Quick Answer: China Perspective — FAQs on ChatGPT and LLMs, 2H23

Published 4 December 2023 - ID G00802478 - 11 min read

By Analyst(s): Ben Yan, Owen Chen, Arnold Gao, Tong Zhang, Mike Fang

Initiatives: Digital Technology Leadership for CIOs in China; Artificial Intelligence; Generative AI Resource Center

CIOs in China can use this research to get quick answers to their questions on ChatGPT and LLMs. We've distilled these questions from client inquiries, media interviews and vendor discussions conducted during the second half of 2023.

Quick Answer

What is Gartner's position on the most frequently asked questions about ChatGPT and large language models (LLMs) from enterprises in China?

- ChatGPT and LLMs have been prominent topics in China since the beginning of 2023. Innovations are evolving rapidly, creating an extremely dynamic market landscape in China. In the second half of 2023, Gartner clients in China shifted from asking general questions about ChatGPT to inquiring more deeply about the exploration and adoption of various types of LLMs.

- This research summarizes our answers to the top questions from enterprises in China during the second half of 2023:
  - How are enterprises adopting LLMs?
  - Should I build or buy?
  - What are the most mature LLM use cases, and what are the emerging use cases?
  - What is the overall enterprise adoption situation in China?
  - Should I fine-tune my own model?
  - How do I compare LLMs?
  - What is the major difference between global LLMs and LLMs in China?
  - What are the key LLM vendors in China?
  - How do I choose between proprietary and open-source LLMs in China?
  - What are the common concerns about LLMs, and how do I mitigate them?
  - How do I control and optimize the cost of proprietary models?
  - How do I improve the observability of the models?
  - What emerging innovations will make a big impact?

More Detail

For this research, we've distilled a set of FAQs on ChatGPT and LLMs from clients, media and vendors, reflecting the latest interests of enterprises in China. Each question has a short answer from our expert analysts.

For audiences who are new to ChatGPT, we highly recommend the global FAQ. See Gartner Addresses Frequently Asked Questions on ChatGPT.

FAQs

1. How Are Enterprises Adopting LLMs?

Enterprises are adopting LLMs in the following ways (listed in ascending order of difficulty):

1. **Deploying LLM-powered applications:** These include embedded enterprise applications (such as Microsoft Copilot) and publicly available LLM applications on the internet (such as ChatGPT):

    - *Enterprise applications* are embedded into enterprise business processes. End users focus only on the functionalities of the applications, without caring about the technical details of the underlying LLMs. This is usually the easiest way for enterprises to leverage LLM capability, especially with the complexity of LLMs. However, most LLM-powered enterprise applications are still new and immature in China, causing enterprises to further explore other options.

    - *Conversational chatbots on the internet* are easy for end users to access, but users need to copy and paste the LLM responses manually into their own daily work environments. Also, the free versions of conversational chatbots have fewer privacy restrictions than the enterprise versions. In the free versions, the LLM vendors may use end users' inputs/prompts for model training/improvements or even share that information with third parties. Because of such security concerns, conversational chatbots on the internet are not the first choice for many enterprises.

2. **Deploying LLMs with prompt engineering:** As an easy start, enterprises can leverage LLM APIs from vendors and arrange internal training programs of basic LLM prompts for their business teams. This approach is a good option when LLM-powered enterprise applications are not mature and internet tools cause security concerns. Enterprises can also leverage LLMs in a deep dive, with advanced prompt engineering and integrations with other IT systems. One popular implementation is retrieval-augmented generation (RAG). This approach enables enterprises to retrieve their internal knowledge/data and feed it into LLMs to generate customized outputs. However, it often requires collaboration across a fusion team of AI, software engineering, IT operations and business members. In addition, the retrieval step can be a daunting task for large-scale adoption.

3. **Fine-tuning models:** Fine-tuning can incorporate additional knowledge into the models or better align the models with human preferences. However, adding new knowledge is still complex .It typically requires a full-parameter fine-tuning approach called “extended pretraining,” leading to significant increases in resources and budget. Lightweight fine-tuning can be an option for some enterprises. For example, fine-tuning open-source models with instruction tuning can change model behaviors (such as making tone adjustments or learning new downstream tasks). However, the use-case scope needs to be limited because the capabilities of many open-source LLMs are currently not as good as those of top proprietary models.

4. **Pretraining LLMs from scratch:** This is not a reasonable choice for most enterprises at the moment, considering the huge investments and uncertain return of investment.

