



Gartner Webinars

Gartner delivers actionable, objective insight, guidance and tools to enable stronger performance on your organization's mission critical priorities

Gartner®

Enhance your webinar experience



**Ask a
Question**



**Download
Attachments**



**Watch
Again**

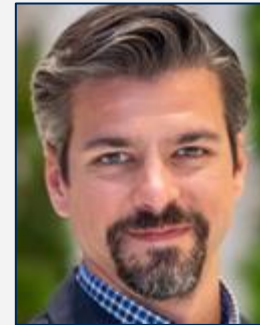
AI Scenarios in Which Small Language Models Outshine Large Language Models

 Connect with Gartner



Birgi Tamersoy

Sr Director Analyst



Radu Miclaus

VP Analyst



Agenda

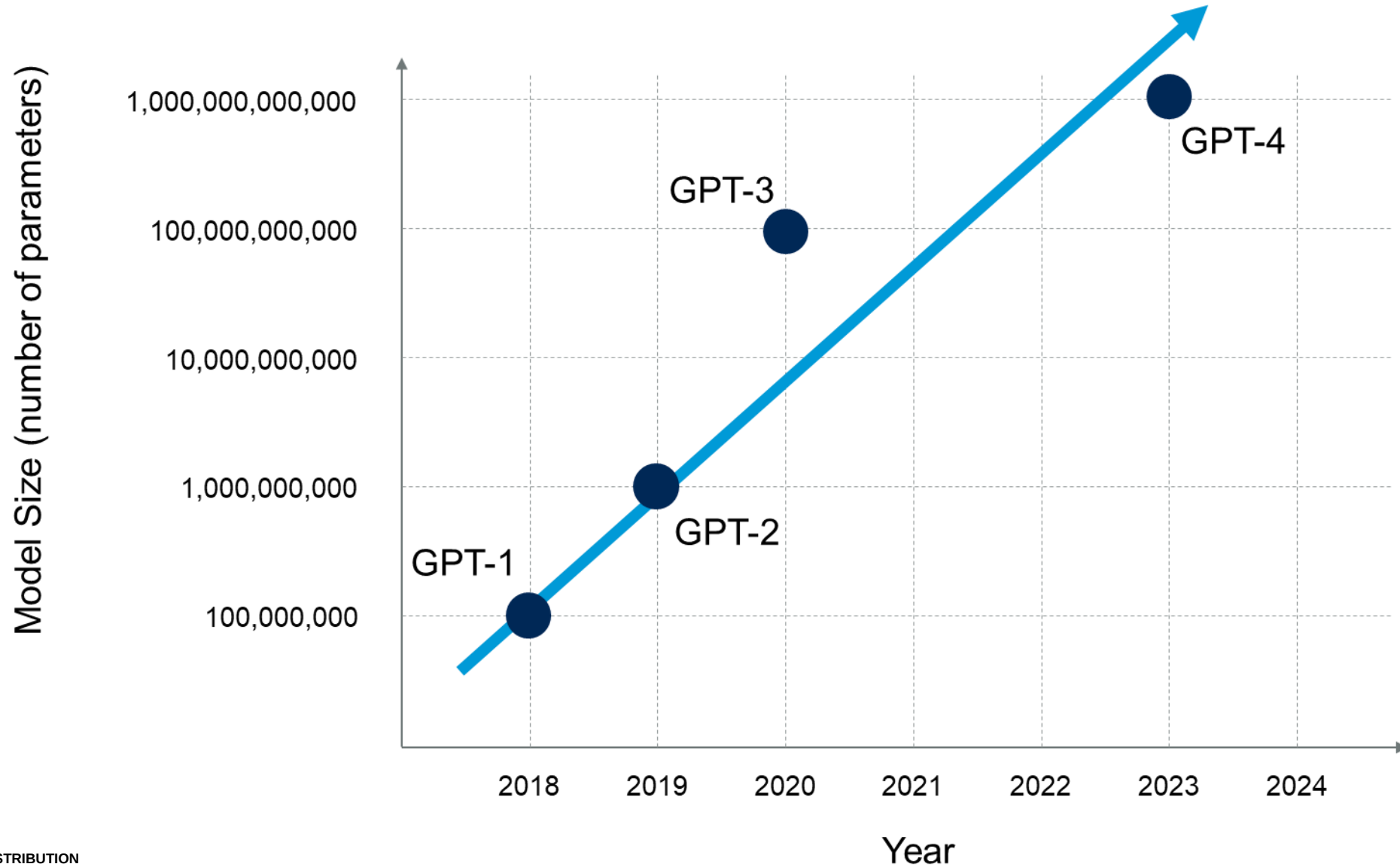
- Introduction
 - What is “Small”?
 - Comparing SLMs to LLMs
- Scenarios in which SLMs outshine LLMs
 - *Edge Experiences*
 - Sensitive Data or Regulatory Restrictions
 - High-User-Interaction Volumes
 - Organizational Language Models
 - Multiple Task-Specialized Models
- Key Findings and Recommendations

Introduction

Scale has been the primary
driving trend in the development
of large language models!



