My conservative prediction is that by **2028**, around **70%** of enterprises will shift from using API-based LLMs today to **open-source** models that are fine-tuned on company-specific data. This shift will be driven by the growing adoption of open-source models, concerns over data privacy, and the high costs of current proprietary models, which will make them less appealing for long-term use by businesses.

The growing influence of Open-Source LLMs

Academia and research are increasingly using open-source LLMs like **LLaMA** as their foundation for **coursework**, **experimentation**, and innovation. Many courses and programs, including our own **CSE 234**, rely on these models for assignments and lectures because they are accessible, transparent, and flexible. Unlike proprietary black-box APIs, open-source models allow students to fine-tune and play with them, giving them hands-on experience in LLMs. As these students graduate and enter the workforce, their familiarity with open models will naturally influence industry adoption, accelerating the shift toward customizable LLM solutions.

API-based LLMs won't scale for Enterprises

The high cost of proprietary LLM APIs makes them **unsustainable** for large-scale enterprise applications. Pricing scales poorly with usage, making long-term reliance on these services costly. In contrast, open-source models like Mistral and LLaMA allow companies (see Spotify case-study[*]) to fine-tune and run LLMs at a fraction of the cost[†], significantly reducing operational expenses after the initial setup.

A case for data Privacy & Security[‡]

Companies often handle sensitive data in sectors like healthcare, finance, legal, etc., that cannot risk exposing their information to external API-based LLMs. Regulatory **compliance** will further push enterprises to host their own AI models rather than relying on third-party providers. With open-source framework, companies can fine-tune and deploy an open-source model on their proprietary financial data ensuring that customer transaction data never leaves its secure environment. This not only enhances security and compliance but also fosters customer trust and loyalty. As concerns around **data ethics** grow, enterprises will prioritize control over their AI systems more than ever before.

Rebuttals

Some might argue that proprietary models will become **cheaper** to serve as advances in GPU **memory bandwidth**[§] and model **quantization** (4-bit, 8-bit inference) improve efficiency. What if API costs decrease due to **competition**? Additionally, in-house models presents challenges like **data drift**, requiring continuous **maintenance** and **monitoring** to remain effective. Some might even argue that closed models offer a **SOTA performance** as compared to open models.

[*] link: https://research.atspotify.com/2023/10/llark-a-multimodal-foundation-model-for-music/

[†] link: https://www.searchunify.com/blog/open-source-llms-pros-and-cons-for-your-organization-adoption/

[‡] link: https://www.sciencedirect.com/science/article/pii/S266729522400014X

[§] link: https://www.nvidia.com/en-us/data-center/technologies/blackwell-architecture/

Addressing these Rebuttals

IF THEY CAN, the enterprises will always prioritize customer data privacy and **control**, even if in-house LLM deployment comes with maintenance costs. Owning the model allows for deep customization and faster **PoCs** (Proof-of-Concept) development, which is critical for innovation. Even if proprietary models become more efficient, this will likely come at the cost of more GPUs for serving (e.g., maybe through **disaggregation**[¶] techniques), leading vendors to **offload** infrastructure costs onto customers to maintain profitability. While proprietary models may offer SOTA performance, the marginal gains in accuracy or efficiency often come at a disproportionate cost in compute resources. Moreover, for most enterprise applications, task-specific fine-tuning on proprietary data matters more than absolute model size or general benchmarks.

Closed models are here to Stay!

The reason I chose a **conservative estimate of 70%** is that closed, proprietary models won't be completely wiped out. Companies like OpenAI will innovate in serving and inference efficiency, improving latency, throughput, **goodput**, and cost-effectiveness. Their models will remain appealing for enterprise applications that require high performance.

- Not all companies have the resources or expertise to fine-tune and maintain self-hosted models. To host and operate LLMs in-house, companies will need significant **engineering talent** to manage infrastructure, fine-tuning, and ongoing updates. This makes managed API solutions more appealing for certain use cases, especially when rapid deployment and ease of integration are key priorities.
- Even though the open-source landscape looks promising, which open-source model should an enterprise **choose**? With a wide range of models available, selecting the right one becomes a complex task that requires expertise and careful consideration.
- Individual developers and early-stage startups won't have the capital to invest in such prospects.

Moreover, the whole LLM field is still in its early stages, and many organizations have yet to fully realize the transformative potential of these technologies.

[¶] link: https://arxiv.org/pdf/2401.09670v1