - **Infrastructure Cost Efficiency:** Keeping cloud spend under control. The board scrutinizes the runway, and cloud compute is the startup's

single largest expense after salaries. Alex has a target to reduce the cost per model training by, say, 30% this quarter. Also an unofficial KPI: keep monthly cloud costs within budget (e.g., <$30k/month right now). If costs grow, Alex needs a story for why (e.g., “we onboarded X new customers” or “we trained a model that improves our accuracy by 5%” to justify it). - **Reliability & Delivery:** Ensure the platform (which includes AI inference for customers) meets SLAs. If their service has downtime or lags due to infrastructure issues, Alex is responsible. So while innovating, Alex cannot compromise production reliability. Uptime and performance metrics for the SaaS (e.g., API response time under 200ms, 99.9% uptime) are KPIs indirectly tied to infra decisions. - **Team Productivity:** Alex also gauges success by how productive the ML and engineering team are. If engineers are spending too much time fiddling with infrastructure (setting up servers, waiting for resources, managing ops), that’s a problem. A qualitative goal is to **minimize devops friction** – the team should focus on model development, not infrastructure firefighting. Time-to-launch for new features is a metric here.

Pains & Frustrations: - *“Our AWS bills are killing us.”* Alex has been frustrated each month seeing AWS charges climb unexpectedly. A specific pain: last quarter, cloud costs jumped 25% while user growth was only 5%*[159]*, squeezing margins. Alex feels that **paying premium prices to AWS for GPU instances that are often underutilized or spend time waiting for data** is wasteful. This causes stress because it shortens the runway or forces Alex to consider raising more capital sooner – something the CTO wants to avoid by optimizing spend now. - *Slow Experiment Cycles:* Currently, training their latest recommendation model on AWS with 4 GPUs takes 3 days. If the team wants to try 5 different hyperparameter combinations, that’s over 2 weeks – too slow. Alex is frustrated that *competitors* (with more resources) might be out-experimenting them simply because they can throw more GPUs at the problem. The pain is a feeling of handicap: *“We have great ideas to improve our model, but we’re bottlenecked by compute – we can’t test them fast enough.”* This was highlighted last month when an important model improvement missed the product release because training wasn’t finished in time. Alex knows faster training would directly speed up feature delivery. - *GPU Availability & Quota Issues:* On AWS, Alex’s startup sometimes hit GPU instance quotas or found the specific new GPU type (like A100 instances) unavailable in their preferred region. They lost half a day filing a support ticket to raise the quota. This bureaucracy irks Alex – a startup needs agility, but big cloud providers make them “wait in line” or justify their usage. *“We’re a small fish to them – but those delays hurt us big-time,”* Alex remarks. The thought of Nebius offering thousands of GPUs with no long wait is very enticing to alleviate this pain*[7]*. - *Vendor Lock-in & Technical Constraints:* Alex is forward-looking and a bit uneasy that their stack was getting too tied to AWS (they use S3, AWS Batch, etc.). It’s a mild frustration that AWS’s ecosystem can be sticky – he’s read stories about exorbitant egress fees and difficult migrations. In meetings, Alex has said: *“I don't want to wake up two years from now stuck with whatever price AWS wants to charge or limited by their roadmap.”* This is both a strategic worry and, philosophically, Alex prefers openness (having come from academic ML, where open-source is valued). The lack of flexibility and fear of being **“at AWS’s mercy”** is a pain point – albeit one partly self-inflicted because AWS was the easiest path initially for the startup. - *Operational Overhead & Distraction:* Another frustration: setting up a complex multi-GPU training on AWS (with their previous approach using spot instances, etc.) took significant DevOps effort. Because their team

is small, Alex often personally had to tinker with Terraform scripts or AWS settings to optimize costs (like use spot, auto-shutdown instances, etc.). That's time Alex would rather spend on core product or architecture. Every hour spent babysitting infrastructure feels like an hour lost. Once, an overnight training job got interrupted because an AWS spot instance was reclaimed; it delayed results by a day and required re-launching manually. These kinds of hassles are painful for Alex. A dream is infrastructure that *"just works"* for heavy compute – minimal babysitting. - *Security/Compliance Anxiety (minor for startup):* While not as heavily regulated as an enterprise, Alex does worry about data compliance for their few European customers – under GDPR they should ensure customer data stays in permitted regions. On AWS, making sure data is only in Frankfurt region etc. was manageable. But as they eye expansion to new geos, Alex is concerned about juggling multiple cloud regions or providers. Nebius's clear regional isolation (e.g., EU versus US) could ease that, but currently it's a niggling worry that as they scale, compliance might become a headache if they stick with one-size-fits-all infrastructure.

Motivations & Aspirations: - *Innovation & Being Cutting-Edge:* Alex is intrinsically motivated by using the best tools to achieve superior outcomes. They take pride in the technical prowess of their startup. Adopting Nebius, with its state-of-the-art NVIDIA hardware and novel approach, appeals to Alex's desire to be on the leading edge. It's exciting: *"We'd be one of the first using this next-gen AI supercloud – that's a competitive advantage."* There's a bit of **status and thrill** in that too; Alex would enjoy being able to talk at industry events about how their startup trained models on a 60,000-GPU European supercomputer[11]. It reinforces their identity as a visionary tech leader unafraid of new solutions. - *Control & Ownership:* As a founder, Alex is highly invested in the company's success and wants control over its destiny. Being locked to one vendor or constrained by resource limits feels like loss of control. Nebius represents taking control back – optimizing costs (thus controlling burn rate), customizing infrastructure to their needs (like getting exactly the cluster config they want), and not being just another customer among millions. The fact that Nebius will likely treat them more attentively (being a smaller provider) also plays to Alex's preference for having influence. Alex aspires to build an infrastructure that is *tailored* to their workload, and Nebius's flexibility (like custom cluster setups, direct line to solution architects) aligns with that. - *Achievement & Hitting Milestones:* On a personal and company level, Alex is motivated by hitting big technical milestones – training the largest model in their sector, achieving a state-of-art result, launching a major product update on time. If Nebius can help achieve those, that's a huge win. For example, Alex dreams that *"with Nebius, we could train a model 2x bigger than we ever could on AWS, and get it to market first."* That would be a professional triumph and also validate Alex's decision-making prowess. - *Financial Prudence & Impressing Stakeholders:* Alex also aspires to demonstrate to the board and investors that they are a good steward of capital. By slashing cloud costs through a novel solution like Nebius, Alex can show tangible savings and extend the runway, which earns trust and perhaps frees budget for other things (like hiring more engineers). There's a motivation of *"doing more with less,"* a hallmark of great startup CTOs. If Nebius helps reduce spend while accelerating output, Alex stands out as an exceptionally effective CTO. This is both rational (company survives longer) and emotional (pride in being savvy and resourceful). - *Team Empowerment & Morale:* Alex cares about the engineering team's morale. Frustrated engineers who wait on slow infra can become demotivated or, worse, leave. Alex is motivated to provide

an environment where the team feels they have **world-class tools** and can do their best work. Adopting Nebius could energize the team: it signals that leadership is willing to invest in cutting-edge tech for them, which is motivating. Alex aspires to be the kind of CTO whose team says, *"We always get to play with the best technologies under Alex's leadership, which lets us move fast and learn."* That reputation helps retain and attract talent – another motivator. - *Personal Growth & Thought Leadership:* As a founder-CTO, Alex also has an eye on personal brand in the tech community. Successfully leveraging Nebius (especially if it's novel) could position Alex as a thought leader in efficient AI infrastructure. Perhaps Alex imagines writing a blog or speaking on a panel about how they achieved X on Nebius with limited resources – that kind of peer recognition is fulfilling. It's not the main goal (the company's success is), but it's a nice side aspiration.

In summary, Alex is driven to **make bold technical decisions** that yield faster innovation and cost advantage – Nebius fits right into that ethos. Alex wants their startup to punch above its weight in AI, and Nebius is a means to do so.

A Day in the Life (Week in the Life) of Alex:
Alex's days are a whirlwind of technical strategy and problem-solving: - **Morning (8-9am):** Alex scans overnight experiment results. For instance, the team might have had a model training on AWS that finished at 3am. This morning, Alex sees that it took 36 hours – longer than hoped – and model accuracy improved only marginally. Over coffee, Alex feels that familiar frustration: "If we could have run 3 experiments in that time instead of 1, we might have found a better variant." This thought is percolating while checking emails. - **Mid-morning (10am):** Daily stand-up with the ML and engineering team. One engineer mentions they need to retrain a model with a different dataset slice, but it will tie up the GPUs for two days. Another mentions our AWS credits (from a promotion) are running out next month, which will triple costs if usage stays the same. You can see Alex frown at that news. Alex and team discuss possibly using lower-cost spot instances or optimizing code to cut training time. It's clear to Alex these are stop-gaps – the team is spending valuable time on cloud cost gymnastics rather than building features. Alex encourages the team, *"Let's list out all options. Maybe we can spin up in another region or reduce precision."* But inside, Alex is thinking: *we need a bigger solution here*. - **Late morning (11am):** A quick call with the CEO and COO to prepare for an investor update. The CEO is concerned about burn rate; the COO notes that cloud spend is projected to be 15% over budget this quarter due to increased ML testing. The CEO asks Alex point-blank: *"Can we get these costs down or find another provider?"* This adds urgency. Alex replies, *"I have a few ideas – including a new provider that could significantly cut our GPU costs. Let me evaluate it thoroughly and get back to you."* The CEO green-lights investigating it, because cost-saving plus faster R&D sounds appealing, but warns: *"Ensure it doesn't risk our delivery timelines or data security."* Alex takes that as a mandate to seriously assess Nebius (which Alex had heard about recently from an engineer's Reddit post discussing Nebius vs CoreWeave). - **Afternoon (1pm):** Alex carves out time to dig into Nebius research. Pulls up Nebius's website and reads about the 80% cost savings claim[*[103]*], and the ability to run thousands of GPUs. Alex joins Nebius's Slack community (if available) or forum, sees some positive chatter from other startups about great support and performance. Alex drafts a quick analysis: *If Nebius costs ~$2/hr for H100 and AWS p4d (8xA100) effectively costs ~$4/hr per 80GB GPU, that's 50% savings. For*

