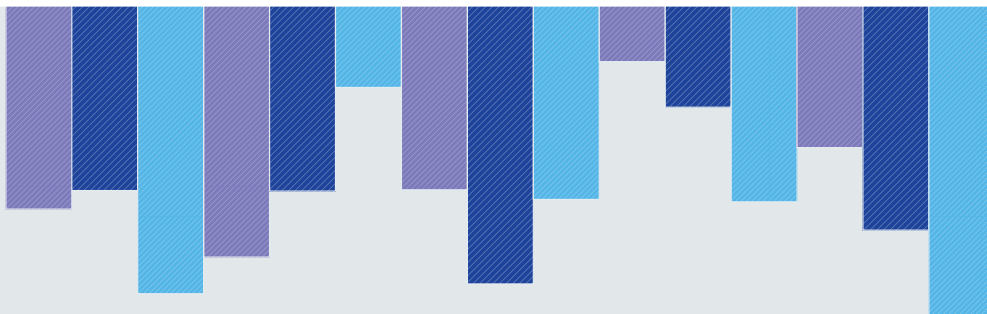


BRIEFING PAPER

Weighing the Open-Source, Hybrid Option for Adopting Generative AI



Sponsored by

CLOUDERA

DELLTechnologies

AMD

SPONSOR PERSPECTIVE

At Cloudera, we recognize the paradigm shift that generative artificial intelligence (AI) technologies are driving. With almost two-thirds of IT decision makers either evaluating or actively using generative AI, there's no denying the disruptive potential of these technologies. From code generation to customer support, the applications continue to expand every day.

However, as underscored by this Harvard Business Review Analytic Services report, this transition isn't without its concerns. Enterprises are grappling with data security, privacy, and the underlying intricacies of using commercial versus open-source large language models (LLMs). The debate surrounding the usage and management of these tools is valid, and these concerns underscore the importance of a tailored approach to adopting generative AI. Whether it's ensuring that sensitive company data remains uncompromised or navigating the intricacies of legal claims against LLM providers, businesses need to be vigilant and agile.

While commercial LLMs offer convenience and superior performance, they might not fit every business's ethical, regulatory, or privacy mold. Conversely, open-source alternatives, though offering transparency and control, demand technical expertise and can entail significant infrastructure costs.

In collaboration with our partners at Dell and AMD, we at Cloudera believe that the answer doesn't lie at the extremes but rather in a nuanced understanding of the technology and its implications. The transformative potential of generative AI is undeniable. However, businesses must adopt a balanced perspective, analyzing their unique needs against the backdrop of evolving LLM capabilities.

This report is a testament to the transformative power and the challenges posed by generative AI. We urge business leaders and technology professionals alike to dive deep, harness the knowledge encapsulated here, and pave the way for a future where AI is not just an operational tool but also a strategic ally.

We at Cloudera remain committed to empowering businesses with the insights and tools they need to navigate this AI era with clarity and confidence. Join us in this journey as we continue to explore and shape the future of generative AI together.



Jake Bengtson
Technical Evangelist
Cloudera, Inc.

Weighing the Open-Source, Hybrid Option for Adopting Generative AI

The vivid interest in ChatGPT, both inside and outside the business world, has put intense pressure on organizations to put generative AI to use in the enterprise. ChatGPT, created by OpenAI Inc., a San Francisco-based nonprofit research laboratory, enables users to interact with the chatbot over the web, powered by a large language model (LLM). With this web-based approach, the LLM is hosted on the public cloud and trained on large volumes of publicly available data to form its responses.

“There’s been a massive amount of focus on the pace of innovation and capabilities of LLMs, which are raising the floor on the kind of work AI can do,” says Bret Greenstein, data and analytics partner at PricewaterhouseCoopers U.S., who notes that the use of generative AI has grown so important to businesses and their stakeholders that a growing number of companies are including their use of AI in their earnings reports. “People are buying into this, in part for what it does today but also what it will do tomorrow.”