2. Should I Buy or Build?

“Buy or build?” is a permanent question in the IT field. The answer depends on the enterprise’s IT strategy to balance ease of use/lower cost (buy) with more flexibility/control (build). For the majority of enterprises, buy is usually a better choice from an efficiency point of view:

- LLM-powered applications are closer to the buy option. Enterprises do not need to care much about the technical details of an LLM: They buy it; they use it. Meanwhile, just like all the other out-of-the-box tools or software, this option lacks flexibility for deeper customization and increases vendor lock-in.
- LLMs with prompt engineering sit in between the buy and build options. Enterprises need to exert a certain amount of effort to build a solution based on LLMs with prompt engineering. This option joins some flexibility with ease of use.
- Fine-tuning or pretraining the model is closer to the build option. In many enterprises, fine-tuning practices are exploring possibilities rather than obtaining concrete ROI. After fine-tuning the models, enterprises probably still need to apply prompt engineering/RAG to build a useful LLM solution.

3. What Are the Most Mature LLM Use Cases, and What Are the Emerging Use Cases?

The most mature LLM use cases are marketing content generation and virtual customer assistants. Emerging use cases include application development and testing, research companion, and digital workspace support (see Use-Case Prism: Generative AI in China). As LLM development quickly evolves, more use cases are being tested and becoming possible. LLM agents and multimodal LLMs are emerging, with huge potential to drive the next wave of AI adoption.

4. What Is the Overall Enterprise Adoption Situation in China?

The 2024 Gartner CIO and Technology Executive Survey [1] indicates that the majority of enterprise adoption of generative artificial intelligence in China is still in the proof of concept (POC) phase, though a few projects have been deployed into production. Compared with the extreme hype during the summer, the atmosphere is calmer now. People have more realistic expectations for the technology. Organizations are combining LLMs with other AI techniques to build more reliable AI solutions, but the results will take time. In one to two years, we will see more mature applications enabled by generative AI available in China. That will be the moment when generative AI brings more value to the business.

5. Should I Fine-Tune My Own Model?

Fine-tuning models is still not an easy task for most enterprises. Although fine-tuning can customize LLMs to obtain better performance in a given task or domain, it is not always the optimal approach (see Quick Answer: When to Fine-Tune Large Language Models). Prompt engineering is still an efficient way to integrate internal enterprise knowledge with LLMs, especially when you need to control user access to the data or have frequently changing data. Open-source LLMs and lightweight fine-tuning are worth trying only when the team has competence, resources, data and valid business cases.

6. How Do I Compare LLMs?

Evaluation benchmarks for LLMs could provide a starting point for LLM comparison. However, this approach has a flaw: The evaluation questions and answers in the benchmarks have already been included in the models' training data. Models that have already "seen" the evaluation questions and answers tend to have higher scores. In addition, there are critical nonfunctional factors to consider (see Quick Answer: China Perspective — How Do I Compare LLMs?).

Due to their black-box nature, LLMs need to be tested and monitored thoroughly across different metrics. Some of the most important metrics are the custom ones that enterprises devise based on their business scenarios. Domain experts need to prepare specific test cases and run them in collaboration with the fusion team. The human check for business-specific use cases is still indispensable.

7. What Is the Major Difference Between Global LLMs and LLMs in China?

LLMs in China are trained or fine-tuned with a larger Chinese language corpus, thus enhancing model performance in Chinese contexts. However, the top global LLMs have better performance from prompt engineering capabilities (such as steerability — the models consistently follow the instructions of system designers during the entire conversation session with the user). Thus, global LLM users favor the prompt engineering approach more heavily than users in China. As LLMs continuously evolve and prompt engineering receives greater attention from vendors, LLM prompt engineering capabilities will gradually improve.

8. What Are the Key LLM Vendors in China?

Because LLMs in China are still in an early and dynamic phase, it is too soon to tell which vendors are the winners. Hyperscale companies, such as Baidu, Alibaba Group and Tencent, all offer LLMs. However, other open-source LLMs, such as those from Zhipu, Beijing Academy of Artificial Intelligence (BAAI) and Baichuan, are also gaining traction (see Tool: Vendor Identification for AI Foundation Models, China). Organizations must measure the entire LLM solution — including supporting services (such as maintenance), retrieval functions, guardrails and observability — instead of only the LLM itself.

9. How Do I Choose Between Proprietary and Open-Source LLMs in China?

A key driver behind the high interest in open-source LLMs in China is control: Enterprises would have more control over the models. The logic of “more control means less risks” applies here. Such enterprises envision choosing a smaller open-source model, fine-tuning it in a lightweight manner and then hosting it on-premises. Although this is an interesting plan, enterprises must consider other factors, such as:

- Additional effort required
- Additional skills required
- Model governance (such as model protection from prompt injection)
- ROI

Self-managed open-source LLMs could ultimately have a higher total cost of ownership (TCO) than proprietary models and solutions. See Quick Answer: What Are the Pros and Cons of Open-Source Generative AI Models?

