



COLUMBIA SPS
Artificial Intelligence Lab

Open Weights, Closed Frontiers:

Small Language Models,
Data Sovereignty,
& the Enterprise
Model-Portfolio Decision

AI Models & Infrastructure

Enterprise Model Strategy



Kim, Kibaek
'25SPS, Applied Analytics

9 October 2026 • WP-MI.EMS-01 v1.0



1 Executive Summary

Currently, businesses obtain AI from two primary ecosystems: closed frontier models provided through cloud APIs, and open-weight models which can be fine-tuned and self-hosted. This article looks at how enterprise technology and AI leaders should distribute their model portfolios between these two choices. Based on a review of existing research, open models dominate experimentation and make up around one-third of real-world token traffic, but closed APIs take most paid enterprise spending and hold a capability lead of about four months. The suggested method is a portfolio that uses closed APIs where up-to-date capability is commercially necessary and small open models where data sovereignty, cost, or latency matter most. Orchestration and federated training support this strategy. A practitioner note illustrates these points with a seven-agent, fully open-weight deployment. However, evidence remains limited, and the paper concludes with a discussion of open questions and a proposed research agenda.

2 Topic Overview & Context

Closed models, such as OpenAI's GPT, Anthropic's Claude, and Google's Gemini, are used via cloud APIs, with the company keeping the model weights, the training data, and control over deployment. In contrast, open-weight models such as Meta's Llama, Google's Gemma, Alibaba's Qwen, DeepSeek, and Microsoft's Phi make their parameters available so that other people can carry out fine-tuning and run them on their own. Here, 'small' means a model that runs on ordinary consumer or edge hardware, typically 1 billion to 30 billion parameters (Belcak et al., 2025).

Enterprise spending on generative AI is expected to rise by nearly three times to \$37 billion in 2025, with \$12.5 billion of that sum being allocated to foundation-model APIs (Tully et al., 2025). Nevertheless, only 20% of the 3,235 executives who were surveyed say that they have experienced any revenue growth as a result of AI (Deloitte AI Institute, 2026), a fact which has made cost and control key factors in sourcing decisions. Regulation also plays a role in these decisions. Since August 2025, the EU AI Act has imposed obligations on providers of general-purpose AI, but has included fewer requirements for truly open models (European Commission, 2025). At present, 89% of organizations include open-source components in their AI systems (Hermansen & Osborne, 2025).

The following three questions arise: What is the position of each ecosystem in terms of usage, paid spending, and top-end capability? Is the remaining advantage of the closed frontier still relevant for use in enterprise applications, or is it possible to eliminate it through system design? And what actual rule should be applied when allocating workloads? To answer them, this paper first reviews the research published to date and then complements the findings of that review with experience from a real-world deployment.

The literature review draws on three types of sources: peer-reviewed and preprint technical papers for capability claims; trend data from organizations such as Stanford HAI, Epoch AI, and

the Linux Foundation; and market data from industry surveys by Menlo Ventures, a16z, Deloitte, KPMG, and Cisco. Preference has been given to primary reports over press coverage and almost all of the sources date from the past 24 months. The survey figures should be treated with care since the sampling frames vary and the same figure can differ between the various trackers; the two references to OpenAI in Finding 1 illustrate this point. If the surveys agree on the direction of change, then that direction is regarded as reliable and the percentages are seen as indicative. The Chatbot Arena, which is used in Findings 2 and 3, is a contested measure (Singh et al., 2025) but has been included since there is no other longer-term series available.