According to Enterprise Technology Research, a New York-based technology market research firm, 66% of the 1,777 senior information technology decision makers surveyed in July 2023 said they were either actively evaluating or in production mode with generative AI. The most common use cases among respondents were code generation and documentation, customer support, writing content and marketing copy, and text and data summarization.¹

Numerous commercial and open-source LLMs have become available in recent months. However, as businesses continue experimenting with a web-based approach to accessing commercial LLMs, it is increasingly apparent that sensitive company information can be exposed and that there is little

HIGHLIGHTS

Businesses are learning they need to **review their own security, privacy, cost, and capability needs** to determine the most suitable way to adopt generative artificial intelligence.

With the **growing availability** of mature and capable open-source large language models (LLMs), there are **several implementation approaches to choose from**.

Running an open-source LLM on the premises **reduces security concerns** of exposing sensitive data and **enables the business to fine-tune the model** themselves.



“There’s been a massive amount of focus on the pace of innovation and capabilities of LLMs, which are raising the floor on the kind of work AI can do,” says Bret Greenstein, data and analytics partner at PricewaterhouseCoopers U.S.

control over the model itself or the content it produces in terms of meeting the business’s own ethical, regulatory, and privacy standards. Further, when LLMs are not fine-tuned or augmented with enterprise-specific or proprietary data sources, they are apt to respond to prompts with an out-of-context or nonfactual response. Lastly, if the business chooses to pay for a license and access a commercial model through an application programming interface (API), the per-use cost of using a commercial LLM could quickly grow, possibly beyond original budget expectations.

Businesses are learning they need to review their own security, privacy, cost, and capability needs to determine the most suitable way to adopt generative AI. With the growing availability of mature and capable open-source LLMs, there are several implementation approaches to choose from. In short, it’s not a matter of whether to bring generative AI into the enterprise but rather how to do it.

“There’s no one-size-fits-all answer,” says Andy Thurai, vice president and principal analyst at Silicon Valley, Calif.-based Constellation Research Inc., a technology research and advisory firm. For this reason, Thurai says, his team advises chief information officers “to look at what you’ve got, what your corporate policies are, what are you trying to achieve, and what the potential use cases are, and then determine what would work for you.” Further, he says, “Don’t choose a tool first and figure out what can you solve with it. The classic problem of when you have a hammer, you will go looking for a nail will apply here.”

This Harvard Business Review Analytic Services report will explore the transformative potential of generative AI, as well as the pros and cons of using commercial LLMs rather than open-source models hosted in the business’s environment, whether that’s on the premises or in the public cloud. The report will also describe alternative ways to create a responsible, reliable generative AI-powered solution, as well as the challenges and benefits of doing so.

Safely Adopting Generative AI

While privacy, data security, and other issues have led some businesses to prohibit employees from accessing commercial LLMs over the web, imposing such constraints is not the final answer. As with social media and mobile computing,

employees have been known to find their way around such prohibitions. Meanwhile, senior executives are under pressure to determine how they can best realize the productivity and innovation advances that generative AI promises to deliver.

“Everyone wants to be more effective at their job and perform better, so a lot of people are doing their own experiments,” says David Greenfield, cofounder of GenAI Partners, a consultancy that helps evaluate, plan, and build generative AI-enabled solutions. “It’s both a benefit and a cause for strife because of the lack of controls that type of usage has.”

A renowned example of the privacy and security risks organizations run when it comes to interacting with commercial LLMs over the web is Samsung Electronics, whose employees put confidential source code into ChatGPT to debug it, as well as transcripts of internal meetings to summarize them.² The issue, Greenfield says, is that this sensitive data would be used to train the model and could surface in a response to someone else’s—even a competitor’s—prompt.

Such a possible scenario is cause for concern, given that 43% of business professionals have used AI tools, including ChatGPT, for work-related tasks, according to a January 2023 survey by Fishbowl, a social media platform for that constituency. Nearly 70% of the 11,700 professionals who responded to the survey on the Fishbowl app say they are using those AI tools without their boss’s knowledge.³

Security fears have led a growing number of companies to ban, place limits on, monitor the use of, or block employees from using commercial LLMs through a web interface. Doing so is only partially effective, however, because the use of ChatGPT has become so prevalent. Says Constellation’s Thurai, “Regardless of the solution you put in place, people will find a way to use it, whether on their home computers or phones.”