Scale Has Been the Driving Trend of LLMs

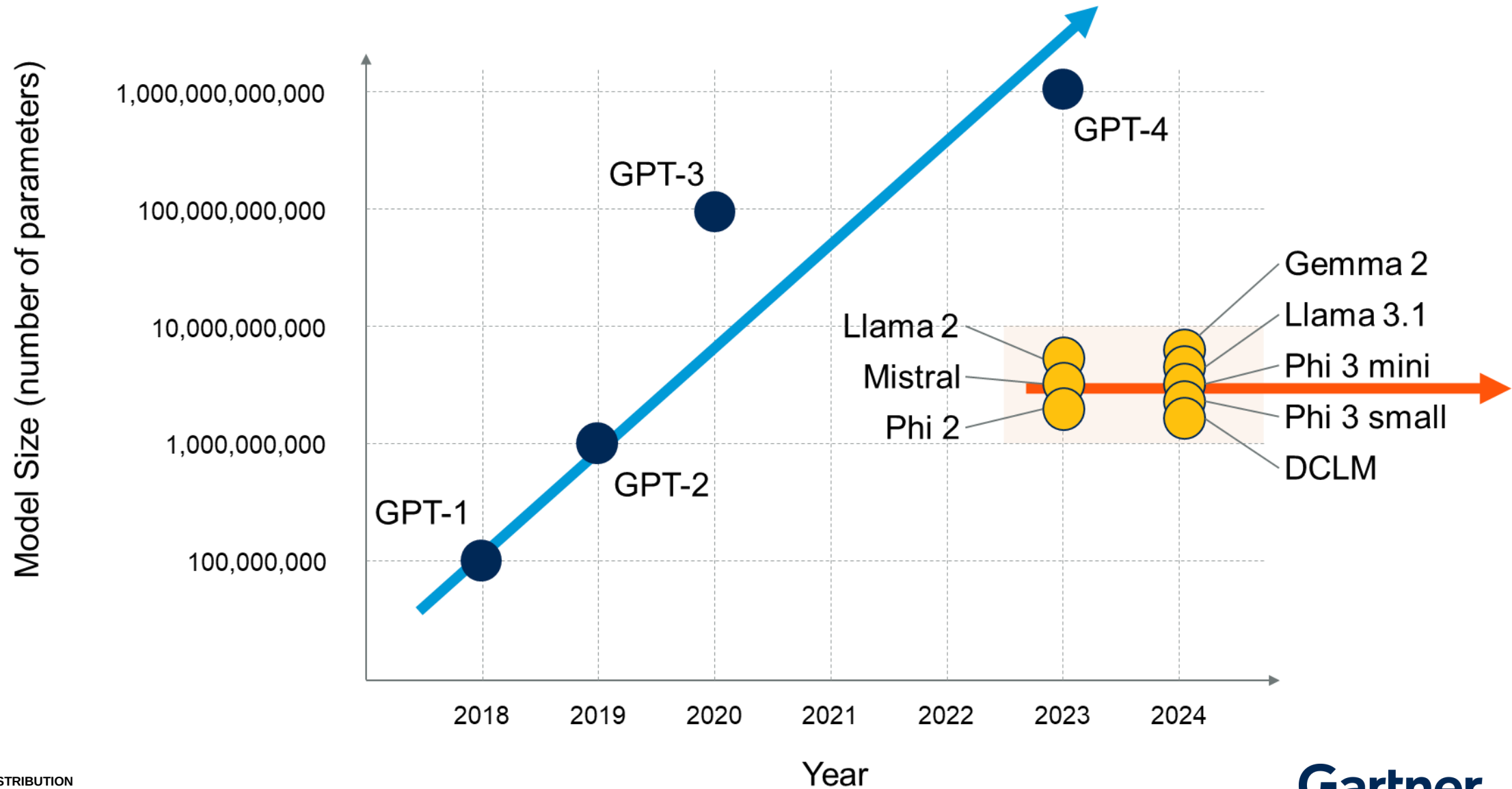


RESTRICTED DISTRIBUTION

Recently a contrarian trend has emerged: development of small language models!



A Contrarian View: Is Small Good Enough?