3 Key Findings from the Literature

Finding 1: Open-weight models are widely adopted, but closed APIs capture most paid enterprise spending

Even though open models are widely used, they account for only a small portion of paid enterprise revenue. Of the more than 100 trillion tokens that had passed through the OpenRouter aggregator by the end of 2025, about one third were processed by open-weight models (Aubakirova et al., 2025). In the enterprise sector, however, expenditure on LLMs linked to open models dropped from 19% in 2024 to 11% in 2025, even though total spending increased threefold (Tully et al., 2025). OpenAI comes out on top when looking at either of the two available measures, although the different tracking methods give conflicting estimates of the size of its lead: Menlo’s spending data indicates that it has 27% of the enterprise market, while the a16z CIO survey suggests it holds around 56% of the wallet share (Tully et al., 2025; S. Wang et al., 2026). The decline overstates the retreat because spending on open models has still risen and most open usage is self-hosted, so it is counted in infrastructure budgets rather than in API spending. The spending figure thus reflects closed providers’ business models more than the actual adoption of open models. The same survey also weakens the case for open models: over one third of enterprises now prefer closed models, partly because of data security concerns; the total cost of ownership of the two is becoming more similar; and enterprises are shifting from fine-tuning to prompting and routing (S. Wang et al., 2026). These findings mean that the recommendations in Section 4 have to be narrowed, since the default option of using open models is appropriate only for teams that plan to invest in domain adaptation or for workloads that are subject to data sovereignty.

Finding 2: Small open models have collapsed the cost of “good enough”

A benefit of having an open ecosystem is that well-designed small models can beat out much larger ones. For instance, Gunasekar et al. (2023) trained a code model called phi-1, with 1.3 billion parameters, on 7 billion tokens of carefully selected and synthesized data; on HumanEval it obtained a pass@1 score of 50.6%, thus surpassing models that are ten times larger and which have been trained on more data. It is obvious that high-quality curation can substitute for raw scale, enabling enterprise teams to fine-tune small models with focused domain data. Table 1 gives an overview of recent developments (Abdin et al., 2024; Ben Allal et al., 2025;

DeepLearning.AI, 2026; DeepSeek-AI, 2025; Gemma Team, 2025). Economic factors also result in the same conclusion. A 7B model costs between 10 and 30 times less to run than a 70 to 175B model (Belcak et al., 2025). Inference prices for models of equivalent capability have dropped by a median of 50 times each year (Cottier et al., 2025), and Meta’s 4-bit Llama 3.2 models are now two to four times faster on smartphones (Meta AI, 2024).

Table 1. Small open models against larger or closed baselines

Open model	Benchmark	Result	Outperforms	Source
phi-1 (1.3B)	HumanEval pass@1	50.6%	Models ~10× larger on far more data	Gunasekar et al., 2023
Phi-4 (14B)	GPQA / MATH	56.1 / 80.4	GPT-4o, its own teacher (50.6 / 74.6)	Abdin et al., 2024
DeepSeek-R1-Distill (14B)	AIME	69.7%	OpenAI o1-mini (63.6%)	DeepSeek-AI, 2025
Gemma 3 (27B)	Chatbot Arena	Elo 1338, top ten	Llama 3.1 405B (15× larger)	Gemma Team, 2025
Qwen3.5 (9B)	Most language benchmarks tested (except reasoning and coding)	—	gpt-oss-120B (>10× larger)	DeepLearning.AI, 2026
SmolLM2 (1.7B)	Mixed; over-trained on ~11T tokens	—	Same-class peers	Ben Allal et al., 2025

HumanEval is employed as a means of testing code generation; GPQA covers graduate-level science questions; the MATH and AIME collections involve competition mathematics; in the case of Chatbot Arena, the models are evaluated through the use of preferences obtained from a crowd-sourced head-to-head comparison, the preferences being expressed in the form of an Elo rating; and AlpacaEval 2.0 examines instruction-following, the judgments for this being supplied by a language model.

Finding 3: The closed frontier holds a narrow but durable lead of about four months, partly for structural reasons

Closed-model systems still have an advantage, as can be seen on Chatbot Arena, where the difference between closed and open models fell from 8.0% in January 2024 to 1.7% in February 2025 before rising to 3.3% by March 2026 (Maslej et al., 2025; AI Index Steering Committee, 2026). Epoch AI claims that the best open models are around four months behind (Edwards & Emberson, 2026). This delay can in part be attributed to structural reasons, since proprietary laboratories can carry out private testing and Google and OpenAI together held nearly 40% of the Arena’s feedback data, while 83 open-weight models accounted for less than 30% (Singh et al.,

2025). The point that needs to be taken into account, however, is the impact that these four-month delays have on enterprise workloads.