Commercial LLM providers have introduced controls that allow users to prevent the data contained in their prompts from training the model. But organizations still need to fortify those controls with policies, best practices, and guidelines for activating the settings, as well as training on acceptable forms of content to include in a prompt.

Even these controls may not be enough for highly regulated businesses that deal with confidential information. “Businesses using it to work on a marketing campaign may be satisfied with these settings, but if you’re doing analytics

on anonymous patient health records or other information that could lead to significant business harm, you wouldn't be as comfortable," Greenfield says, adding that the commercial providers are continuing to release controls aimed at securing enterprise use.

Generative AI as Subject-Matter Expert

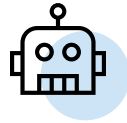
Another important consideration for adopting generative AI is how to enable the model to respond to prompts within the context of the company's own proprietary data. Most commercial LLMs are trained with generic web-scale data, which makes them highly adept at responding to general-purpose types of questions. For business use, however, the model needs to act almost like a subject-matter expert, providing guidance on, for instance, how to assemble a transformer at a specific location or detailing the process for reimbursing an unhappy customer.

To enable these more-contextually relevant responses, the models either need to be fine-tuned with business data sets or the business data needs to be made available to the model during the prompt. One technique for making the data available to the model is "prompt engineering," in which the query is injected with additional data that steers the model's response. An example of this approach is connecting the LLM to a business database that sits outside the model and contains relevant content that is retrieved and passed to the LLM. Often, these databases are "vector" databases, which can store and retrieve unstructured data and are optimized to work with LLMs for fast, easy searches and data retrieval.

For a price, commercial LLM providers are beginning to offer fine-tuning services, which involve retraining the model on specific data sets to adapt it to the specific business context, and these pricing models will vary, according to Thurai. Because fine-tuning services from the commercial LLM providers are just starting to appear, it will take time for them to reach a high stage of maturity or be available on the latest models, Greenfield adds.

For organizations choosing to make business-specific data available to the LLM for fine-tuning or through prompts or vector databases, the data security concerns of doing so could be addressed by using a commercial LLM hosted in a virtual private cloud inside the company's firewall, which the leading cloud vendors now offer. "Each of the cloud vendors has the ability to host an LLM in a virtual private cloud," Greenstein says.

As he explains, this setup should alleviate most of the pressure on data security. "Everything you can do on a public-cloud instance, you can do on a private-cloud instance," Greenstein asserts. "There's a cost to it, but you don't lose capability, and you gain security and trust and the ability to manage the model, monitor it, and extend it."



Another important consideration for adopting generative AI is how to enable the model to respond to prompts within the context of the company's own proprietary data.

Many businesses, however, may opt for another equally viable option—and one that is increasingly in the spotlight: open-source LLMs. Unlike commercial LLMs, open-source LLM providers make the model's architecture, parameters, and training data available. Particularly with Meta's announcement that its latest version of LLaMa would be released as open source—which was widely seen as in direct competition with popular commercial LLMs—the open-source approach is gaining recognition.

"Meta's offering is the first that is fully open-sourced and free to use commercially, truly democratizing AI foundational models," Thurai says. "It is easier to retrain and fine-tune these models at a much cheaper cost than massive LLMs."

Model repositories such as Hugging Face enable businesses to download an open-source, pretrained LLM and run it on their own systems and platforms. Additionally, organizations can choose to deploy the open-source LLM on a hybrid platform that involves both public cloud and on-premises, private systems. In this case, the organization could use the public cloud for quick experimentation at a low cost to prove concepts, and then move to the on-premises system to deploy the model with their data secure. Running an open-source LLM on the premises reduces security concerns about exposing sensitive data and enables the business to fine-tune the model themselves. In addition to fine-tuning, "there are other aspects you might want to control, too—like system availability and latency—that the commercial vendors may not give you the opportunity to do," Greenfield says. "By having a dedicated infrastructure and your own model, it gives you more control over all these technical aspects."