Pros and Cons of Language Model Size

LLM

- + Impressive *generic* natural language understanding and generation capabilities.
- + Impressive *in-context learning* capabilities. State-of-the-art results can be achieved with prompt engineering.
- May be costly to serve and customize for a particular domain or task.

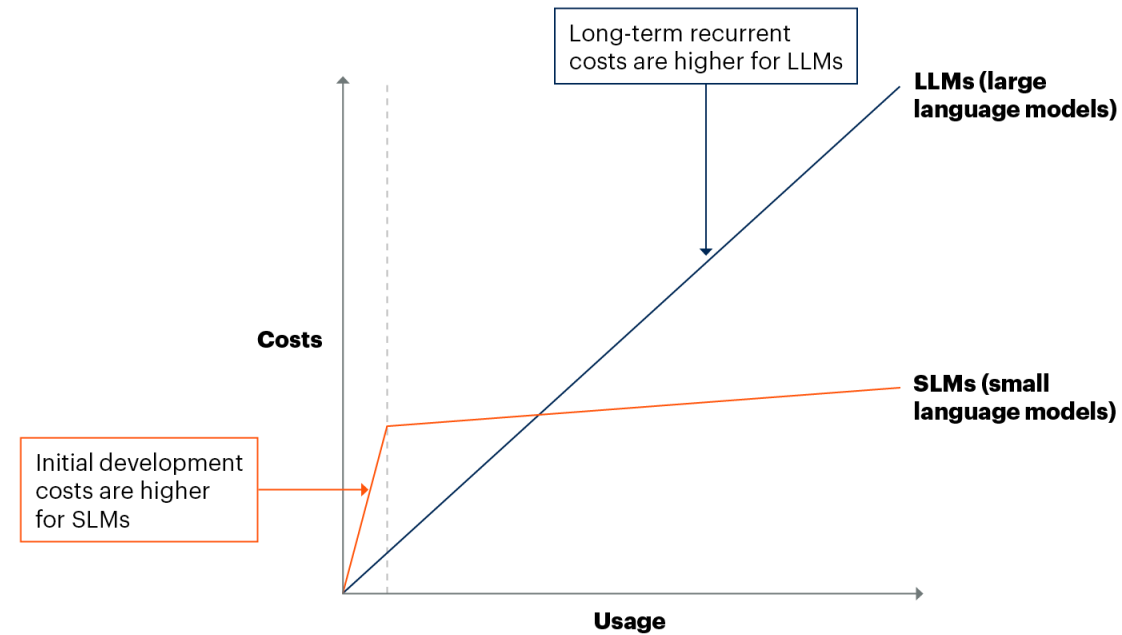


SLM

- + Cost-effective to serve on-device, on-premises, on a private cloud or on a public cloud.
- + Makes domain or task specialization technically and economically feasible through better aligned models, which are easier to control in terms of inputs and outputs.
- May require customization to achieve a particular performance level, which requires more investment compared to prompt engineering.

SLM vs. LLM Costs

SLM vs. LLM Costs
Illustrative



Source: Gartner

Note: The dotted line represents the end of development and the beginning of deployment for SLMs.