Finding 4: "Open" is a spectrum, policed now by licenses and regulators

It is in dispute what should be considered as open. When the Open Source Initiative applied its Open Source AI Definition, it found that only a small number of models met the criteria; Llama, Phi-2, Mixtral, and Grok do not (Vidal, 2024). An audit of 46 so-called "open" systems introduced the term "open-washing" to describe the discrepancy between the labels and the actual licenses (Liesenfeld & Dingemanse, 2024). For example, the license for Meta's Llama 3.1 includes different terms for services that have more than 700 million users per month (Meta Platforms, Inc., 2024). A similar situation occurred within one company in August 2026, when Alibaba released Qwen3.8-27B (repository created on August 5, 2026) under the Apache 2.0 license while Qwen3.8-Flash-Next (created on August 24, 2026) was given a custom license (Qwen Team, 2026a, 2026b). The Linux Foundation's Model Openness Framework looks at 17 different aspects of a release (White et al., 2024). The EU AI Act provides exemptions on the basis of open-source status only if both the license and the disclosures are truly open (European Commission, 2025).

Finding 5: Orchestration, not parameter count, closes most of the remaining capability gap

The delay of four months is based on a per-model, per-prompt, and per-answer basis, while this is not how actual systems function. J. Wang et al. (2024) created the Mixture-of-Agents (MoA) method, which involves a sequence of language models in which each one sees the answers given by the previous layer. An all-open-weight MoA achieved 65.1% on AlpacaEval 2.0, beating GPT-4 Omni's 57.5%, which shows that an open ensemble can outperform the best closed model on certain benchmarks. As for cost, Chen et al. (2023) developed FrugalGPT, a system that uses a cascade approach starting with a less expensive model and only proceeding to more expensive ones when needed, and thus attains GPT-4 level performance at a cost that is as much as 98% lower. A similar production system mentioned in 2026 was able to keep 97–99% of its top-tier accuracy while handling the majority of queries by using cheaper models (Moslem et al., 2026). This comparison thus supports self-hosting since, when using an API, extra model passes are charged for whereas with self-hosting such passes simply make use of spare capacity on hardware that is owned. There are, however, some limitations: the additional passes cause greater latency, the MoA results predate the 2026 benchmarks, and many of the tests depend on preferences judged by LLMs.

Finding 6: Federated training makes sovereignty compatible with improvement

It might seem that a model cannot be improved if the data's owner refuses to share the data; however, McMahan et al. (2017) addressed this issue by using the Federated Averaging algorithm, which leaves the training data on the client devices and only combines the model updates, thus reducing the number of communication rounds by a factor of 10 to 100 while at the same time being capable of handling unbalanced, non-IID data. Federated training is only feasible when the weights are open, since the different organizations need to have control over

the parameters. This approach is in line with the view of the 90% of privacy professionals who consider it safer to keep the data stored locally (Cisco, 2025). Sovereignty lies on a spectrum and currently closed providers do offer in-region, single-tenant, or even air-gapped options. The main difference in the case of open weights is that the model itself can be examined, modified, run without the vendor, and improved through collaboration. It is exactly because of this openness that federated training can occur.