Performance is another reason on-premises solutions may provide an advantage, as long as the organization has access to enough graphics processing units (GPUs) and compute capacity, Greenstein explains. "At the moment, clouds are shared among a lot of people, and there's a huge amount of demand. So performance remains a characteristic that might lend itself to an on-premises implementation," he says. This approach is particularly relevant for businesses creating a

product based on generative AI intended for use by millions of users versus an enterprise creating a tool for just thousands of people to use, he adds.

According to Enterprise Technology Research, the percentage (32%) of respondents planning to host a generative AI model on the premises (private infrastructure) is equal to that of those choosing the cloud (public infrastructure). **FIGURE 1**

The Open-Source Choice

According to Greenstein, capability and performance, not data security, are the two main reasons to opt for an open-source LLM approach because hosting an LLM in a virtual private cloud is as secure as doing so on-premises. “The choice of an open-source or commercial LLM is going to entirely depend on whether the open-source models advance in capability faster than the commercial models,” he says.

The commercial models currently exceed open-source models on most benchmarks, but the gap is shrinking. “But if that ever flips, it will be a very interesting discussion,” he says. A reversal in the status quo could start to tip the balance in favor of open-source models, indicating they would have an edge on commercial LLMs in terms of performance and capability.

Ashu Garg, general partner, and Jaya Gupta, partner, at Foundation Capital, an early-stage venture capital firm based in Palo Alto, Calif., wrote in an August 2023 blog that the choice between open-source and proprietary models isn’t binary. “You may find value in a hybrid approach that leverages the strengths of both” open-source and commercial models, they wrote.⁴

According to Garg and Gupta, the main advantage of commercial LLMs is their impressive performance across



“The companies training their own models are in the top regulated sectors, where developers typically execute with an excess of caution, as regulations may change,” says David Greenfield, cofounder of GenAI Partners.

a wide variety of tasks. “They’re also easy to use and come with managed infrastructure, which allows you to get up and running quickly,” they wrote in their blog post.

Open-source models, meanwhile, offer code that anyone can access and modify. “They are quickly catching up to their proprietary counterparts, as the performance of Meta’s LLaMa 2 attests,” Garg and Gupta wrote. Open-source LLMs’ transparency and flexibility may be considered critical in highly regulated industries, they added.

According to Garg and Gupta, open-source and commercial LLMs present pros and cons that businesses will need to carefully weigh against their own specific requirements. For example, while open-source models enable greater control over data and privacy, as well as an array of model sizes that could result in faster speed, the technical expertise and costly talent and infrastructure costs open-source impose put them at a disadvantage against commercial LLMs.⁵ **FIGURE 2**

Another reason organizations might opt for an open-source/on-premises approach, Greenfield explains, is to avoid vendor lock-in. “When they see a technology as so powerful and transformative, they may not want to be overly reliant on something they can’t control and that’s controlled by one large, external commercial party,” he says.

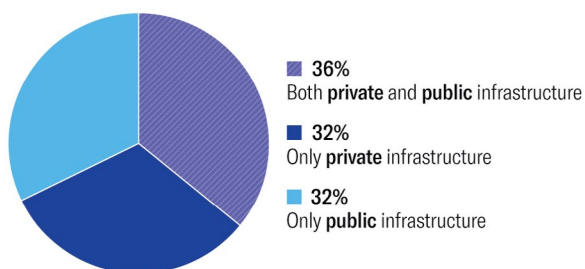
In other cases, a business might view its generative AI implementation as promising so much competitive advantage that it wants to go beyond out-of-the-box functionality. “If it performs 2% better than an out-of-the-box offering, that could start you down the road of producing your own AI model,” he says.

Lastly, organizations in highly regulated industries—such as government, financial services, and health care—might decide to run an open-source, on-premises model to comply with statutes prohibiting data from running on public or virtual private clouds. “The companies training their own models are in the top regulated sectors, where developers

FIGURE 1

Working In and Off the Cloud

Businesses plan to use a mix of private and public infrastructure for generative artificial intelligence (AI)



Source: Enterprise Technology Research survey, July 2023

FIGURE 2

Choosing Your Model

Open-source and proprietary large language models raise different considerations

OPEN-SOURCE MODEL

Advantages	Disadvantages	Considerations
Greater control over data and privacy	Lower quality, but gap is closing	How sensitive is the data?
Local execution	Requires technical expertise and machine learning (ML) teams	What level of control over the model's training is needed?
Depending on use cases, models can be smaller and, therefore, faster	Costly infrastructure and talent requirements	Can you hire and retain ML talent for maintenance/support?