818216_C

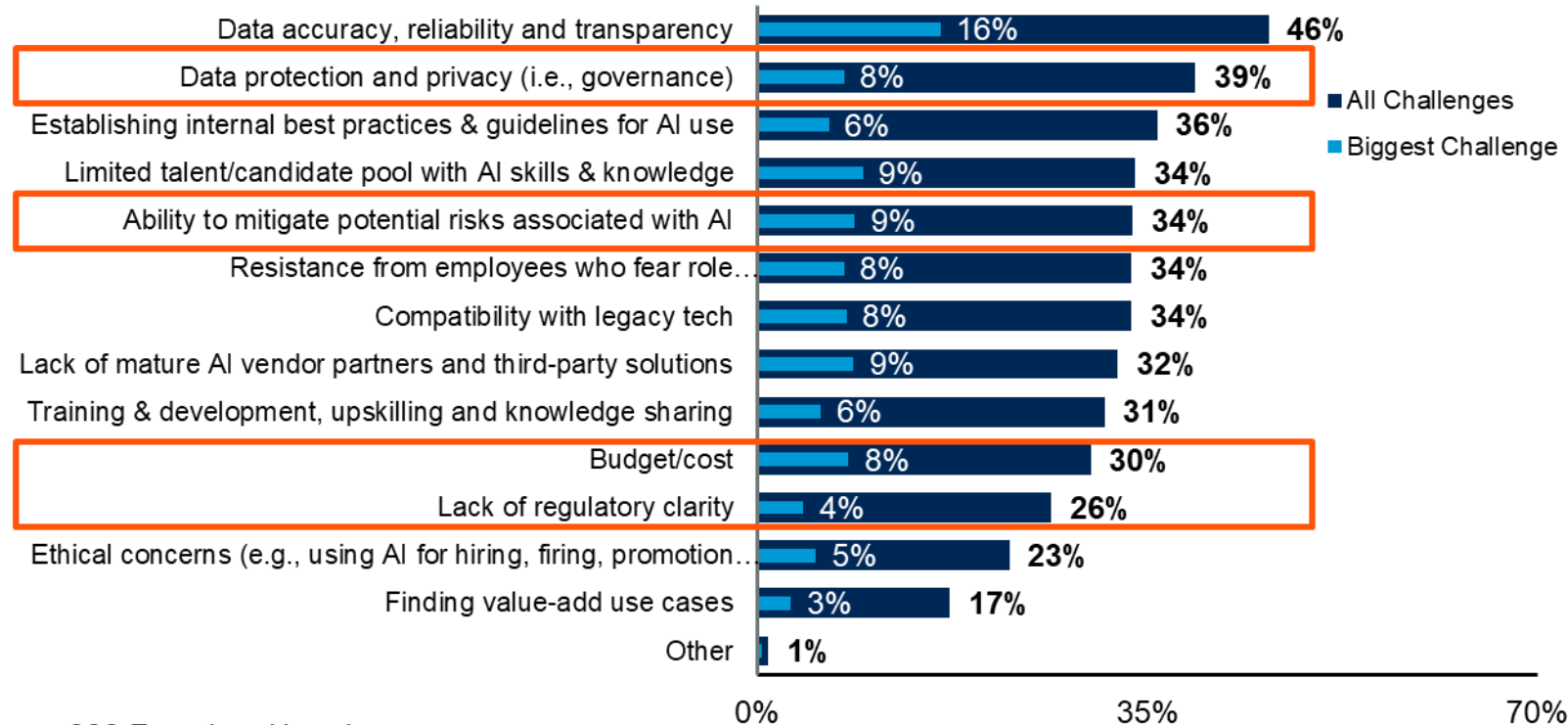
Gartner.

Gartner®

SLMs Address Key GenAI Challenges

Implementation Challenges From Generative AI

Multiple responses



n = 822 Functional Leaders

Q13a: In your understanding, what are some potential challenges with the implementation of Generative AI in your organization?

Q13b: What is the biggest challenge with the implementation of Generative AI in your organization?

Source: Gartner Generative AI 2024 Planning Survey

Footnotes

ID:



Emerging Tech: Enhance Edge Experiences With GenAI



Summary

Light GenAI models are creating new opportunities in use cases across industry and consumer segments. This research