Table 2. How the two ecosystems compare on enterprise decision criteria

Decision dimension	Open-weight ecosystem	Closed ecosystem
Frontier capability	Trails by ~4 months on average (Edwards & Emberson, 2026); Arena gap 3.3% as of Mar 2026 (AI Index Steering Committee, 2026)	Leads: six of the top ten Arena models are closed-weight
Cost per capability	7B models 10–30× cheaper to serve; fixed-capability prices falling a median of ~50×/year (~200×/year post-2024)	Premium API pricing; \$12.5B enterprise API spend (2025)
Paid enterprise share	11% of enterprise LLM market, down from 19% (Tully et al., 2025)	OpenAI ~27% of tracked spend (Tully et al., 2025) or ~56% of surveyed wallet share (S. Wang et al., 2026); methodologies differ
Real-world token traffic	~One-third of 100T-token OpenRouter traffic (late 2025)	~Two-thirds of routed traffic
Data sovereignty	Self-hosted, on-premises, or air-gapped; the artifact itself is inspectable and modifiable	In-region, single-tenant, and on-premises options exist; residency is contractual and architectural rather than owned
Licensing clarity	Contested: few models meet the OSI definition	Clear commercial terms of service
Edge/device deployability	Quantized 1–3B models run on phone- and embedded-class CPUs	Requires connectivity to provider infrastructure
Orchestration headroom	Open-model ensembles scored 65.1% on AlpacaEval 2.0 vs. 57.5% for GPT-4 Omni; extra passes cost utilization and latency, not dollars	Same techniques apply, but every additional pass is metered as marginal API cost

Improvement on non-exportable data	Federated fine-tuning across independent parties is available	Single-tenant fine-tuning is available; cross-party federation is not, since participants cannot each hold the weights locally
------------------------------------	---	--

4 Business & Technology Implications

In the short term (0–2 years), most companies are employing a number of different models: 37% of CIOs say that they are using five or more, and LLM budgets are increasing by around 75% each year (S. Wang et al., 2025). Regulatory compliance has now become the top barrier to the deployment of generative AI (Deloitte, 2025), and 36% of organizations plan to enhance data-sovereignty audits across their partnership ecosystems (KPMG International, 2026). In the near term, most organizations route customer-facing work through closed APIs and run regulated or high-volume tasks on small, self-hosted open models.

Over a longer period (3 to 5 years), data sovereignty will become an integral part of the infrastructure. In 2025, spending on AI infrastructure doubled to \$318 billion, and it is forecast to exceed \$1 trillion per year by 2029 (Seminara, 2026). Edge-AI revenue is expected to nearly triple and reach \$157 billion by 2030 (STL Partners, 2025). A workstation with 128 GB of unified memory, such as NVIDIA’s DGX Spark or AMD’s Ryzen AI Max+ 395 (AMD, 2025), is able to run a quantized 70B model locally, with memory bandwidth rather than compute being the limiting factor (Zhou & Chen, 2025). Since the models are to be embedded within sovereign or edge infrastructure, they have to be inspectable, modifiable, and capable of being run on-site, rights that only open licenses provide.

The ethical aspects and broader social impact of the model portfolio decision extend well beyond mere regulatory compliance. While energy trade-offs exist between datacenter-based serving and smaller edge deployments, the paramount social concern is long-term accountability. Because self-hosting eliminates the vendor’s safety buffer between the raw model and its output, core governance functions such as bias audits, safety guardrails, and mechanisms for human redress become the deploying organization’s direct responsibility. Self-hosting thus shifts the societal impacts of AI deployment, including algorithmic fairness and public trust, directly onto the enterprise, making ethical oversight an operational imperative.

The best course of action is to group workloads according to how sensitive the data is, the capabilities required, and the need for cost efficiency. In cases where sovereignty is a concern or the work is being carried out at the edge, use small open-weight models and reserve the closed frontier APIs for situations in which a months-long delay in capability really leads to financial loss. When designing a system, you should plan for a cascade prior to buying raw capability (Chen et al., 2023), and set the latency budget in advance. If it is necessary for the data to remain

decentralized, then use federated fine-tuning. You must always check that the label ‘open’ conforms with the OSI definition or the Model Openness Framework.