PROPRIETARY MODEL

Advantages	Disadvantages	Considerations
High-quality outputs; best performance currently	Data privacy concerns	What is the desired time to market?
Lower barriers to entry	Limited control; model weights/access can change	What are the cost implications?
Quicker to implement	Costly at scale	For my use case, what performance do I need from my outputs?
	Reliability concerns	

Source: Foundational Capital. August 2023

typically execute with an excess of caution, as regulations may change,” Greenfield says.

In Greenstein’s view, there are three key drivers of going the open-source LLM route. One driver is when the company’s culture and data science capabilities are such that it would never want to use a model unless the company had built and fine-tuned the LLM itself. Another is when the business is building a product embedded with generative AI and needs a high degree of control over the training data and fine-tuning of the output. A third is when the business has a unique knowledge base on a specific topic that it can train a model on and then commercialize it. Greenstein contends that some organizations in the finance sector, in particular, have embraced the third driver, and at least one organization he knows of has built a generative AI tool around GPS data.

To make an open-source LLM perform, Greenstein says, “You have to do an enormous amount of fine-tuning and probably some reinforcement learning around it.” When it comes to running the model on-premises, he adds, “you’ll

need to buy a lot of GPUs to do it at scale.” Many enterprises, he believes, may choose to focus on prompting and embedding data and integrating generative AI in their workflows instead of continuously fine-tuning the model.

Greenfield agrees that fine-tuning requires a high degree of effort. “The more you change the model, the more expertise you need,” he says, adding that PhD-level expertise in machine learning would be required for training—not fine-tuning—the foundational model.

Uncertainty Surrounding Commercial Large Language Models

In addition to reliability and control issues, an ongoing concern about commercial LLMs is the growing number of lawsuits that contend the model makers improperly used the plaintiffs’ intellectual property. In June and July 2023 alone, at least five class-action lawsuits were filed against LLM providers, and these were preceded by several others earlier



“[C-level executives] are talking to their corporate governance boards, legal teams, and HR teams to figure out the right use case where they can deploy AI and, if so, what things they need to be careful about,” says Andy Thurai, vice president and principal analyst at Constellation Research Inc.

in the year.⁶ Authors, artists, and software developers claim commercial LLM providers have trained models with their work without giving them credit or payment.⁷

“If the commercial LLM providers win these lawsuits, it changes the whole landscape. But if they lose, it also creates the opportunity for companies to be a little worried about it,” Thurai says. “Almost every C-level executive we speak with is excited and scared at the same time about the possibility of AI. So, they are talking to their corporate governance boards, legal teams, and HR teams to figure out the right use case where they can deploy AI and, if so, what things they need to be careful about.”

A final area causing uncertainty about commercial LLMs is cost. Once businesses adopt a commercial LLM through an API, they are charged on a per-use basis, both for how much text they ask it to process inbound and how much text it generates outbound, Greenfield explains.

“Everybody is going crazy about OpenAI, but what people don’t realize is that after your proof of concept, there is a usage model that is not cheap, and [costs] can add up,” Thurai says. “It can get quite expensive, so companies are looking for alternatives with open-source models.”

One company that Thurai spoke with was spending tens of millions of dollars on creating a generative AI tool based on a commercial LLM and had yet to see a meaningful return on the expenditure. “Some companies have figured out how to monetize their spend on LLMs, and others haven’t,” he says. “The ones that haven’t are burning through cash fairly quickly.”

According to Foundation Capital, commercial models’ consumption-based cost structure can be about 10 times more expensive than their open-source counterparts for some tasks. “For simple, focused tasks, their performance may be more than you need, and smaller open-source models might be more cost-efficient,” Garg and Gupta wrote.⁸

For some businesses, the use case itself will determine whether an open-source, on-premises approach is best. This scenario is particularly true for businesses that want a specialized LLM for a specific use case. Thurai points out that open-source models like LLaMa 2 offer an array of parameters and sizes that are much smaller than those of some of the commercial models. “Though they are smaller, they tend to

be more accurate if they are fine-tuned for a specific task, especially with your own corpus of data,” Thurai says.