10. What Are the Common Concerns About LLMs, and How Do I Mitigate Them?

Loss of confidential data, hallucinations and prompt hacking are the top concerns for clients in China. Some global proprietary models are physically hosted outside China, and enterprises may not be willing to send data (even as prompts) to other countries. This issue is even more serious for enterprises with strict regulatory requirements that do not even use public cloud services. In these cases, enterprises need to find domestically deployed models and/or deploy models on-premises.

11. How Do I Control and Optimize the Cost of Proprietary Models?

Currently, pricing for proprietary models is typically based on the consumption of tokens — namely, the length of the text in the prompts and in the completed tasks. Organizations need to manage the length of the inputs and outputs to control the operational cost. Note that current global LLMs usually count every single Chinese character as a token. Hence, processing such languages could generate higher costs than processing English. Additionally, fine-tuned proprietary models could cost much more than standard models. Thus, organizations need to balance the benefits and costs of fine-tuning as well.

Concision is the rule of thumb both for inputs and outputs, and it is a key target for prompt engineering to achieve. Other approaches include:

- Creating templates of “concise” prompts and mapping users’ original prompts to the predesigned prompts to avoid irrelevant and/or unnecessary texts
- Prompt caching, where repetitive prompts are served from cache, thus saving API calls

12. How Do I Improve the Observability of the Models?

AI observability is the ability to manage and assess the behavior of an AI model to gain better understanding and control of the output. LLM adoption requires organizations to set up guardrails to monitor and control models. Such guardrails monitor number of requests, token usage, toxicity, readability of prompts, personally identifiable information (PII) leakage, citation of responses, evaluation of responses, and even leverage of other LLMs to evaluate existing models. Organizations must establish policy and technical mechanisms to improve the observability.

13. What Emerging Innovations Will Make a Big Impact?

More organizations are realizing the limitations of LLMs. Thus, they are shifting from trying everything on LLMs to combining LLMs with other AI techniques (such as composite AI). For example, the LLM is working as a moderator in the AI solution, but it still relies on traditional models to perform the subtask in a more efficient way. More mature LLM-enabled enterprise applications will be available in the next one to two years.

Other innovations are already emerging, including multimodal LLMs, which will reach production environments in two to five years. In addition, autonomous agents have huge potential to take intelligence and automation to another level. However, these will take more time (five to 10 years) to be ready for large-scale adoption.

Evidence

[1] **2024 Gartner CIO and Technology Executive Survey:** This survey was conducted online from 2 May through 27 June 2023 to help CIOs determine how to distribute digital leadership across the enterprise and to identify technology adoption and functional performance trends. Ninety-seven percent of respondents led an information technology function. In total, 2,457 CIOs and technology executives participated, with representation from all geographies, revenue bands and industry sectors (public and private), including 38 from China. Disclaimer: The results of this survey do not represent global findings or the market as a whole, but reflect the sentiments of the respondents and companies surveyed.

Document Revision History

Quick Answer: China Perspective — Frequently Asked Questions on ChatGPT and Large Language Models - 2 March 2023

Recommended by the Authors

Some documents may not be available as part of your current Gartner subscription.

Quick Answer: China Perspective — How Do I Compare LLMs?

Use-Case Prism: Generative AI in China

Hype Cycle for Data, Analytics and AI in China, 2023

Tool: Vendor Identification for AI Foundation Models, China

How to Choose an Approach for Deploying Generative AI

© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. Gartner is a registered trademark of Gartner, Inc. and its affiliates. This publication may not be reproduced or distributed in any form without Gartner's prior written permission. It consists of the opinions of Gartner's research organization, which should not be construed as statements of fact. While the information contained in this publication has been obtained from sources believed to be reliable, Gartner disclaims all warranties as to the accuracy, completeness or adequacy of such information. Although Gartner research may address legal and financial issues, Gartner does not provide legal or investment advice and its research should not be construed or used as such. Your access and use of this publication are governed by Gartner's Usage Policy. Gartner prides itself on its reputation for independence and objectivity. Its research is produced independently by its research organization without input or influence from any third party. For further information, see "Guiding Principles on Independence and Objectivity." Gartner research may not be used as input into or for the training or development of generative artificial intelligence, machine learning, algorithms, software, or related technologies.