5 Practitioner Note: Seven Agents on Self-Hosted Open Weights

The following section outlines the way in which I have put into practice a cosmetics regulatory compliance system. It must be understood as an example only and not as a general statement. The system makes use of seven specialized open-weight agents: intake classification, retrieval of regulatory documents, domain-specific question answering, formulation generation, reading of labels and artwork, rating of ingredient risks, and the anonymization of any text which identifies clients. Models with up to 20 billion parameters are operated on a private network consisting of Apple Silicon machines, and none of the data from regulated clients leaves the organization.

First observation: Specialization proved to be better than size, in agreement with the results from the phi-1 model. When asked five regulatory questions, the model that had been fine-tuned for the specific domain achieved a score of about 4.6 out of 5, whereas the general model, which was running on the same hardware, scored 2.5. In one case, the general model wrongly stated that a prohibited substance was allowed at low levels. The scoring was carried out by one rater and since there were only five questions, the findings are indicative rather than conclusive. The reason for the superior performance was the improved training data rather than the size of the model.

Second observation: Orchestration dealt with the problems that individual models could not handle. For instance, the generator at times gave ingredient percentages that added up to more or less than 100%, so another downstream agent adjusted the totals while making sure the regulatory limits were enforced. The smaller general model produced the first draft and the bigger, domain-fine-tuned model was only used when it was necessary.

Third observation: The main cost associated with data sovereignty was operational, not computational. The throughput amounted to about 17 tokens per second, which was enough for asynchronous review, and as Zhou and Chen (2025) point out, this was constrained by memory bandwidth. This had an effect on the system’s practice of selectively escalating requests. One failure was significant. On two out of three hosts, the same model weights returned their answers in different response fields because of inconsistent configuration flags in the serving runtime. When individuals choose to self-host, they shift the responsibility for reliability from the vendor’s service agreement to their own internal operations, a point that is frequently ignored in the debate between open and closed systems.

6 Open Questions & Research Gaps

The gap in capability between open and closed models narrowed in 2024–25 but once again widened in 2025–26. In August 2026 Alibaba launched the Qwen3.8 family (Qwen Team, 2026a,

2026b), a development that has attracted attention since the vendor-reported figures suggest that it is approaching the closed-model frontier. However, the models' own configuration files show that they are based on the same underlying architecture as the Qwen3.5 generation which was released earlier that year, and there has so far been no independent assessment of their performance on standard benchmarks or in enterprise applications. More broadly, the record is too brief to determine whether the observed gap represents a permanent trend or merely a short-term fluctuation, especially since a great deal of the evidence is based on vendor surveys and the disputed Chatbot Arena benchmark (Singh et al., 2025). The situation with regard to licensing enforcement remains unclear: the OSI definition together with the EU's open-source exemptions could result in stricter standards, or 'open-weight' might in the end just be a marketing term. Moreover, the orchestration results showing that open models are on a par with closed ones are mainly the result of LLM-judged preference tests carried out at the technological frontier of 2024. There is still a lack of large-scale, independent evidence concerning how these ensembles perform when it comes to tool use or complex, multi-step tasks where accuracy cannot be averaged out.

7 A Research Agenda for the SPS AI Lab

Another area that deserves further investigation is the infrastructure layer underneath the application. Open LLMs are currently being used for network management. A fine-tuned Llama 2–7B beat existing algorithms on three networking tasks (Wu et al., 2024), and LLM agents can convert natural language into switch configurations, though with limited first-pass accuracy (Lin et al., 2025). Whether quantized sub-2B open models can run effectively on the control-plane CPUs of network operating systems such as SONiC is untested, and only open weights make the test possible. A companion study could federate telemetry summarization across a network without exporting operator data.

A more ambitious approach embeds the model below the application layer, where semantic communication replaces payload bytes with compact indices drawn from a common codebook; the transport protocols currently in use for this target wireless links (Y. Wang et al., 2025; Y. Wang et al., 2026). Multi-agent systems might dispense with text altogether, since exchanging KV-cache data improves accuracy by 3.1 to 5.4% at 2.5 times lower latency (Fu et al., 2025). Whether a MiniLM-scale encoder (W. Wang et al., 2020) can carry out this encoding at layer 4 of a wired datacenter stack is an open question, and a model integrated into a protocol stack must be open weights, since it cannot depend on external APIs.