our usage of ~2000 GPU-hours/month, we'd save around $4k/month. That's significant. Alex also notes Nebius offers InfiniBand – a potential performance boost. The more Alex reads (case studies, e.g. one where a company cut training time by 30% using Nebius*[59]*), the more excited he gets. - **Mid-afternoon (3pm):** Meeting with the lead ML engineer, *Bella*, who is the team's performance guru. Alex shares the Nebius idea: *"Bella, I'm seriously considering trialing Nebius cloud for our next big training. I hear we could get H100s with InfiniBand – might cut our training time a lot."* Bella is intrigued (she's heard of Nebius's reputation). They whiteboard what it would take to try it: migrating data (maybe use Backblaze for easier transfer), setting up the environment (Bella is happy that Nebius supports MLflow, which they already use*[129]*). They identify a candidate model training next week as the pilot on Nebius. Bella's only worry: *"Is Nebius stable? It's not as known as AWS."* Alex addresses this: Nebius has NVIDIA's backing and even positive press about being a top AI cloud*[14]*. Alex also points out they will run both AWS and Nebius in parallel initially as a safety net. This planning session leaves them both a bit energized – it feels like problem-solving that could relieve their pains. - **Late Afternoon (5pm):** Alex revisits tasks: reviewing a contract for a new hire, checking on a production issue (one of their inference endpoints on AWS had a spike in latency – possibly noisy neighbor issues on AWS VM). Alex notes ironically that a specialized platform like Nebius might avoid such multi-tenancy issues by design (InfiniBand for internal comms, etc.). Before ending the workday, Alex emails Nebius sales/support with a couple of detailed questions (SLAs, data security measures like encryption) to ensure due diligence. Almost surprisingly, a Nebius solutions architect replies within an hour with thorough answers and even offers to set up a Zoom to walk through how to migrate (this responsiveness already contrasts with AWS support tickets that take a day). Alex is impressed – *"They're treating us like a big customer even though we're small. This could be a true partnership."* - **Evening (8pm):** At home, Alex is catching up on tech Twitter and forums. Sees a tweet about a competitor raising a huge round and claiming they have a massive GPU cluster to train ever-larger models. Instead of feeling despair, Alex feels a bit of optimism because if Nebius works out, tomorrow their startup could punch in that weight class compute-wise. Alex messages the CEO on Slack: *"I'm looking into an AI cloud solution that might drastically cut our costs and also allow us to train bigger models – will update tomorrow, but initial signs are promising."* The CEO reacts with a thumbs-up emoji and "excited to hear more."

Throughout the week, Alex's interactions revolve around balancing immediate product needs with strategic improvements. The pain of cloud cost and slow experiments is recurring daily, and Nebius is now the leading potential solution in Alex's mind. Alex is preparing a proposal to present to the team and CEO: perhaps a one-month Nebius trial with measurable goals (e.g., "reduce training time by 30%, reduce cost by 50% on X workload"). Alex's day-in-life shows constant context-switching – from technical problem to managerial discussion to strategy – and in each context, Nebius addresses something: technical (InfiniBand speed), managerial (cost savings), strategic (outpacing rivals).

Watering Holes & Info Sources: - Alex stays current on technology primarily through developer communities: frequenting **Reddit (r/MachineLearning, r/aws, r/devops)** and Hacker News. In fact, Nebius first came to Alex's attention via a Hacker News thread discussing new AI infrastructure options, where someone mentioned Nebius's performance*[96]*. That piqued Alex's

curiosity to research more. - Alex also reads **tech blogs and press**: like TechCrunch, IEEE Spectrum, and Medium posts by thought leaders. The DCD (DataCenterDynamics) article on Nebius's origin*[161][10]* was something Alex skimmed earlier – now revisiting it to glean insights on Nebius's stability and backing. - Conferences & talks are big for Alex. He attends **NVIDIA GTC** online sessions; interestingly, maybe at GTC he heard Nebius was deploying H200 GPUs in Europe*[5]*. That gave credibility. Alex also networks in a Slack group for startup CTOs, where cost-saving tips are exchanged. Recently another CTO posted *"We moved some training off AWS to Nebius and saw big savings"* – peer validation like that strongly influences Alex. - For cloud-specific queries, Alex uses **Stack Overflow** and vendor documentation. E.g., in assessing Nebius, Alex downloaded Nebius's tech docs (checking API compatibility, Terraform provider availability, etc.). The ease or difficulty found there will influence Alex's confidence. - **Consultation with the Team:** Bella (lead ML engineer) and others are internal "watering holes" – Alex trusts their opinions deeply. If Bella was skeptical of Nebius's technical merit, Alex might reconsider. But Bella is actually enthusiastic to try Nebius (she loves new tech), giving Alex further confidence. - Vendor conversations: Typically Alex avoids pushy vendor sales, but Nebius's solution architect felt more like a helpful engineer. That conversation itself became a source of information – learning about Nebius's roadmap and support approach. Alex actually appreciates that Nebius folk can talk nerdy (InfiniBand metrics, etc.) rather than just sales fluff. - Summarily, Alex's information diet is highly technical and peer-driven. Nebius has entered that sphere via community chatter and credible articles, which is exactly how to win someone like Alex.

Tech Adoption & Risk Profile: - Alex is an **Innovator/Early Adopter**. Starting a company itself shows Alex's risk tolerance. Specifically with tech, Alex was among the first in prior jobs to try new frameworks (for instance, adopting PyTorch in its early days, or using Kubernetes when it was new). This trend continues: Alex is willing to pilot Nebius even though it's not the industry default yet, because the potential payoff is high. Of course, Alex manages risk by doing a phased trial rather than a full switch overnight. But relative to most CTOs, Alex is comfortable being on the cutting edge and even evangelizing it if it works. - That said, Alex does perform due diligence on critical matters. He will ensure Nebius meets security needs (encryption, compliance for GDPR, etc.) before committing production data. He won't gamble the company's core assets recklessly. It's a controlled risk-taking: "Let's try this new thing with a fallback plan." - Alex also knows when to **sell the risk-taking upstream** – e.g., framing Nebius to the CEO not as a crazy experiment but as a smart cost-saving measure with high upside and manageable risk (the CEO sees Alex's plan as fairly prudent given the context). - If Nebius shows clear benefits in the pilot, Alex will push for aggressive adoption and may become a champion of Nebius, possibly outpacing some team members' comfort. But Alex will address concerns with evidence (metrics from the pilot). This persona is the type to **influence others to adopt new tech** by showing it in action – a positive force for Nebius's word-of-mouth growth if satisfied.

Content Consumption & Communication: - Alex prefers **short, information-dense content**. For example, a well-written case study or a comparison table will grab Alex more than a generic marketing video. (He actually really liked Nebius's tech blog post where they published some performance benchmarks – hypothetically Nebius might have one – because it spoke to his engineer's mind with real data). - He also values **community discussions**: seeing Q&A on

Nebius's Slack or hearing a rep answer tough questions on a webinar. Alex watched part of a webinar Nebius did with SpringDB/Backblaze about multi-cloud AI infrastructure[37] – it resonated because it talked about escaping hyperscaler traps (exactly Alex's pain). - For communication style: Alex is straightforward and data-driven. When Nebius's architect gave numbers like "105.1M revenue, 625% YoY growth – we're scaling fast"[64], that gave Alex confidence Nebius isn't fly-by-night. Alex noted those details mentally to mention to the COO (to reassure that Nebius is financially robust / credible). - Internally, Alex communicates Nebius's value to the team via logic and enthusiasm: "Look, if this works, we get results faster *and* save money – that's huge for us." Alex addresses any pushback (like "what if Nebius goes down?") with practical mitigation (multi-cloud fallback, Nebius's SLA). - If pilot succeeds, Alex will likely present at an all-hands or engineering meeting: *"Kudos team – we tried Nebius on that last model and it trained in 18 hours vs 36, at half the cost. We're going to integrate it into our workflow."* This would boost team morale, painting the adoption as a win for everyone, not just a cost cut. - Alex is articulate and not shy to share experiences externally too (maybe a Medium post on the startup's engineering blog about their multi-cloud journey after a few months, giving Nebius a positive shout-out). That would be a win for Nebius's marketing indirectly – Alex becomes an advocate if satisfied.

Decision Criteria for Choosing Nebius: When Alex formally evaluates Nebius vs alternatives (like sticking with AWS, or trying another like CoreWeave or Oracle Cloud), he uses several criteria: - **Cost vs. Performance:** The prime criterion is the **cost-per-training and cost-per-inference** improvement. Alex is essentially computing ROI: if Nebius reduces training cost by 50% and training time by 30%, that is extremely compelling. So far Nebius is checking that box from initial tests. If Nebius had been only a tiny bit cheaper, Alex might not bother migrating – but the promise of 50%+ savings[110] is a game-changer. - **Ease of Integration:** How much work to get up and running on Nebius? Alex assesses that Terraform scripts can be adapted (Nebius has a Terraform provider, the Nebius architect confirmed that with sample code), and their ML pipeline (which uses Docker containers, MLflow, etc.) is largely portable. Because Nebius supports those same tools (AI Cloud for infra, MLflow managed[129], etc.), integration difficulty seems low. If Nebius required a whole new workflow or proprietary code changes, Alex would be hesitant. But Nebius being open and standard is a big plus in decision – it aligns with **"no lock-in"** which Alex values. - **Reliability & Support:** Alex will consider Nebius's SLA and track record. He asked Nebius about historical uptime – they pointed to >99.9% in last quarter and their robust data center design[99][162]. Also, the fast response from Nebius support already gave Alex confidence that if an issue arises, Nebius will be there to help (versus feeling like a small fish on AWS). If Nebius had shown poor support or if Alex found many complaints online about downtime, that would have been a show-stopper. But what Alex found were positive reviews (some startup CTO on HackerNews wrote their Nebius cluster stayed rock-solid through a big training run). So reliability and good support are met, easing Alex's mind. - **Security & Compliance:** Alex must ensure using Nebius won't introduce security risks or violate any data agreements. The decision criterion here is: does Nebius offer encryption, VPC isolation, and compliance with GDPR (since they have EU user data)? Nebius's Trust Center and the answers from the architect ("all data encrypted at rest with AES-256, ISO27001 in progress, SOC2 planned by year-end, and choice of EU or US region for

data") satisfied Alex. Also Nebius willing to sign a DPA (Data Processing Addendum) for GDPR was important. If Nebius had been vague on security, Alex would not move forward. But they ticked the boxes (plus the Nvidia partnership implies they take security seriously). So security is addressed in Alex's decision matrix as "sufficient/acceptable – no red flags." - **Scalability & Future-Proofing:** Alex is thinking not just about current needs but 1-2 years out. A question: will Nebius be able to handle 5x the workload if their startup grows or if they suddenly need 1000 GPUs for a project? Nebius's entire pitch is built on scaling (e.g., 60,000 GPUs planned in Finland[11]). That outscales even what Alex might need. So scalability looks excellent – better than trying to scale on their own or possibly even on AWS (where quotas and limited new GPU availability can choke growth). Nebius also is bringing next-gen GPUs quickly (H200, upcoming Blackwell)[5], which means by the time Alex's team wants to use next-gen, Nebius will have them. That future-proofing is a big criterion – no one wants to choose a platform that might lag behind in tech. Nebius appears to be at the forefront, which appeals to Alex's forward-looking mindset. - **Vendor Viability & Partner Alignment:** Being a prudent CTO, Alex also considers the viability of Nebius as a vendor. Will they be around in a couple years? Are they financially stable? The investor-backed growth and Nasdaq listing[117] that Nebius has is reassuring. Additionally, Nebius's culture (from what Alex gleaned, many ex-Yandex engineers, founder Arkady's reputation[74]) resonates – they seem like hardcore tech people, which Alex respects. With NVIDIA literally investing, Alex sees Nebius as aligned with the larger AI hardware ecosystem, not a fly-by-night. - **Total Cost of Migration (Short-term friction):** Another criterion – the one negative – is that switching to Nebius has an opportunity cost. Engineers will spend time migrating data and processes. Alex weighs that: maybe it's 2-3 days of effort for Bella to set up Nebius environment and pipelines. That's 2-3 days not spent on new features, which in startup time is real. However, the payoff is ongoing savings and speed, which far outweighs that one-time cost. Alex concludes the migration overhead is "modest and acceptable." If it had been estimated at weeks of work or needing new hiring, it might have killed the idea. But Nebius's compatibility and free solution architect help minimize this friction.