“Some models are small enough to fit on an iPad,” Greenstein says. However, most enterprises have 300 to 500 use cases, “not just three,” he points out. “If I had just three use cases, I would do it with open source and run it on my desktop, but an enterprise wouldn’t want to have different models running all over the place. They’d pick a standard model with broad applicability.”

Conclusion


For all the excitement about generative AI, some real choices need to be made when it comes to enterprise implementation—and a key decision is whether to use a commercial LLM or an open-source model.

“C-level executives have the board breathing down their necks, asking when they’re going to adopt generative AI and why [it’s] taking so long,” Thurai says. “Everyone is moving fast,” he says, but he adds that very few are at the production stage of rolling out an enterprise implementation.

What is also moving fast is the open-source and commercial model landscape, with new developments regularly coming to the market. In addition to assessing the capabilities of the open-source and commercial options, businesses will need to first identify how they can best put generative AI to use in their own enterprise and then how much they can afford to invest, their appetite for risk, how much control they need over the LLM and its performance, and whether they want to rely on commercial providers or trust the open-source community for the tools and technologies they need.

“It’s a bit of a race [between] building it yourself with open source and waiting for the commercial providers,” Greenfield says.

He compares the situation to the early days of the cloud, when it took time for fully mature cloud solutions to evolve. “Most businesses are being pressed pretty hard to work with generative AI quickly,” Greenfield continues, “so there’s not a high tolerance to wait three months to satisfy any constraints that commercial models might impose on their specific need for generative AI.”



“Most businesses are being pressed pretty hard to work with generative AI quickly, so there’s not a high tolerance to wait three months to satisfy any constraints that commercial models might impose on their specific need for generative AI,” says Greenfield.

Endnotes

- 1 Enterprise Technology Research, "July 2023 Macro Views Summary," July 2023. <https://etr.ai/featured-research/macro-views-summary/>.
- 2 DeGeurin, Mack, "Oops: Samsung Employees Leaked Confidential Data to ChatGPT," *Gizmodo*, April 6, 2023. <https://gizmodo.com/chatgpt-ai-samsung-employees-leak-data-1850307376>.
- 3 Jackson, Sarah, "Nearly 70% of People Using ChatGPT at Work Haven't Told Their Bosses About It, Survey Finds," *Business Insider*, March 21, 2023. <https://www.businessinsider.com/70-of-people-using-chatgpt-at-work-havent-told-bosses-2023-3>.
- 4 Ashu Garg and Jaya Gupta, "Applying Generative AI to Enterprise Use Cases: A Step-by-Step Guide," *Foundational Capital*, August 4, 2023. <https://foundationcapital.com/applying-generative-ai-to-enterprise-use-cases-a-step-by-step-guide/>.
- 5 Ibid.
- 6 Eric Hanson, Dr. Sylvia Lorenz, Yixin Gong, et al., "AI Legal News Summer Roundup: Edition 1," *White & Case*, July 20, 2023. <https://www.whitecase.com/insight-our-thinking/ai-legal-news-summer-roundup-edition-1>.
- 7 De Vynck, Gerrit, "AI Learned From Their Work. Now They Want Compensation," *Washington Post*, July 16, 2023. <https://www.washingtonpost.com/technology/2023/07/16/ai-programs-training-lawsuits-fair-use/>.
- 8 Garg and Gupta, "Applying Generative AI to Enterprise Use Cases."



Harvard Business Review

ANALYTIC SERVICES

ABOUT US

Harvard Business Review Analytic Services is an independent commercial research unit within Harvard Business Review Group, conducting research and comparative analysis on important management challenges and emerging business opportunities. Seeking to provide business intelligence and peer-group insight, each report is published based on the findings of original quantitative and/or qualitative research and analysis. Quantitative surveys are conducted with the HBR Advisory Council, HBR's global research panel, and qualitative research is conducted with senior business executives and subject matter experts from within and beyond the *Harvard Business Review* author community. Email us at hbranalyticservices@hbr.org.

hbr.org/hbr-analytic-services