8 References

- Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., Lee, J. R., Lee, Y. T., Li, Y., Liu, W., Mendes, C. C. T., Nguyen, A., Price, E., de Rosa, G., Saarikivi, O., ... Zhang, Y. (2024). *Phi-4 technical report* (arXiv:2412.08905). arXiv. <https://arxiv.org/abs/2412.08905>
- AI Index Steering Committee. (2026, April). Technical performance. In *The 2026 AI Index report*. Institute for Human-Centered AI, Stanford University. <https://hai.stanford.edu/ai-index/2026-ai-index-report/technical-performance>
- AMD. (2025). *AMD Ryzen AI Max+ 395: A leap forward in generative AI performance with consumer PC* [Technical article]. <https://www.amd.com/en/developer/resources/technical-articles/2025/amd-ryzen-ai-max-395--a-leap-forward-in-generative-ai-performanc.html>
- Aubakirova, M., Atallah, A., Clark, C., Summerville, J., & Midha, A. (2025, December 4). *State of AI: An empirical 100 trillion token study with OpenRouter*. OpenRouter & Andreessen Horowitz. <https://openrouter.ai/state-of-ai>
- Belcak, P., Heinrich, G., Diao, S., Fu, Y., Dong, X., Muralidharan, S., Lin, Y. C., & Molchanov, P. (2025). *Small language models are the future of agentic AI* (arXiv:2506.02153). arXiv. <https://arxiv.org/abs/2506.02153>
- Ben Allal, L., Lozhkov, A., Bakouch, E., Martín Blázquez, G., Penedo, G., Tunstall, L., Marafioti, A., Kydliček, H., Piqueres Lajarán, A., Srivastav, V., Lochner, J., Fahlgren, C., Nguyen, X.-S., Fourier, C., Burtenshaw, B., Larcher, H., Zhao, H., Zakka, C., Morlon, M., ... Wolf, T. (2025). *SmolLM2: When Smol goes big—Data-centric training of a small language model* (arXiv:2502.02737). arXiv. <https://arxiv.org/abs/2502.02737>
- Chen, L., Zaharia, M., & Zou, J. (2023). *FrugalGPT: How to use large language models while reducing cost and improving performance* (arXiv:2305.05176). arXiv. <https://arxiv.org/abs/2305.05176>
- Cisco. (2025, April 2). *Cisco's 2025 Data Privacy Benchmark Study: Privacy landscape grows increasingly complex in the age of AI* [Press release]. <https://newsroom.cisco.com/c/r/newsroom/en/us/a/y2025/m04/cisco-2025-data-privacy-benchmark-study-privacy-landscape-grows-increasingly-complex-in-the-age-of-ai.html>
- Cottier, B., Snodin, B., Owen, D., & Adamczewski, T. (2025, March 12). *LLM inference prices have fallen rapidly but unequally across tasks*. Epoch AI. <https://epoch.ai/data-insights/llm-inference-price-trends>
- DeepLearning.AI. (2026, March 20). Alibaba's latest flagship models are open-weights MoE performers in sizes from less than 1B parameters. *The Batch*. <https://www.deeplearning.ai/the-batch/alibabas-latest-flagship-models-are-open-weights-moe-performers-in-sizes-from-less-than-1b-parameters>
- DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning* (arXiv:2501.12948). arXiv. <https://arxiv.org/abs/2501.12948>
- Deloitte. (2025, January 21). *The path to sustainable generative AI value balances passion, pragmatism and patience, finds new Deloitte survey* [Press release]. <https://www.deloitte.com/us/en/about/press-room/state-of-generative-ai.html>
- Deloitte AI Institute. (2026, January 21). *The state of AI in the enterprise: 2026 AI report*. Deloitte. <https://www.deloitte.com/us/en/what-we-do/capabilities/applied-artificial-intelligence/content/state-of-ai-in-the-enterprise.html>