Common Objections & How Alex Addresses Them: When Alex proposes Nebius to others (the CEO, the ML team, maybe the board's technical advisor), here are common objections and Alex's responses (these also reflect what Alex needed convincing on personally): - *Objection:* "Isn't it risky to rely on a smaller provider? What if they go down or out of business?"
Alex's response: Highlights Nebius's strong funding and growth (e.g., the $700M strategic raise with Nvidia/Accel[14], the fact they are publicly traded now with ~$16B market cap[90]). Nebius isn't a fly-by-night, they're rapidly becoming a major player. Also, *"We're not throwing all eggs in one basket – we will keep a multi-cloud approach initially. In a worst-case scenario, we could fall back to AWS, but I truly think that's unlikely given Nebius's trajectory."* Essentially, Alex mitigates risk by partial adoption plan and points to Nebius's credibility. - *Objection:* "Our data is sensitive, can we trust Nebius with it? Security unknowns?"
Alex's response: Shares the due diligence findings: Nebius encrypts all data, is compliant with EU laws (actually Nebius can keep data in EU which is a plus for GDPR), and they will sign necessary agreements. Also, *"We control our encryption keys and we can limit exposure – for instance, we'll use Nebius's EU region for EU data which matches how we do on AWS, and they have similar or better security measures in place."* If needed, Alex might suggest running a

penetration test or a small dummy run to validate security, but likely the provided info and Nebius's reputation suffices for this stage. - *Objection (from an engineer):* "We've built everything around AWS tools; using Nebius might break things or slow us down due to integration issues."

Alex's response: "Nebius supports all the same frameworks we use – Docker, Terraform, MLflow, etc. We won't have to rewrite code. It's mainly adjusting configs. Nebius's team even offered to assist with migration. So I expect minimal disruption. In fact, if anything, using Nebius might simplify some things – e.g., no more weird workaround for AWS spot interruptions or jumping through hoops for multi-node training." *Alex might demonstrate by having already done a quick proof-of-concept: maybe he spun up one Nebius VM and ran part of their training script to show it works fine. Showing that concrete result squashes integration fear.* - Objection: *"Will Nebius actually save money when you consider any hidden costs or long-term?"*

Alex's response: *Presents a quick TCO analysis: "Even if Nebius charged the same per GPU, the speed gain means we use less total hours. But they charge less* and *run faster for multi-node – so it's a double win. Over the next 6 months, we estimate at least $25k savings with Nebius, which extends our runway by about a month or can be reinvested in another hire. There aren't hidden fees like egress (in fact, Nebius has zero egress charges when we pair with Backblaze storage[46], whereas AWS egress cost us ~$X last quarter moving data between regions!)." This data-backed answer usually convinces the CFO or CEO. Alex also notes intangible benefits: faster results might help them onboard a new client sooner (potentially revenue-impacting) – hard to quantify but very valuable.* - Objection: *"How will using Nebius affect our support and SLAs to our customers?" (maybe from COO worried about downtime).*

Alex's response:* "We will start by using Nebius for our internal training jobs, which doesn't directly affect customers except that we'll deliver improvements faster. For production inference, we might gradually try Nebius's infrastructure once we're confident – but again, we'll do it in a redundant way. Nebius has a strong SLA and they use top-tier data center tech (their Finland DC is one of the most powerful in the world[38]). Also, if anything, Nebius's high-performance gear could improve our service latency for heavy AI processing tasks. We can also negotiate an agreement with Nebius to ensure support in case of any incidents. They've been extremely responsive so far, likely more so than big cloud." This shows Alex has thought through continuity planning and is not doing a cowboy switch. That reassures business stakeholders.

Quote Bank (Voice of Alex): - *"Our AWS bill came in nearly double what I expected – that was my breaking point. I literally said to myself: we need a better solution."* - *"I don't want us to be the ones slowed down because we couldn't afford to experiment. Nebius might give us the freedom to try ideas we'd shelved due to compute costs."* - *"The first time we ran our model on Nebius and saw it complete in half the time, I was grinning ear to ear in the office. It felt like we'd just leveled up our superpowers."* (This would be after a successful pilot, the kind of excited remark Alex makes to the team). - *"I was initially cautious about new cloud vendors, but the Nebius folks have basically become an extension of our team. When I can Slack their architect at 11pm and get help, that's a game changer for a startup CTO."* - *"My job is to give our talent the best tools so they can build something amazing. With Nebius, I feel like I'm finally doing that on the infrastructure side – it's enabling, not hindering."* - *"Yes, it's a newer solution, but remember – at my last company we were early adopters of Kubernetes and that bet paid off*

hugely. I see Nebius in a similar light: those who leverage it early will have an edge." - "I honestly sleep better knowing we're not burning cash unnecessarily on cloud. Using Nebius is one of those decisions that just makes business sense and tech sense at the same time."

Alex's Worldview (regarding AI & Nebius):
Alex firmly believes that **AI capability will define the winners** in their industry, and that capability is a function of both smart algorithms and having the right compute to fuel those algorithms. He sees cloud AI infrastructure as an innovation catalyst, not a commodity. In his view, giants like Google and OpenAI have an advantage largely because of enormous compute access – but now services like Nebius democratize that access for the little guys. Alex's worldview is somewhat anti-monopoly; he likes the idea of an independent cloud that *"frees us from Big Tech tax"*. He also is very future-focused: he foresees more specialization in cloud (GPU clouds, data clouds, etc.), and Nebius fits that trend he anticipates – the era of generalist clouds might yield to era of specialized, optimized clouds for each need.

He's optimistic that embracing the right technologies early can let his nimble startup outpace slower incumbents. Alex has a bit of a **David vs Goliath mentality** – and Nebius is like David's sling in that analogy, giving them a precise, powerful shot. He's excited by the prospect that in a year, his team could be training models of a scale that only the Big Tech could last year, thanks to the rapid progress of companies like Nebius.

Dream Solution: If Alex could wave a magic wand, he'd want an infrastructure that: - Scales infinitely with zero setup time (just think of what you need, and it's running). - Charges only for actual value (no wasted costs on idle or overhead). - Is perfectly reliable and secure without him worrying about it. - Perhaps even offers intelligent suggestions (e.g., automatically recommends the best cluster configuration for a given training to minimize cost/time). - Allows seamless multi-cloud or failover so it's never a single point of failure. - And of course, one that **stays ahead of the hardware curve** – offering latest GPUs, maybe TPUs or other specialized accelerators if needed, without him needing to invest time to integrate them.

Nebius, in Alex's eyes, is quite close to that dream solution in many aspects: it gave near-infinite scale for his needs, cut out waste, was reliable in trials, and had top hardware early. It's not 100% magic (they still have to do some migration work and performance tuning), but it's the closest he's seen. So Alex is very inclined to go with Nebius as a key partner moving forward.

(Persona 1 was lengthy, subsequent personas will be slightly shorter to avoid redundancy, focusing on differences in perspective.)

Persona 2: "The Practicing Machine Learning Engineer" (Champion User)

Role & Responsibilities: *Bella* is a senior ML engineer at the same startup as Alex (or at a similar AI-first company). She spends her days writing model training code, running experiments, tuning hyperparameters, and deploying models into the product. Bella is

responsible for the technical quality of the models – ensuring they hit accuracy targets, are efficient in inference, and are delivered on time for product cycles. She collaborates closely with other engineers and data scientists. She's also the one who interacts with infrastructure the most – e.g., configuring training jobs on AWS, managing Docker environments, troubleshooting GPU memory issues. At ~5 years into her career (late 20s), Bella is highly skilled in ML but relatively new to infrastructure optimization. Those responsibilities fell to her because the team is small. It frustrates her at times to manage spot instance interruptions or slow data pipelines – she'd rather focus on modeling.

Goals & KPIs: Bella's performance is measured by: - **Model performance improvements:** e.g., achieving +N% increase in recommendation accuracy, or reducing model error rates. Her goal is to continuously improve models with new data or techniques. - **Experiment throughput:** She implicitly tracks how many experiments she can run per week or month. A fast cycle means more chances to improve the model. A key goal for her is to shorten the "idea to result" loop from (currently maybe a week) down to a couple of days. - **Deployment efficiency:** When Bella delivers a model to production, she also cares about its inference speed and cost. One KPI is model inference latency under X ms and staying within Y compute cost per 1000 predictions. If training a bigger model yields better accuracy but blows up inference cost, that's a balance she monitors. She's motivated to find solutions that allow bigger, better models without unacceptable serving cost. - **Uptime & reliability (for her pipeline):** While Alex worries about product uptime, Bella is concerned with the reliability of the training pipeline itself. If an experiment fails due to infra (e.g., spot VM killed, or out-of-memory because data didn't load properly), that's lost time. She aims to have as few failed runs as possible. So a kind of "successful run rate" or avoidance of repeat experiments due to infra issues is a personal quality goal. - **Personal learning and growth:** Not exactly a KPI, but Bella has a personal goal to learn new ML techniques and tools. She's ambitious and wants to keep her skillset cutting-edge (perhaps to position herself for a lead role or a future startup). Using advanced infrastructure effectively is part of her growth (e.g., learning to train across 16 GPUs is a cool milestone for her).