helps product leaders identify use cases that will deliver the highest benefit from GenAI and extend their offerings to differentiate and increase customer satisfaction.

Impacts

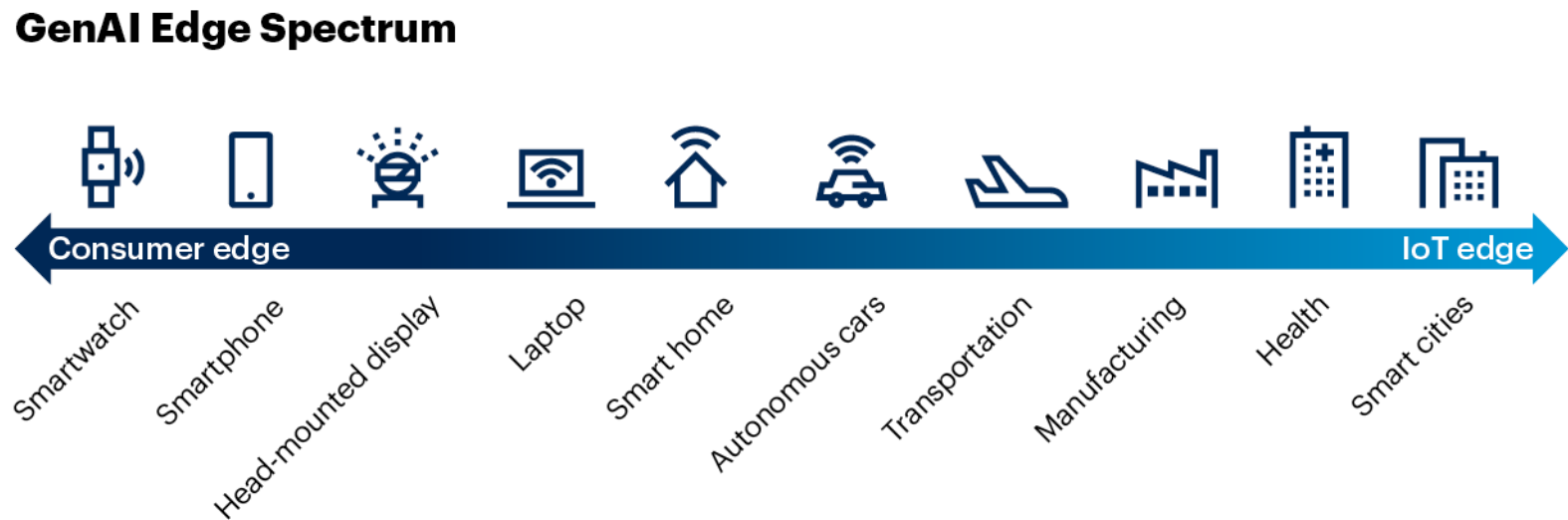
- Advances in AI inference chips and improved light generative AI (GenAI) models will enable an explosion in edge AI use cases.
- Consumer technology will initially use GenAI to enhance user interfacing in areas such as health, productivity, entertainment and mobility, with fitness and health monitoring being most impacted.
- Business technology edge cases will use GenAI to augment experiences in manufacturing, healthcare, transportation and smart cities.