- Edwards, J., & Emberson, L. (2026, May 29). *Open models are 4 months behind the state-of-the-art closed models*. Epoch AI. Retrieved August 29, 2026, from <https://epoch.ai/data-insights/open-closed-eci-gap>
- European Commission. (2025, November 11). *Guidelines on obligations for general-purpose AI providers* [FAQ]. Shaping Europe's Digital Future. <https://digital-strategy.ec.europa.eu/en/faqs/guidelines-obligations-general-purpose-ai-providers>
- Fu, T., Min, Z., Zhang, H., Yan, J., Dai, G., Ouyang, W., & Wang, Y. (2025). *Cache-to-cache: Direct semantic communication between large language models* (arXiv:2510.03215). arXiv. <https://arxiv.org/abs/2510.03215>
- Gemma Team. (2025). *Gemma 3 technical report* (arXiv:2503.19786). arXiv. <https://arxiv.org/abs/2503.19786>
- Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., & Li, Y. (2023). *Textbooks are all you need* (arXiv:2306.11644). arXiv. <https://arxiv.org/abs/2306.11644>
- Hermansen, A., & Osborne, C. (2025, May). *The economic and workforce impacts of open source AI*. The Linux Foundation. <https://www.linuxfoundation.org/research/economic-impacts-of-open-source-ai>
- KPMG International. (2026, January). *KPMG Global tech report 2026: Leading in the Intelligence Age—Excelling today, shaping tomorrow*. <https://assets.kpmg.com/content/dam/kpmgsites/xx/pdf/2026/01/global-tech-report.pdf>
- Liesenfeld, A., & Dingemanse, M. (2024). Rethinking open source generative AI: Open-washing and the EU AI Act. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1774–1787). Association for Computing Machinery. <https://doi.org/10.1145/3630106.3659005>
- Lin, S., Zhou, J., & Yu, M. (2025). *An LLM-based agentic framework for accessible network control* (arXiv:2509.20600). arXiv. <https://arxiv.org/abs/2509.20600>
- Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., ... Oak, S. (2025, April). Technical performance. In *The AI Index 2025 annual report*. Institute for Human-Centered AI, Stanford University. <https://hai.stanford.edu/ai-index/2025-ai-index-report/technical-performance>
- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Proceedings of Machine Learning Research, Vol. 54, pp. 1273–1282). PMLR. <https://proceedings.mlr.press/v54/mcmahan17a.html>
- Meta AI. (2024, October 24). *Introducing quantized Llama models with increased speed and a reduced memory footprint* [Blog post]. Meta AI Blog. <https://ai.meta.com/blog/meta-llama-quantized-lightweight-models/>
- Meta Platforms, Inc. (2024, July 23). *Llama 3.1 community license agreement*. https://github.com/meta-llama/llama-models/blob/main/models/llama3_1/LICENSE
- Moslem, Y., Kacmajor, M., Nedumpozhimana, V., Abbas, A., Panahi, S., Lynch, D., Nie, Z., Agapitos, A., Milenovic, A., Song, H., Shi, Y., Pan, Y., Buffini, P., & Kelleher, J. D. (2026). *Cluster, route, escalate: Cascaded framework for cost-aware LLM serving* (arXiv:2606.27457). arXiv. <https://arxiv.org/abs/2606.27457>