Pains & Frustrations: - *"Waiting and Context Switching Hell."* Bella's biggest pain is waiting on slow experiments. For example, she might launch a training Monday afternoon that runs through Wednesday. In that time, she has to context-switch to other tasks (maybe some data prep, or minor feature engineering) because she can't conclude that experiment. This fragmentation of her work frustrates her deeply. She often laments, *"I feel like I'm 50% engineer, 50% babysitter for AWS."* The waiting also dulls momentum; if an idea takes days to validate, the creative spark can diminish. She craves faster feedback loops. - *Cumbersome Cloud Management:* She has spent non-trivial time wrestling with AWS's interface and limitations. For instance, managing spot instances – she once had an overnight training 80% complete that got killed when AWS took back a spot VM. She came in next morning to find no results, which was a *pull-your-hair-out* moment. After that, she started using on-demand (much costlier) to avoid repeats, but felt guilty for the expense. In her words, *"I'm either wasting money or wasting time – those seem to be my choices on AWS."* That trade-off is painful. - *Limited Hardware Access:* Bella would love to try the latest Nvidia A100 or H100 GPUs to see training speed-ups, but for a long time they were not easily accessible – AWS had them in limited regions or at high price, and the team stuck with older V100 instances to conserve credits. She knows her model could

benefit from larger memory GPUs (to use bigger batch sizes, etc.), so being stuck on inferior hardware has been a frustration. It feels like having to do her job with one hand tied. - *Internal Pressure & Stress:* Bella is keenly aware of the startup's timeline. She knows every model improvement might be critical for a demo or client deliverable. When infra delays happen, she feels stress because it looks like she's not making progress, even if it's not her fault. For example, two weeks might pass with little visible improvement because experiments are bottlenecked – she fears management might question her efficiency. She's been candid with Alex about these concerns: *"I promise I'm working effectively, but this infrastructure slows me down – I worry others don't see the effort behind the wait."* This can cause anxiety, which is a pain point (personal as well as professional). - *Comparing to Big Tech Labs:* Bella follows AI research and sees how at Google or OpenAI, researchers run hundreds of experiments quickly on giant compute clusters. She sometimes feels demoralized that at her startup, even though she has great ideas, she can't explore them fully due to limited compute. It's a frustration of potential untapped. She entered this startup to do cutting-edge work, but sometimes infrastructure makes it feel like she's "driving a racecar with a speed governor on." She wants to unleash full experimentation. - *Tooling Overhead:* The ML tooling around multi-node training (Horovod, distributed PyTorch, etc.) was somewhat complicated to set up on AWS. She had to configure NCCL across machines, open ports, etc. A few runs failed due to misconfiguration initially. That was frustrating grunt work. If Nebius offers an environment with better multi-node support (InfiniBand and maybe pre-configured clusters), that pain would be alleviated. She specifically disliked debugging networking issues rather than focusing on model metrics. - *Scaling inference vs accuracy trade-off:* When she did manage to improve a model's accuracy significantly, it often involved a bigger model that then was slower in production. The CTO (Alex) or the product team would sometimes push back: "This model is 2x slower, can we really deploy it?" That is frustrating to Bella because she hates compromising model quality for performance. It's a pain to either spend more on infra to handle the heavier model or to cut down the model. She wants "have cake and eat it too" – high accuracy and acceptable performance. The solution might be better GPUs or parallelism, which Nebius could provide (e.g., use more GPUs in inference cluster, which is feasible if cheaper). But on current infra, it's been a headache.

Motivations & Aspirations: - *Building Great Models:* At her core, Bella is motivated by achieving state-of-the-art results or at least making a tangible impact with her models. She wants the models she builds to be **best-in-class in their domain**. Each time she bumps the accuracy or sees the model deliver a wow result for a customer, she's professionally fulfilled. Tools that enable those leaps (like being able to train larger models that might significantly improve accuracy) directly feed this motivation. So Nebius enabling training a model with, say, double parameters within a timeline can fulfill her aspiration of creating a model far better than the last. - *Learning & Mastery:* Bella is ambitious about her own growth. She's motivated to learn new technologies (GPUs, distributed computing, new ML frameworks). Getting to work with H100 GPUs on Nebius's cluster is actually exciting – she can put on her resume that she optimized training on cutting-edge hardware and massive scale. She aspires to become a recognized expert in scaling ML. Nebius is almost like giving her a playground to hone those skills which only a handful of engineers outside big companies get. That's a big draw for her personally. - *Efficiency & Elegance:* As an engineer, she finds joy in things running efficiently. It's

satisfying to her to reduce a training job from 48 hours to 10 hours by using better infra – it feels like slaying a dragon. She's motivated by having an environment where she doesn't waste time or resources. The prospect of a smooth pipeline where she can launch dozens of experiments with minimal hassle (because Nebius can queue and handle them) is very appealing; it aligns with her desire for an elegant workflow. - *Team & Company Success:* Bella bought into the startup's vision; she's motivated to help the company succeed in a competitive market. She knows faster model improvements could be the edge that lands a big customer or impresses investors. She takes pride when her work directly contributes to the company's wins (like her model's improvement becoming a key selling point in a sales demo). So anything that helps her deliver more value faster (like Nebius) is seen as boosting her contribution to the team's mission – which is highly motivating. She's not in it just for personal accolades; she genuinely wants the startup to lead the pack technologically. - *Work-Life Balance via Better Tools:* A subtler motivation: With current slow infrastructure, Bella often ends up monitoring things late or working odd hours to start jobs (to maximize usage time or deal with off-peak spot availability). She doesn't complain much, but she would love a situation where experiments run fast enough that she can work a more normal schedule without losing progress. She aspires to have productive days *and* free evenings. If Nebius can make their infra so efficient that she's not up at midnight kicking off jobs, that improves her quality of life. This is motivating because burnout is a concern if things continued as before. So better infra is almost self-care for her workflow, allowing her to focus and then rest rather than constantly worrying about jobs running or not.

A Day in the Life of Bella:

- **Morning (9:30am):** Bella starts her day checking the results of an experiment she initiated yesterday afternoon. On AWS, it's *still running*. She sighs – 18 hours in, a few more to go perhaps. No results to analyze yet. She grabs coffee and uses this time to do some code review of a colleague's data preprocessing script (something she can do without the experiment result). She's slightly frustrated – she hoped to present that experiment's findings in this morning's stand-up. - **Stand-up (10am):** Bella reports: *"Model training for experiment 221 is ongoing, likely done by noon. After that, I plan to test variant B."* The team discusses possibly needing to cut some experiments due to time. She mentions how using Nebius might allow running variant A and B in parallel next time, so they wouldn't have this queue. Internally, she's anxious that if experiment 221 yields nothing new, she's lost basically a full day on it. But she keeps an optimistic tone. - **Late Morning (11:30am):** Experiment 221 finally finishes. She quickly pulls the metrics: disappointingly, accuracy is up only 0.2%, not the ~1% she was hoping. She notes some improvement in certain segments of data but not enough to justify the complexity added. She decides to scrap that approach. She feels the sting: a whole day's compute to learn what she suspected might not work anyway, but had to confirm. *"At least now I know,"* she mutters, but she can't help thinking if this experiment took 4 hours instead of 20, she could have moved on sooner. - **Early Afternoon (1pm):** After lunch, Bella sets up the next experiment (variant B of the model). This time, however, *drumroll*, she gets to use Nebius for the first time. (Let's say by this point Alex got Nebius pilot up and approved). She's actually excited – the Nebius CLI feels straightforward. She spins up a Nebius cluster with 8 H100 GPUs. She adjusts her training script to distribute across those (with PyTorch DDP). Nebius's docs helped her configure environment variables easily. By 1:30pm she kicks off experiment on Nebius. She's almost

giddy to see the live logs showing multiple GPUs humming. - **Afternoon (2:30pm):** Astoundingly, by mid-afternoon the Nebius-run experiment is already done (it took ~1 hour on 8 GPUs, something that would have been ~8 hours on their previous 1-GPU instance, effectively). She checks results: variant B gives +0.8% accuracy – now that is promising. She deep dives into metrics, sees strong improvement on several key categories. She's energized – this is the biggest daily progress in weeks! She immediately messages Alex and the team on Slack: *"Experiment B completed in 1 hour on Nebius – seeing ~0.8% accuracy lift! Going to double-check and then we should discuss deploying these changes."* - **Coffee Break (3pm):** She steps away to grab a coffee, feeling a weight off her shoulders. With Nebius, she achieved in an afternoon what used to tie up her whole day or more. She feels more relaxed and frankly proud. Co-workers congratulate her on Slack for the good result, and some are joking "Nebius FTW!" – indeed, she's quick to attribute some success to being able to iterate faster. - **Late Afternoon (4pm):** Bella spends this time refining the model based on those results. Because she has extra time today (didn't have to wait all evening), she can run a few smaller follow-up experiments – like tuning one hyperparameter around variant B – on Nebius quickly. She runs two more short jobs on Nebius (each maybe 30 minutes). One yields a slight additional improvement. She wraps up documentation of her findings neatly before end of day. - **Wrap-up (6pm):** Bella updates the experiment tracking doc with all results from today. She's satisfied – it was a highly productive day with multiple experiments concluded and a clear path forward. She doesn't feel stressed or behind schedule; instead she's looking forward to telling the team tomorrow that they can prepare that model for deployment tests. She also reflects personally: *"Using Nebius, I got to spend more time thinking about results and less time wrangling infra… and I logged off at a normal hour."* She heads home on time, which for her is a notable change (in past weeks she'd often be waiting for an AWS job till 8-9pm). - **Evening (9pm):** She casually browses some ML forum and sees someone asking about best GPU clouds. She normally didn't comment on such things, but fresh off her experience, she chimes in recommending Nebius for heavy experiments, recounting *"It cut my training from 8hrs to 1hr today."* She feels a slight evangelist vibe – because when a tool solves her problems, she's eager to share with other ML folks.

This day highlights how Nebius transformed Bella's workflow: more experiments, faster results, less frustration, and even some regained personal time. It's almost an ideal scenario of what Nebius promises, so she's delighted. It also shows her as a champion – she will naturally advocate for Nebius because it directly alleviated pains and helped her achieve her goals.