Recommendations

Technology product leaders should consider the following when planning for edge GenAI use cases:

- Increase the performance of GenAI applications and reduce operational costs by offloading models into edge hardware whenever practical.
- Improve the consumer experience by creating a GenAI interface platform that can be integrated across new and existing applications.
- Track GenAI adoption at the edge by monitoring the development of applications in manufacturing, healthcare, transportation and smart cities, as these industries will likely be early adopters.

Figure 1: GenAI Edge Spectrum

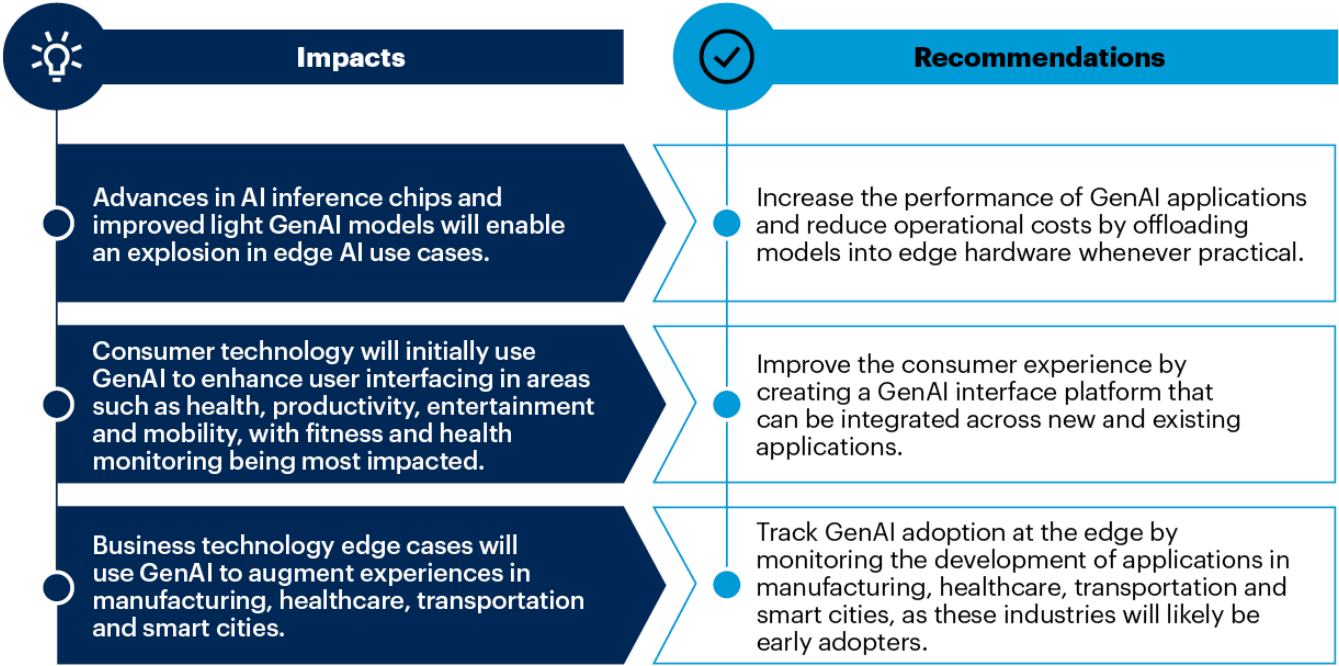


Source: Gartner
811201_C

Gartner.