- Qwen Team. (2026a, August 5). *Qwen3.8-27B* [Model]. Hugging Face. <https://huggingface.co/Qwen/Qwen3.8-27B>
- Qwen Team. (2026b, August 24). *Qwen3.8-Flash-Next* [Model]. Hugging Face. <https://huggingface.co/Qwen/Qwen3.8-Flash-Next>
- Seminara, J. P. (2026, April 16). *AI infrastructure spending caps historic year at ~\$90 billion in Q4 2025; 2029 spending to eclipse \$1 trillion* [Blog post]. IDC. <https://www.idc.com/resource-center/blog/ai-infrastructure-spending-caps-historic-year-at-90-billion-in-q4-2025-2029-spending-to-eclipse-1-trillion/>
- Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., Üstün, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N. A., Ermis, B., Fadaee, M., & Hooker, S. (2025). *The leaderboard illusion* (arXiv:2504.20879). arXiv. <https://arxiv.org/abs/2504.20879>
- STL Partners. (2025, January 29). *Edge AI revenue will reach USD157bn by 2030 globally, STL Partners predicts in brand new forecast* [Press release]. <https://stlpartners.com/press/edge-ai-revenue-to-reach-usd157bn-2030/>
- Tully, T., Redfern, J., Das, D., & Xiao, D. (2025, December 9). *2025: The state of generative AI in the enterprise*. Menlo Ventures. <https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/>
- Vidal, N. (2024, December 17). *2024 end-of-year review: Open Source AI Definition v1.0* Open Source Initiative. <https://opensource.org/blog/2024-end-of-year-review-open-source-ai-definition-v1-0>
- Wang, J., Wang, J., Athiwaratkun, B., Zhang, C., & Zou, J. (2024). *Mixture-of-Agents enhances large language model capabilities* (arXiv:2406.04692). arXiv. <https://arxiv.org/abs/2406.04692>
- Wang, S., Kahl, J., & Xu, S. (2026, January 30). *Leaders, gainers and unexpected winners in the enterprise AI arms race*. Andreessen Horowitz. <https://a16z.com/leaders-gainers-and-unexpected-winners-in-the-enterprise-ai-arms-race/>
- Wang, S., Xu, S., Kahl, J., & Erten, T. (2025, June 10). *How 100 enterprise CIOs are building and buying gen AI in 2025*. Andreessen Horowitz. <https://a16z.com/ai-enterprise-2025/>
- Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). *MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained Transformers* (arXiv:2002.10957). arXiv. <https://arxiv.org/abs/2002.10957>
- Wang, Y., Ma, S., Shen, B., Shi, S., Wu, Y., Shi, G., & Cheng, X. (2026). *SPAT: A semantic port-aware adaptive-rate transmission protocol for semantic communication* (arXiv:2605.00897). arXiv. <https://arxiv.org/abs/2605.00897>
- Wang, Y., Ma, S., Wu, Y., Shi, G., Cheng, X., Liu, Y., & He, P. (2025). *SITP: A high-reliability semantic information transport protocol without retransmission for semantic communication* (arXiv:2512.09291). arXiv. <https://arxiv.org/abs/2512.09291>
- White, M., Haddad, I., Osborne, C., Liu, X.-Y., Abdelmonsef, A., Varghese, S., & Le Hors, A. (2024). *The Model Openness Framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence* (arXiv:2403.13784). arXiv. <https://arxiv.org/abs/2403.13784>
- Wu, D., Wang, X., Qiao, Y., Wang, Z., Jiang, J., Cui, S., & Wang, F. (2024). *NetLLM: Adapting large language models for networking*. In *Proceedings of the ACM SIGCOMM 2024 Conference* (pp. 661–678). Association for Computing Machinery. <https://doi.org/10.1145/3651890.3672268>
- Zhou, J., & Chen, R. (2025, October 13). *NVIDIA DGX Spark in-depth review: A new standard for local AI inference*. LMSYS Org. <https://www.lmsys.org/blog/2025-10-13-nvidia-dgx-spark/>

About the SPS AI Lab

Founded in 2026, Columbia University's School of Professional Studies (SPS) AI Lab is an innovative and comprehensive resource and community for students, alumni, and faculty. Its mission is to advance and accelerate Applied AI knowledge, adoption and skills to lead in an AI-driven future.

For more information visit:

ai.sps.columbia.edu



COLUMBIA SPS
Artificial Intelligence Lab