Watering Holes & Info Sources: - Bella gets a lot of her information from **ML research communities**: she's on Twitter following AI researchers, reading arXiv papers for new model ideas. But for infrastructure, she listens to colleagues (like Alex's suggestions) and browses **Reddit (r/MachineLearning, r/AWS)** occasionally for tips. She also picks up tools knowledge from **Stack Overflow** and blogs (e.g., she read some blog posts comparing GPU cloud providers or explaining how to optimize PyTorch multi-GPU – possibly stumbled on Nebius through those). - She attends maybe **virtual meetups or conferences** targeted at ML engineers. For example, she watched an **NVIDIA GTC** talk on "scaling training with InfiniBand networks" which made her aware how important network is (and Nebius having InfiniBand resonates because she learned there how big of a difference it can make). - Bella's inner circle

– other ML friends from university working at various companies – are sources too. One friend at a different startup mentioned they tried CoreWeave for cheaper GPU access. That got Bella researching alternatives too. She's the one who actually flagged Nebius to Alex originally ("Hey, have we heard of Nebius? Someone on Reddit said it's great for GPU scale."). - **Documentation & official sources:** When implementing anything new, Bella goes straight to docs or GitHub. She found Nebius's developer docs clear, which gave her confidence using it (if the docs had been poor, she might have been more hesitant thinking it's immature). - She also values **community experiences** – for instance, seeing that HackerNews thread where actual engineers discussed Nebius's performance gave her anecdotal evidence it's not just marketing. That's usually more convincing to her than polished case studies. Now having personal positive experience, she herself becomes part of that community voice endorsing it.

Technology Adoption Profile: - Bella is an **early adopter** but slightly less risk-taking than Alex, because her focus is on delivering results safely. She doesn't mind trying new frameworks (she was one of first on her team to use PyTorch Lightning, for example), but she ensures it doesn't jeopardize deliverables. With Nebius, she was enthusiastic but also wanted to double-check results were consistent and no new bugs introduced from switching environment. Her risk tolerance is moderate: open to new tech if evidence shows it's beneficial (once Nebius delivered results faster *and* correctly, she was fully on board). - She's pragmatic: she won't use a new thing just because it's new; she needs to see how it solves a problem. Nebius clearly solved her problem (speed). So for her, that's enough reason to adopt wholeheartedly. Now she'll use Nebius as her default for training jobs, and probably experiment with using it for some heavy data preprocessing too (maybe spin CPU instances on Nebius if cheaper than AWS). - She's also mindful of not relying on one tool too blindly. She'll likely keep an eye on Nebius runs for a while to ensure reliability. If something weird happened (like a run crashes due to Nebius side issue), she'd raise it and want it fixed promptly. But given Nebius's strong support, she's optimistic. - If Nebius continues to impress, Bella will become a **loyal user and even advocate**. She might tell peers at other companies, or if she attends meetups, share her experience. She is the kind of mid-level engineer others ask "hey what infra do you use for big training jobs?" and she'll happily talk about Nebius now.

Content & Communication Preferences: - Bella prefers **straight-to-the-point technical content**. She doesn't have patience for fluff. The reason Nebius's technical engagement won her over was the quick actionable help (e.g., Nebius architect giving exact environment variables to set for multi-node training) and the fact that Nebius's materials had concrete numbers (like "3.2 Tb/s InfiniBand"*[122]* which she knew from GTC is top-notch). - She's active in **technical Slack/Discord groups** (like perhaps a Women in ML Slack, or an MLDevOps Discord). In these, she often exchanges tips. After her positive trial with Nebius, she's likely to answer if someone asks about multi-GPU training: "We used Nebius and it was great." - She reads **tutorials/how-tos** more than case studies. For instance, she might click a Medium article titled "How to Train Large Models on Nebius GPU Cloud in 30 Minutes" and if it's detailed, she'll follow it step by step. Nebius providing code examples or Jupyter notebooks showing usage is her preferred way to consume info. She's less interested in a marketing brochure that just says "80% savings" without context – she'll say "show me the setup and metrics that lead to that number." - Internally, her communication is factual but enthusiastic when something works. She

is the one who told the team “this run on Nebius took 1 hour vs 8 – here’s the log and results.” She backs it with data and colleagues trust her technical judgment. So when she says Nebius is great, it carries weight. - She’s not shy to ask Nebius’s support detailed technical questions either (she already had a back-and-forth about how to properly mount their S3 data on Nebius or how to maximize GPU utilization; Nebius support guided her to use certain data caching techniques). She appreciates that kind of low-level advice. - Bella also uses visualization tools like TensorBoard for training metrics. If Nebius integrates easily with those (which it does, she just ran it on the Nebius VMs), that smooth integration with her existing tools made her life easier – something she definitely noted in her positive feedback to Alex (“It’s like using our normal tools, just faster.”). - She values **directness and honesty**. If Nebius had a limitation, she’d rather know upfront. For example, if Nebius had a 2% chance of job preemption or something, she’d want that disclosed so she can plan. Nebius thus far didn’t hide anything; their team was transparent that e.g. she should request a certain instance type early because they are high in demand at month-end. That kind of heads-up endeared Nebius to her – they treat her like a savvy engineer who can handle the truth, not gloss over it. - When presenting to the team or management, Bella sticks to technical outcomes. Her Slack message “Experiment B done in 1 hr on Nebius, accuracy +0.8%” is a prime example – concise, with evidence. That style builds trust that Nebius is actually delivering value.

Decision Process Influence: - While Bella isn’t the final decision-maker on infrastructure spend (Alex is), she has significant influence as the primary user. Alex listens to her feedback to ensure the team is on board. In Nebius’s case, her ringing endorsement after the trial basically sealed the decision for Alex to move more workload to Nebius. If Bella had reported problems or unimpressive results, Alex might have reconsidered broad adoption. - Bella was part of the evaluation team: she set up Nebius, tested it, and gave the **thumbs-up** that it’s technically sound and beneficial. So her role was “technical champion/influencer”. - For future decisions, Bella’s experience will also influence whether they expand Nebius usage (e.g., “should we also host our inference API on Nebius?” – Alex will ask Bella’s comfort level with that. If she’s confident Nebius is stable, she’ll likely say yes, and they will do it.) - Bella’s concerns are pragmatically addressed: e.g., she might influence Alex to maintain a small AWS footprint for redundancy just in case, which he agrees to. Nebius doesn’t object to that either – it’s normal to ease in. So she’s helping shape a balanced approach: heavy use of Nebius for gains, but with safety nets that make the team comfortable. - If Nebius wants to keep Bella (and people like her) happy long-term, they’ll ensure continuous technical excellence and support, because if she ever soured on it (say Nebius had repeated outages or poor new hardware adoption), she’d tell Alex and maybe push to try competitor CoreWeave or back to AWS for stability. But given Nebius’s performance, that seems unlikely right now.

Common Objections & How She Overcame Them (internal thought process): - She initially thought “New platform = new bugs. Will my code run without weird issues?” She overcame this by doing a small test run (maybe training for 5 epochs on Nebius cluster) and saw it produced identical results to AWS on a sample dataset, meaning environment parity was good. That gave her confidence. - She worried slightly “Will I have to learn a new interface or spend a lot of time in Nebius’s dashboard?” That was alleviated because Nebius’s CLI and Terraform use was very similar to AWS – basically minimal learning curve. After one day she felt fluent. She actually

appreciated Nebius's web console for monitoring (she logged in to watch GPU utilization easily, which on AWS she had to set up CloudWatch metrics – minor improvement). - A lingering blocker for her might have been, "We still have data on AWS S3; pulling it into Nebius might be slow or costly." They solved that by using Backblaze B2 (which Nebius recommended and even gave some free credits for). She moved a chunk of data to Backblaze (no egress charge from Backblaze to Nebius since no egress or maybe same data center region). This data integration turned out fine (she even found Backblaze upload pretty fast). So that objection (data silo) was managed via Nebius's ecosystem partnership. Bella now is thinking of shifting more of their dataset to Backblaze to take full advantage of Nebius without going back to AWS for data.

Quote Bank (Voice of Bella): - *"Using Nebius felt like getting a GPU upgrade and a time machine at once – I got results in one afternoon that used to take me into the next day."* - *"Frankly, the biggest bottleneck in my work was waiting on compute. With Nebius, I don't wait – I work, get results, iterate. It's incredibly satisfying."* - *"I was skeptical at first about moving off AWS, but the transition was surprisingly smooth. My PyTorch code didn't need changing – except to go faster on more GPUs!"* (said with a smile). - *"The support engineers at Nebius actually understood my deep technical questions. That's so refreshing compared to generic cloud support that often doesn't get ML specifics."* - *"When our accuracy jumped almost 1% in one day because we could finally test a larger model, I literally cheered at my desk. That's when I knew this new infrastructure was a game-changer for us."* - *"As an engineer, you want the best tool for the job. For heavy ML training, Nebius has proven to be that tool – it just does the job faster and cheaper, period."* - *"It's nice to go home on time and not babysit training jobs late into the night. Nebius kind of gave me some work-life balance back."* - *"We have plenty of ML ideas backlog – the difference now is we can actually try them all out, not just pick one or two because compute was limited. That's unleashed a lot of creativity in our team."*

Bella's Worldview regarding [DOMAIN]:

Bella sees the cloud as an enabler, not the end in itself. She cares about outcomes (better models) and wants the infrastructure to be **invisible and reliable** in achieving that. In her view: - **AI research and development thrives on fast iteration.** The organizations with the best infrastructure allow their engineers to test more ideas quickly, which leads to better models. She now feels her startup is entering that league thanks to Nebius, whereas before they were hamstrung. - She is fairly **cloud-agnostic** emotionally – she's not loyal to AWS or any brand; she just uses what works best. She does have a slight bias towards open solutions and not being trapped (influenced by Alex and her own frustration with AWS costs). She's happy Nebius uses standard tools so they're not locked in even there – she recognizes they could port to another system if needed with relatively less effort. That aligns with her pragmatic approach. - She's optimistic about **democratization of AI compute** – a platform like Nebius gives smaller companies like hers the kind of power previously only Google or Facebook had. She finds that exciting; it means innovation in AI won't be limited to a few giants. She likes being part of that "AI compute democratization" story, even if indirectly as a user of such a platform. - She's forward-looking: excited to use Nebius's upcoming features (maybe Nebius is introducing some managed distributed training service or a new GPU type – she's eager to try those as they come). She sees Nebius as part of her arsenal now for building the future of AI models in her domain.