Figure 2: Impacts and Top Recommendations for GenAI at the Edge

Impacts and Top Recommendations for GenAI at the Edge



Source: Gartner
811201_C

Gartner

Figure 3: Consumer Edge Use Cases for GenAI

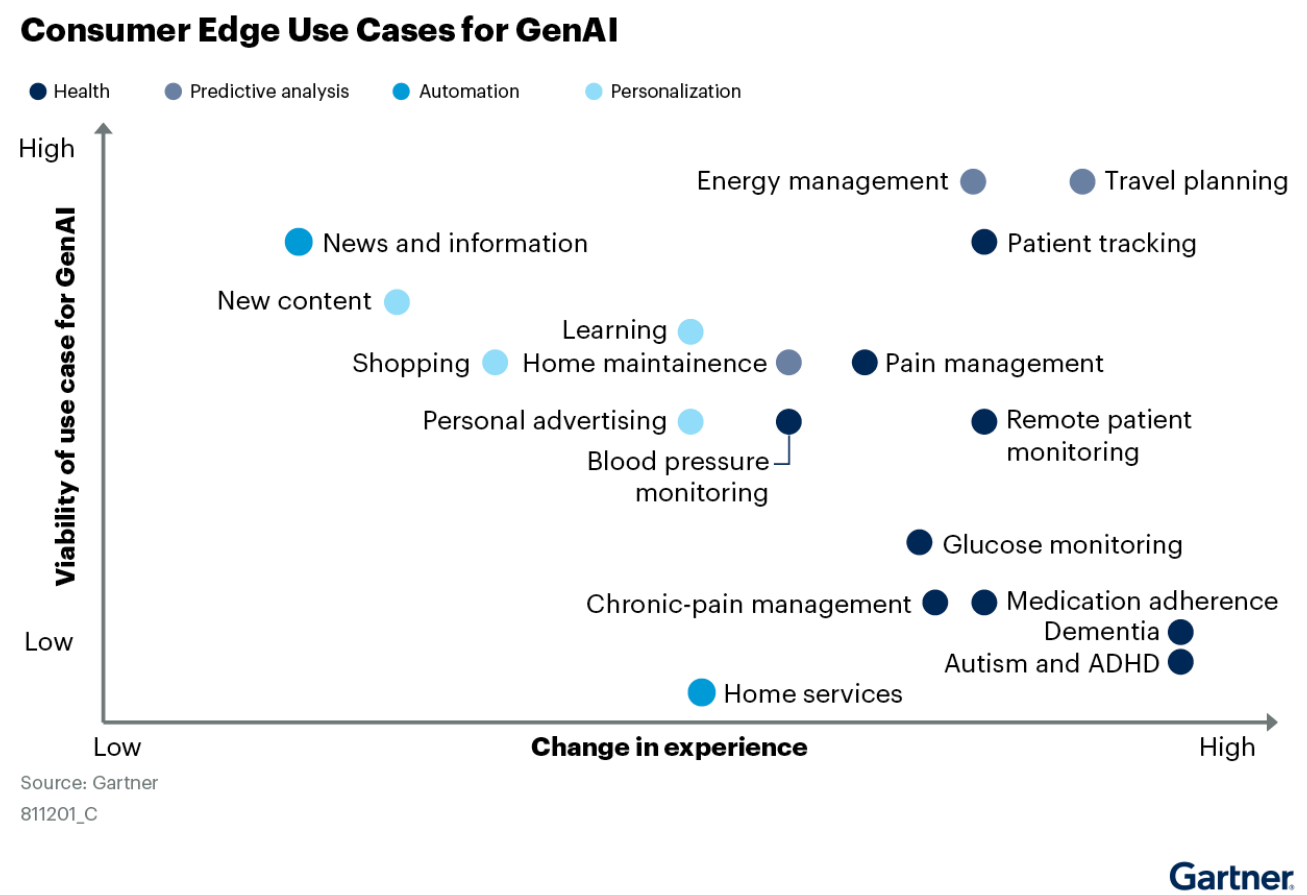
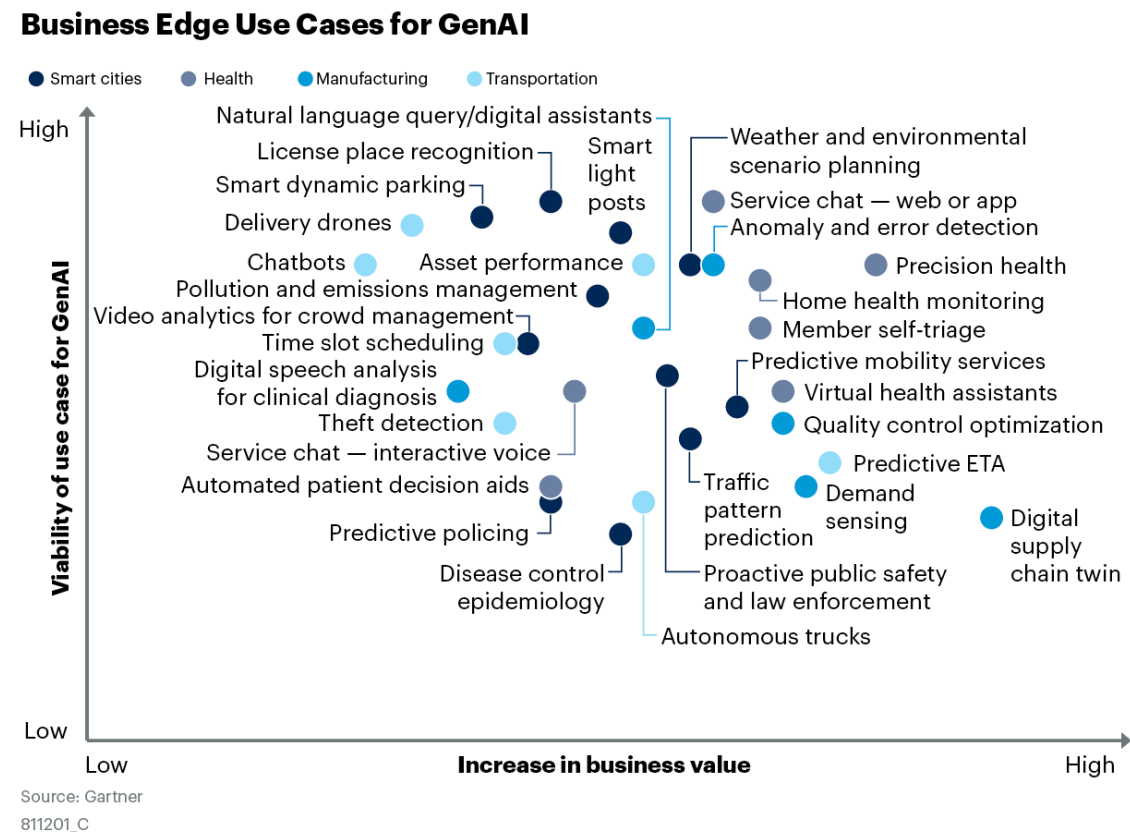