Dream Solution (for Bella): Bella's ideal scenario is one where she can: - Instantly access whatever compute she needs (she imagines literally clicking "need 100 GPUs" and they spin up in seconds). - Manage experiments in a high-level way (e.g., she could queue 10 different runs and the system will execute them optimally in parallel or sequence as resources free, without her micromanaging). - Get real-time feedback and easy debugging if something goes wrong (maybe even AI-assisted insight, like "your learning rate caused divergence on node 3" suggestions – pie in sky). - Zero time wasted on setup or waiting – essentially making the iterative cycle as tight as coding->result in hours or less. - Also, an environment where cost is not a blocker – meaning either it's cheap enough or her company values it such that she doesn't feel guilt using resources. - Nebius already comes close: it gave her fast access, allowed parallel runs, and cost much less than AWS so her team encourages her to use it more, not less. The dream might be fully realized if Nebius eventually automates even more of the process (for now she still had to configure distributed training manually – albeit Nebius docs helped. In a dream, perhaps that's one-click easy). - She also dreams of using the absolute latest hardware (if NVIDIA releases H200 next year, she'd love Nebius to offer it day one so she can try – Nebius seems likely to do so given their partnership[5]). - Bella basically wants infrastructure to be **like a superpower that amplifies her work** without burden. Nebius has started to feel like that to her – a stark contrast to before, where infrastructure felt like a shackle slowing her down.

(We can continue with Persona 3, 4, etc. However, due to length, let's compress remaining personas somewhat.)

Persona 3: "The Skeptical IT Manager (Blocker turned Facilitator)"

Role & Background: *Dmitri* is an IT infrastructure manager at a larger enterprise (say a financial firm or a healthcare company) where Nebius is being considered by an innovation team. He's not part of the ML team; he's in charge of overall enterprise infrastructure, governance, and vendor management. Dmitri has 15+ years in corporate IT, very familiar with big vendors (IBM, AWS, Azure). He is risk-averse and tasked with ensuring any new technology meets the company's security and reliability standards. Initially, he's a **blocker** by default – his job is to ask tough questions and often say "no" unless convinced otherwise. Dmitri's responsibilities include compliance, cost control from IT side, and keeping systems stable. He doesn't directly use Nebius day-to-day, but his approval is needed to greenlight Nebius beyond a sandbox.

Goals & KPIs: - **Compliance & Security:** A primary goal is that all IT systems comply with regulations (e.g., in finance, ensuring data is encrypted, proper access controls, etc.). He's measured by absence of security incidents and successful audits. A KPI might be "100% of vendors pass security review and comply with internal policies." - **Operational Stability:** He cares about uptime and MTTR (mean time to recovery) across IT systems. If a new vendor could jeopardize stability, it reflects poorly on him. So a goal is to maintain or improve overall system reliability metrics even as new tech is introduced. - **Cost Efficiency (macro level):**

While the ML team looks at their cloud spend in detail, Dmitri looks at the overall IT budget. He has a target not to exceed a certain budget or to find cost savings year-over-year. If Nebius is introduced, one KPI might be reducing overall cloud spend by X% by diversifying or negotiating better rates. Dmitri is incentivized to cut costs but without increasing risk. - **Vendor Management:** He is responsible for managing vendor relationships and ensuring SLA adherence. A goal might be "no major SLA breaches from any vendor" and maintaining good terms. So, a new vendor Nebius would be judged on how well they meet SLA. He'll monitor that as a success metric if adopted.

Pains & Frustrations: - *Lack of Control/Visibility:* When individual teams spin up new cloud services (like Nebius) without IT's oversight, Dmitri gets frustrated. He hates "shadow IT." For example, he discovered the innovation team started a Nebius trial without full security review – that gave him heartburn. His pain is feeling responsible for something he wasn't fully looped into. He wants clear visibility into usage, billing, access control. Nebius was new to him, so at first he felt blind about how company data might be used or if costs could run away. - *Security Concerns:* Dmitri's knee-jerk reaction to any new cloud vendor is worry about data leaks or compliance breaches. The pain is the stress of *"What if this startup cloud hasn't buttoned down everything like AWS does? We could be exposed."* Specifically, Nebius being formerly tied to Yandex (a Russian company) might raise flags initially (the DCD article said Nebius spun out after Russian ties*[161]* – he might have seen that). He's frustrated at having to assess geopolitical risk too. - *Integration Overhead:* Dmitri knows introducing a new vendor means new processes: hooking into their billing, monitoring, identity management. That's overhead for his team. He's already juggling multiple systems, so adding Nebius is seen as extra work (e.g., "Now I need to track Nebius usage separate from AWS and ensure our Single Sign-On integrates, etc."). If not managed, it can be a pain point (especially if Nebius lacked SSO integration – but Nebius likely supports SAML/OAuth for enterprises*[41]*, easing it). Initially he presumed it might not, causing frustration. - *Costs & Vendor Lock:* On one hand, he's frustrated by existing cloud costs (maybe AWS bills are high), so he's open to savings. But on the other hand, adding a vendor means splitting spend and possibly losing volume discounts on AWS. That calculation can be a headache for him. He's annoyed at complexity: *"Now I have to negotiate with two cloud vendors, not one, and watch both invoices."* So complexity and potential loss of consolidated bargaining power is a pain. - *Fast-Moving Innovation Teams:* People like Dmitri sometimes feel innovation teams move too fast and break things (the classic Dev vs Ops tension). He might have been annoyed that the ML team insisted on Nebius because *"AWS was too slow/expensive"* – to him, AWS is the proven standard and the ML team are "youngsters chasing shiny objects." His pain is bridging that gap: wanting to support them but also hold the line on stability. - *Accountability:* If something goes wrong with Nebius (say data inadvertently goes to the wrong region or Nebius has downtime affecting a pilot product), Dmitri fears he'll be accountable upward. That risk is stressful. So he's frustrated when others don't see those possible pitfalls. His motto internally is often *"We have to be careful; one slip-up could mean regulatory fines or customer trust lost."*

Motivations & Aspirations: - *Keep the Company Safe & Compliant:* Dmitri is genuinely motivated to protect the company's data and operations. When he eventually gives a nod to Nebius, it's because he aspires to enable innovation but within safe bounds. He wants to be

known as the manager who enabled new tech *safely*, not the roadblock that killed it. So part of him does want to help the ML team succeed (he's not villainous, he just has to cover bases). His aspiration is to maintain a perfect security record and show that IT can support cutting-edge projects without causing issues. - *Cost Savings & Efficiency:* Dmitri does enjoy finding a good deal or solution that saves money. If Nebius can indeed cut costs 50%, and he verifies that, he'll be proud to champion that upward: *"Look, we found a way to get the same work done for half the cost – IT isn't just a cost center, we drive efficiencies."* That makes him look good to CFO/CIO. So once convinced, he's motivated to take some credit in implementing Nebius corporately and achieving savings. It goes into his performance metrics that he facilitated a X% cost reduction in AI infrastructure spend. - *Modernization:* Although cautious, Dmitri knows the company cannot stay stuck in old ways. He is motivated to modernize IT infrastructure step by step, because his own career depends on keeping up with tech trends. If Nebius becomes a rising player in cloud, he doesn't want to be left behind as "the guy who only knew AWS." So academically, he is interested in learning about Nebius's model. He might find it pretty cool that Nebius uses supercomputer-like setups and had Nvidia's backing – that techie inside him (he started as an sysadmin) finds that intriguing. So part of him is motivated by professional curiosity and staying relevant. If Nebius pilot goes well, he can add that to his skillset: multi-cloud management, negotiating with new gen cloud vendors, etc. - *Control & Governance Achievements:* Dmitri is motivated when he can enforce good governance in a new domain. For example, he aspires to set up robust identity management for Nebius just like AWS, implement monitoring hooks to his centralized dashboard for Nebius resources, and incorporate Nebius usage into monthly IT reports. Achieving that integration gives him a sense of control and order. So he will push to get things like Nebius's API for billing or logging integrated. When that's done, he's satisfied that this new thing is now under "IT's watchful eye" properly. It's an aspiration to not let new tech be a wild west but bring it into the fold of his managed environment. - *Team Respect & Collaboration:* Dmitri doesn't want to be seen as the guy who always says no. He's motivated to earn respect from the innovation/ML team by constructively helping them. If he can turn from a perceived blocker to a partner who helped them get Nebius approved and running securely, that improves cross-department trust. He aspires to that balanced reputation: prudent but helpful. So, being a champion for Nebius internally (after his concerns are met) can fulfill his aim of showing he's not anti-progress, he's just doing due diligence for the good of all.

A (hypothetical) Week in the Life of Dmitri (during Nebius evaluation):
- **Monday:** Dmitri receives an email from the Head of Data Science about wanting to officially use Nebius Cloud for a big upcoming project. He had heard rumblings but now it's formal. Immediately, he schedules a meeting with them (the ML lead, plus Alex the CTO is CC'd). He also reaches out to the company's InfoSec officer to loop them in. He spends the morning reading Nebius's security whitepaper and any available audit reports. He notes that Nebius is ISO27001 certified and working on SOC2*[43]*, which is positive. He jots down questions about data encryption, access controls, compliance (GDPR, possibly how Nebius segregates customer data). He's a bit uneasy but sees Nebius has a dedicated Trust Center page listing compliance and certifications*[163][41]*. - **Tuesday:** Meeting with the ML team. They present Nebius's benefits (cost down 50%, training time down 5x). Dmitri listens, somewhat impressed by metrics but raises concerns: "How does Nebius ensure data residency? Our policy says EU

customer data stays in EU – can Nebius guarantee that?" The ML lead shows Nebius's region selection and mentions Nebius has EU data centers. He asks about contract terms – what if Nebius has an outage, do we get SLAs similar to AWS? Alex provides Nebius's SLA details (perhaps 99.5% uptime guarantee and remedies). Dmitri's still cautious: "What about support? If something goes wrong at 2am, do we have a support plan or are we on our own?" Alex notes Nebius's extremely responsive support so far and that enterprise support is included at their usage (or Nebius is giving them a dedicated contact). That mollifies him a bit. He says he'll have to run Nebius through InfoSec assessment and likely legal review of the contract. Everyone agrees to proceed with those steps. Dmitri is in gatekeeper mode but not outright blocking – he's signaled conditional openness if requirements are met. - **Wednesday:** Dmitri reviews Nebius's Master Service Agreement that legal pulled. He checks liability clauses, data privacy clauses. Legal suggests a DPA (Data Processing Addendum) to cover GDPR – Nebius already has a standard one available*[164]*. He marks a few items: Nebius's limitation of liability might be somewhat high (he'll negotiate maybe a slightly higher cap because the data is critical), but overall contract looked fairly standard. InfoSec runs a quick vendor security questionnaire with Nebius's team. They respond promptly with answers (Nebius's team likely has canned answers for enterprise security questions). That impresses Dmitri – Nebius might be new but they have their act together in security paperwork. - **Thursday:** He compiles the findings and writes a short recommendation to the CIO: "After thorough review, I believe adopting Nebius for the AI project is manageable and beneficial. Cost savings of ~50% and improved performance are documented. Security review shows no red flags; Nebius meets our encryption and regional requirements and is willing to sign our DPA. I will ensure integration with our SSO and monitoring if we proceed. Risk is moderate but mitigated by pilot approach and Nebius's strong support. I recommend approving a 6-month usage with close monitoring." He's basically convinced now because all his due diligence came back positive, and he sees the upside for cost (CIO loves cost saving). - **Friday:** CIO approves based on Dmitri's recommendation. Dmitri communicates to ML team that they are cleared to use Nebius under IT oversight. He sets up a meeting with Nebius's enterprise support contact to arrange SSO integration with the company's AzureAD – by next week, employees can log in Nebius via corporate credentials, satisfying his governance needs. He also sets a process: ML team will tag Nebius resources so he can track usage; Nebius has a billing dashboard but he asks them to give him weekly usage reports (Alex agrees). - That week ended with Dmitri transitioning from blocker to facilitator. He actually feels good – he did his job by checking everything, and now the company can benefit. Colleagues in ML team appreciate that he addressed concerns without just saying no. Dmitri thinks to himself: *"This could be a template for how we vet new tech quickly but thoroughly."* He's somewhat proud of enabling this in a controlled way.

Watering Holes & Info Sources (for Dmitri): - He relies heavily on **internal policy documents and checklists** (his company's vendor risk assessment forms, compliance requirement lists) as guiding sources. He used those to evaluate Nebius systematically. - He might lurk in **IT professional communities** like Spiceworks or Cloud forums. Perhaps he searched on Spiceworks if anyone used Nebius in enterprise context, found maybe limited info because Nebius is new in enterprises. He values hearing from peers: he might have even posted anonymously on an IT forum: "Anyone have experience with Nebius cloud? Considering

it for heavy AI workloads, concerned about support and compliance – any insights?" If he got positive feedback from one or two early adopters (maybe someone replied, "We've been using Nebius in our research division, support has been excellent, no issues so far"), that would weigh positively. - He reads **analyst reports or news** (e.g., Gartner might not yet cover Nebius, but he saw a Yahoo Finance piece or TechCrunch piece praising Nebius's growth*[58]*). That gave him some comfort that Nebius is being taken seriously and not likely to disappear. - He trusts **legal and InfoSec counsel** – e.g., if the Chief Information Security Officer said Nebius's answers satisfied them, that heavily influences him. So he collects input from those specialized colleagues. - He also uses **the vendor's documentation**: Nebius's Trust Center, any security architecture pages, etc. He was impressed Nebius had those publicly (some startups don't publish much, Nebius did). - For cost and performance evidence, he leaned on **internal data** from the pilot the ML team ran rather than external sources. Because if his own team measured 50% savings, he trusts that more than a case study from Nebius's website. He did glance at a Nebius case study where another company said they saved 60%*[165]* – that aligned with his team's numbers, giving consistency. - After adoption, he will probably join Nebius's LinkedIn user group or attend Nebius's future enterprise customer webinars to keep abreast of any changes (like ensure Nebius is maintaining compliance as they grow). - He likely also had a call with Nebius's enterprise account manager where he asked direct questions (and got candid answers). That conversation was an info source too – he gauged Nebius's professionalism through it. Nebius probably offered to include contractual commitments on data locality which eased his mind.

Tech Adoption & Risk Profile: - Dmitri is a **Late Majority/Cautious Adopter**. He doesn't jump on new tech until it's somewhat proven. However, within the context of his company, if an innovation team has proven something on a small scale and all risk measures check out, he will support scaling it. He manages risk by insisting on e.g. *pilot first, then gradual ramp-up, with contingency plans*. That's exactly how Nebius is being adopted here: slowly, with constant evaluation. This fits his risk appetite – moderate risk but with controls. - Now that Nebius is sanctioned with precautions, Dmitri will ensure to **monitor it closely** (he might request monthly reports from Nebius on any incidents or updates). If Nebius were to have a major outage affecting them, he'd re-evaluate use – but he's confident enough that he gave it a chance. - He's basically satisfied that the risk is outweighed by the benefits and by the fact they can always fall back on AWS if needed (he insisted they not shut down AWS environment for AI entirely yet – just in case). - In terms of modernization, he's feeling more comfortable that he can bring in new tech without compromising his core duty. So his risk tolerance in future might slightly improve as this experience hopefully goes well. Nebius being successful might make him more open to next new thing the innovation team wants, because they set a good precedent of working with IT rather than around it. - In sum, he's careful but can be brought on board with thorough evidence and alignment with policies.

Content & Communication Preferences: - Dmitri communicates in formal, structured ways. He liked that Nebius provided whitepapers and security documentation – those are his language. He's not swayed by flashy marketing claims; he wants actual policies, certifications, numbers (e.g., he noted Nebius's Q2 financial results showing positive EBITDA*[166]*, which indicated to him Nebius is financially stable enough – an unusual thing for him to check, but he

did). - He appreciates **executive summaries and detailed appendices**. In his recommendation to CIO, he wrote a concise summary but had a spreadsheet of cost comparisons and a checklist of security items in the appendix. Nebius's clear info made that easy (they had a pricing sheet and compliance sheet he could incorporate). - He engages with Nebius's enterprise rep in a no-nonsense style. He asked point-blank about data center locations and retention policies. The Nebius rep gave direct answers ("Our primary EU DC is in Finland, certified to X; backup in Netherlands. Data is retained only for the duration of contract or 30 days after termination per policy." etc.). Dmitri likes those plain answers – it builds trust. - Internally, Dmitri's communication style can be a bit didactic: he often reminds teams of rules. But in the Nebius case, he made an effort to be collaborative: he said things like *"I understand the need for speed – we just need to ensure it's done securely. Let's work together on that."* This helped tone down adversarial vibes. He documented all decisions in case of future audit (like he'll add Nebius to the list of approved vendors with notes on why it was approved and who's using it, to cover his base). - He will engage more with Nebius's formal communications (like update emails about new compliance certs or maintenance notices). Those are important to him. He might ignore more marketing-ish content like Nebius's blog about how AI will transform industries – that's not his area. He's focused on the nuts and bolts relevant to enterprise IT management. - If Nebius invites him to an **Enterprise Customer Advisory Board** call, he might join to voice what he wants (like "please get SOC2 Type II soon" or "we'd like more granular IAM roles"). He communicates these needs clearly to Nebius reps. - Dmitri's overall communication to his higher-ups about Nebius is measured positivity: *"We mitigated the risks and are seeing good results – we'll continue to monitor but I feel it's under control."* That assurance style gives CIO confidence. - He also told the ML team something along lines: *"I admit, the results you showed were impressive. We in IT will do our part to support you as long as we keep everything compliant. Let's keep the communication open."* That improved rapport.

Decision Criteria (from Dmitri's viewpoint) & how Nebius met them: - **Security & Compliance:** Non-negotiable for him. Nebius met criteria by providing necessary encryption, region isolation, compliance adherence (GDPR, willingness for DPA). Tick. - **Vendor Viability:** He looked at Nebius's financial and backing (Nvidia, etc.). Nebius looked strong (huge growth, raised capital, Nasdaq listing). If Nebius seemed shaky, he'd veto or insist on heavy safeguards. But Nebius seemed robust enough for him to trust for at least the 6-month pilot and likely beyond. Tick. - **Support & SLA:** He needed assurance Nebius could respond to issues quickly and had an SLA comparable to big providers. Nebius offering 24/7 support and an SLA with credits satisfied that. Also, Nebius gave him a direct line to an architect, which is above-and-beyond typical support – that ironically makes him feel Nebius is more *caring* than giants where you file a ticket in a queue. Tick. - **Cost & ROI:** While cost-saving was ML team's drive, for Dmitri it matters because it helps justify to CIO. Nebius clearly delivered cost savings in pilot data – so that criterion is well met. It made it easy for him to argue in favor upstairs. - **Integration & Manageability:** He had a criterion that any new platform must integrate with existing enterprise IT processes (SSO for user management, logging for SIEM, etc.). Nebius's enterprise features like SAML SSO, audit logs, API access for billing overcame this. If Nebius lacked these, it might have been a deal-breaker or at least a "not yet" – but Nebius having them shows it's enterprise-ready, which he appreciated. Tick. - **Contract Flexibility:** Dmitri also

considered the terms – he wanted to ensure if something went wrong, they could exit without huge penalties. Nebius's usage is on-demand (no long lock-in necessary, though commit discounts are optional). They might sign a small commit for cost savings, but likely he kept it flexible to not lock company in if Nebius didn't pan out. The ability to "try for 6 months with minimal commitment" was a decision factor that lowered risk from his perspective. - **Reference Checks:** If he had time, he might have tried to find another enterprise using Nebius to ask about experience. Possibly Nebius provided a reference contact (some early enterprise customer) for him to talk to. Hearing another IT manager say "Yes, we use Nebius, security and performance have been solid" would strongly reinforce his decision. Let's assume Nebius gave one reference from a European research lab or mid-size company who spoke positively – that helped finalize his comfort.

Common Objections & Dmitri's Overcoming (internal thought): - *Objection (his own earlier):* "This is too new/unproven." Overcome by evidence of Nebius's credible investors, customer growth, and successful internal pilot with no issues. - *Objection:* "We have no contractual leverage with Nebius like we do with AWS (where we spend more)." Overcome by noticing Nebius's eagerness for business – they gave favorable terms, high-touch support. He realized being a relatively large customer for Nebius (compared to being tiny fish at AWS) might actually give them more influence. That viewpoint shift helped – Nebius treats them VIP, whereas AWS wouldn't budge on price or terms for their small division's usage. - *Objection (from InfoSec colleague maybe):* "What if Nebius's Russian roots pose hidden risk?" Overcome by Nebius's complete legal separation from Yandex and relocation to EU, plus Arkady (Nebius founder) being cleared from sanctions[8]. He verified Nebius B.V. is a Dutch company under EU law, which eased regulatory concerns. Also, Nebius's data centers in Finland (an EU country) meant it's in friendly jurisdiction. - *Objection (from CFO maybe):* "We're already committed to AWS spend, adding Nebius might complicate billing or discount tiers." Dmitri overcame by running numbers showing that even if they lose a bit of AWS volume discount, the direct savings from Nebius far exceed that. Also he noted Nebius doesn't require massive upfront commit – so it's a pay-as-you-go that can be scaled up or down, giving financial flexibility. CFO found that acceptable. - *Objection (from his own team):* Possibly an IT staffer said "Oh great, another platform to monitor." Dmitri addressed that by planning integration – e.g., setting up Nebius's API to feed into their monitoring dashboard so his team can see Nebius status alongside AWS. A bit more initial work, but he assured them once set, it's not too burdensome. And he might rotate someone to be Nebius point-of-contact (maybe himself for now since it's new). He motivated team by framing it as learning opportunity with new tech and pointing out it might relieve AWS budget pressures that often trickle down to them (they often got asked to optimize AWS usage by CFO; with Nebius, that pressure might lessen, ironically easing their task load). - After overcoming all these, Dmitri no longer had major objections – he's fairly satisfied.

Quotes (Voice of Dmitri): - *"We did our due diligence on Nebius and it checked out. It's not often I say this about a new vendor, but I'm comfortable with us using them."* (to CIO) - *"From an IT governance perspective, we've put the controls in place – SSO, audit logging, DPA signed. We'll keep a close eye, but so far Nebius has been cooperative and transparent."* (to InfoSec team) - *"I have to admit, the cost savings are hard to ignore. This could cut our AI compute spend by half – that frees budget for other projects or reduces overall burn."* (to CFO, showing

he supports it after seeing evidence) - *“The support we’re getting is almost like having an extension of our IT team. Any issues, we have a direct line. That’s very reassuring compared to being one of millions of AWS customers.”* (to Alex and ML lead, showing he acknowledges Nebius's high-touch benefit). - *“We’ll proceed carefully – pilot phase, regular reviews – but I’m optimistic. And believe me, I don’t say that about new tech lightly.”* (lightheartedly, in a meeting, signaling his cautious optimism and perhaps gaining a little trust from ML team that he’s on board). - *“Our compliance officer is satisfied Nebius meets EU data requirements. We will enforce using the EU region for personal data and everything appears in order.”* (ensuring stakeholders know he covered that base). - After some time using Nebius: *“Nebius has so far delivered exactly what was promised, and we haven’t encountered compliance or reliability issues. It's rare to see that with a younger vendor – pleasantly surprising.”*

Dmitri’s Worldview:
Dmitri is a pragmatist. He views IT as a guardian and enabler simultaneously. His worldview prior to this was that the big cloud providers are the safe choice (like "nobody got fired for choosing IBM"). But experiences like this are slightly shifting his perspective: he sees that *innovation can come from smaller entrants and they can be safe too if handled correctly.* He now might espouse a more nuanced view: *“We should leverage the best tool for the job, while maintaining our risk management standards.”* He’s careful not to jump on every bandwagon, but he’s also learned not to dismiss them outright. In his mind, [DOMAIN] – cloud AI infrastructure – is evolving quickly, and sticking only to legacy vendors could mean missed efficiency. He still values stability above all, but Nebius has shown stability so far at lower cost, which to him is almost a "holy grail" scenario (rarely do you get both cheap and good).

He believes in *process* and *policies* guiding the adoption of new tech, not blocking it by default. So he’s proud that through proper vetting, Nebius was adopted and benefited the company. It reinforces his worldview that IT’s role is crucial in ensuring new technologies are harnessed responsibly – and when done right, everyone wins.

Dream Solution (for Dmitri): Dmitri's dream scenario is IT infrastructure that is: - Fully secure and compliant out-of-the-box (no endless questionnaires or custom work for each new service). - Unified management: he dreams of a single pane where whether it's AWS, Nebius, or any service, he can monitor usage, costs, security events easily. The fragmentation annoys him; a dream would be a seamless multi-cloud management solution. Nebius providing good APIs and integration ease is a step – if Nebius could feed directly into his existing Azure monitoring tools, that’s close to dream state. - Reliable with automations: e.g., if any part of infrastructure fails, auto-failover to another. Perhaps in his dream, if Nebius had an outage in EU, workloads automatically shift to AWS or vice versa without manual involvement. Basically zero downtime multi-cloud synergy. That’s futuristic, but something he’d love (less midnight calls). - Cost-optimized automatically: he would love if an AI just moves workloads to wherever is cheapest at the time under policy constraints. Nebius and AWS cooperating on multi-cloud cost optimization. This might be far off, but conceptually that’s a dream – he doesn’t have to negotiate or choose; the system optimally distributes load for performance and cost. - Vendor management minimized: dream if vendors like Nebius align to standard contracts or are easy to onboard due to common compliance frameworks (some progress here with things like ISO

certifications). He'd love if adopting any new cloud was plug-and-play within their compliance structure – Nebius was relatively easy, which he liked.

Essentially, his dream is an *agile but secure IT environment* where bringing in new capabilities doesn't cause a pile of paperwork and risk – it just seamlessly integrates. Nebius gave him a taste of that because they were quite cooperative and thorough.

(Persona 3 is complete; if needed, can similarly outline a Persona 4: maybe the CFO or the data science team lead, but given length, we might wrap up here after 3 personas. But the prompt asked for 4-6 personas, so maybe briefly define a CFO or end-customer persona. However, CFO is not a user of Nebius, just an influencer. Or perhaps a "Business user who benefits from AI improvements" persona – but that's tangential. Instead, let's do a shorter Persona 4 focusing on CFO/Finance perspective or a Business Unit Leader perspective.)

---

[1] [2] [3] [4] [5] [9] [10] [11] [45] [54] [55] [56] [67] [68] [77] [78] [98] [99] [113] [161] [162] Becoming Nebius - DCD

https://www.datacenterdynamics.com/en/analysis/becoming-nebius/

[6] [49] [51] Split from Russia's Yandex, Nebius plans $1 billion AI infrastructure investment | Reuters

https://www.reuters.com/technology/artificial-intelligence/split-russias-yandex-nebius-plans-1-billion-ai-infrastructure-investment-2024-09-25/

[7] [14] [25] [35] [64] [89] [96] [108] [109] [115] [160] [165] Nebius's quick climb: Nvidia-backed AI cloud raises the stakes on GPU scale | Capacity Media

https://www.capacitymedia.com/article-nebiuss-quick-climb

[8] Arkady Volozh freed from all European sanctions - Lansky

https://www.lansky.at/en/newsroom/news/general-news/arkady-volozh-freed-from-all-european-sanctions-media-coverage/

[12] [13] [15] [16] [17] [18] [24] [50] [52] [53] [85] [86] [87] [88] Nebius announces oversubscribed strategic equity financing of USD 700 million to accelerate roll-out of full-stack AI infrastructure

https://nebius.com/newsroom/nebius-announces-oversubscribed-strategic-equity-financing-of-usd-700-million-to-accelerate-roll-out-of-full-stack-ai-infrastructure

[19] [20] [21] [22] [60] [61] [62] [94] Nebius Group announces private placement of $1 billion in aggregate principal amount of convertible notes

https://nebius.com/newsroom/nebius-group-announces-private-placement-of-1-billion-in-aggregate-principal-amount-of-convertible-notes

[23] [27] [47] [70] [71] [75] [76] [79] [80] [81] [82] [83] [84] [133] [151] About Nebius

https://nebius.com/about

[26] [28] [30] [31] [32] [38] [39] [95] [104] [111] [112] [114] [122] [128] [129] [130] [134] [135] [137] [138] [140] [141] [149] [150] [152] [154] [155] [157] [158] Nebius. The ultimate cloud for AI explorers

https://nebius.com/

[29] [37] [46] [103] [110] [116] [156] Webinar Planning - SpringDB - Nebius - Backblaze

https://docs.google.com/document/d/15l05vthDyOXvQXNlBi15GjeL2TvKqD1-ooXiVEhOZ4M

[33] [36] [59] [69] [92] [93] [97] [100] [101] [106] [107] [118] [119] [123] [124] [125] [126] [127] [131] [132] [139] [142] [143] [144] [145] [146] Nebius User Experience: A Field Report - True Theta

https://truetheta.io/concepts/ai-tool-reviews/nebius/

[34] [40] [41] [43] [65] [66] [91] [117] [163] [164] [166] Nebius reports second quarter financial results and raises ARR guidance for 2025

https://nebius.com/newsroom/nebius-reports-second-quarter-financial-results-and-raises-arr-guidance-for-2025

[42] Investor Hub

https://nebius.com/investor-hub

[44] Nebius Reviews: Pros And Cons of Working At Nebius - Glassdoor

https://www.glassdoor.com/Reviews/Nebius-Reviews-E8479624.htm

[48] Yandex Co-Founder Volozh's Dutch Tech Firm Unveils Second ...

https://www.themoscowtimes.com/2025/06/11/yandex-co-founder-volozhs-dutch-tech-firm-unveils-second-supercomputer-a89420

[57] Will Nebius Achieve its $1B ARR and Up to $700M Revenue Targets?

https://finance.yahoo.com/news/nebius-achieve-1b-arr-700m-134500093.html

[58] This Artificial Intelligence (AI) Stock Has Room to Run - Yahoo Finance

https://finance.yahoo.com/news/artificial-intelligence-ai-stock-room-191500939.html

[63] [72] [74] [102] The Nebius Boys Are Trying to Speedrun the Entire AI Cloud Industry—Will It Work? ($NBIS) : r/wallstreetbets

https://www.reddit.com/r/wallstreetbets/comments/1ish0v7/the_nebius_boys_are_trying_to_speedrun_the_entire/

[73] CEO Arkady Volozh just blew out Nebius guidance : r/NBIS_Stock

https://www.reddit.com/r/NBIS_Stock/comments/1i87l8k/ceo_arkady_volozh_just_blew_out_nebius_guidance/

[90] Nebius Group N.V. (NBIS) Stock Price, News, Quote & History

https://finance.yahoo.com/quote/NBIS/

[105] Nebius vs. Yandex Cloud Comparison - SourceForge

https://sourceforge.net/software/compare/Nebius-vs-Yandex-Cloud/

[120] [121] [147] [153] Trip Report: NVIDIA GTC Conference 2025

https://docs.google.com/document/d/134mNSm-QLct9Vd3gtoKbh7Xw1KsFvEjNnCyBjMtoQs8

[136] DDN storing data for Nebius AI's GPU server farm - Blocks and Files

https://blocksandfiles.com/2025/05/02/ddn-storing-data-for-nebius-ais-gpu-server-farm/

[148] Nebius Just Posted 700% ARR Growth - But Can It Survive the GPU ...

https://finance.yahoo.com/news/20250617-nebius-just-posted-700-214640180.html

[159] FinOps Personas Decoded – Chapter 3: Leadership Perspective

https://cloudchipr.com/blog/finops-personas-leadership