Figure 4: Business Edge Use Cases for GenAI



Gartner

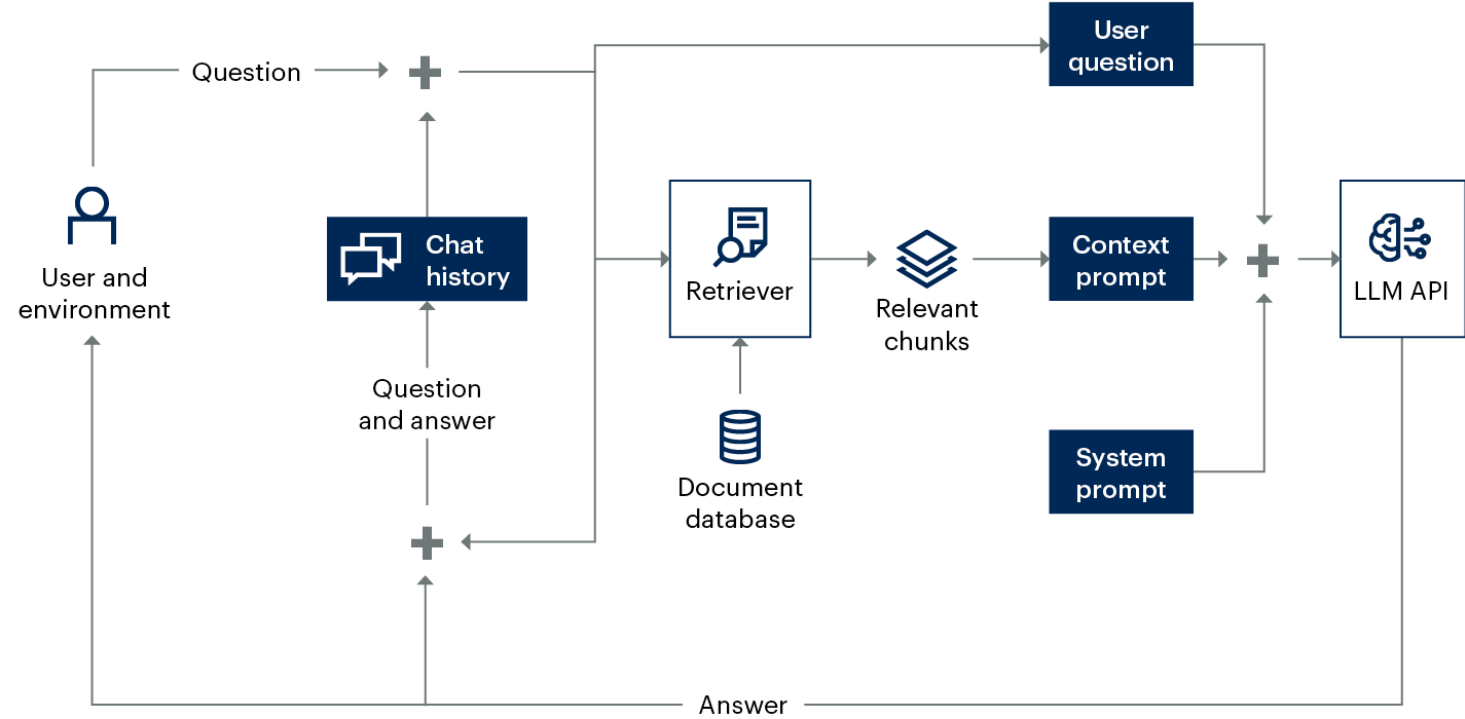


Other Common Scenarios



Document Chatbot

Document Chatbot



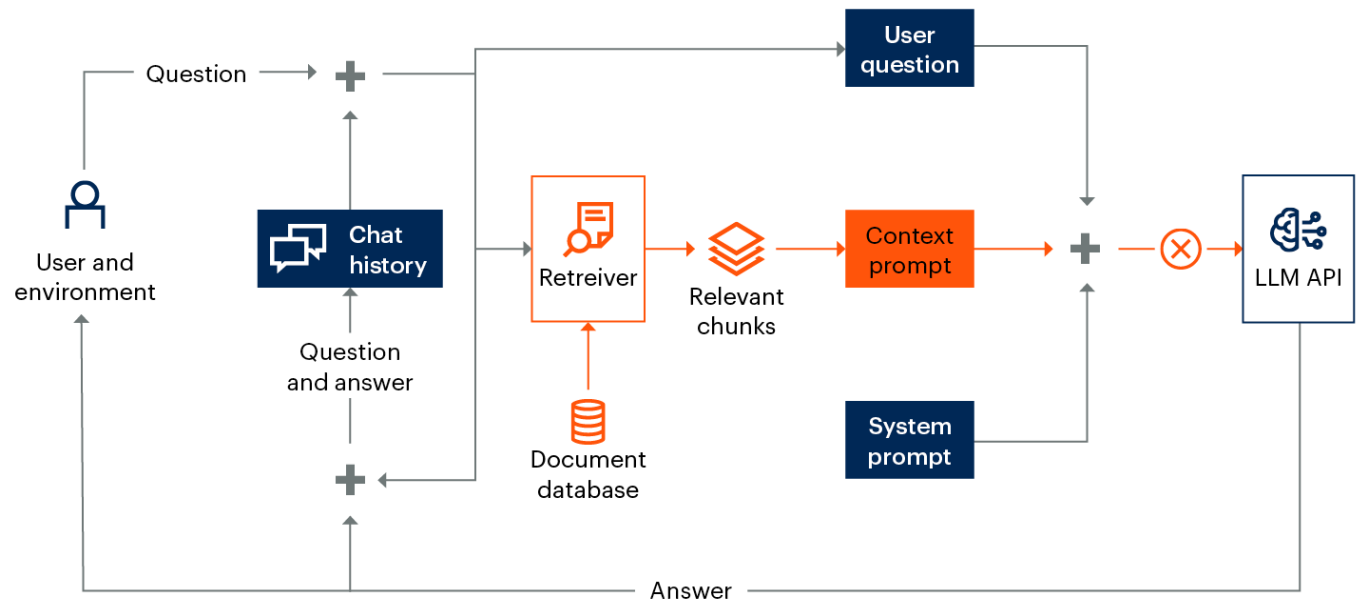
Source: Gartner
818216_C

Gartner.

Gartner®

Sensitive Data or Regulatory Restrictions

Sensitive Data or Regulatory Restrictions



Source: Gartner
818216_C

Gartner

Implementation Challenge With LLMs

Organizational data is an important asset! Sharing parts of this asset with third-party cloud providers through LLM APIs is not always desirable.

Local laws and regulations may also prohibit sharing of internal information for highly sensitive use cases.

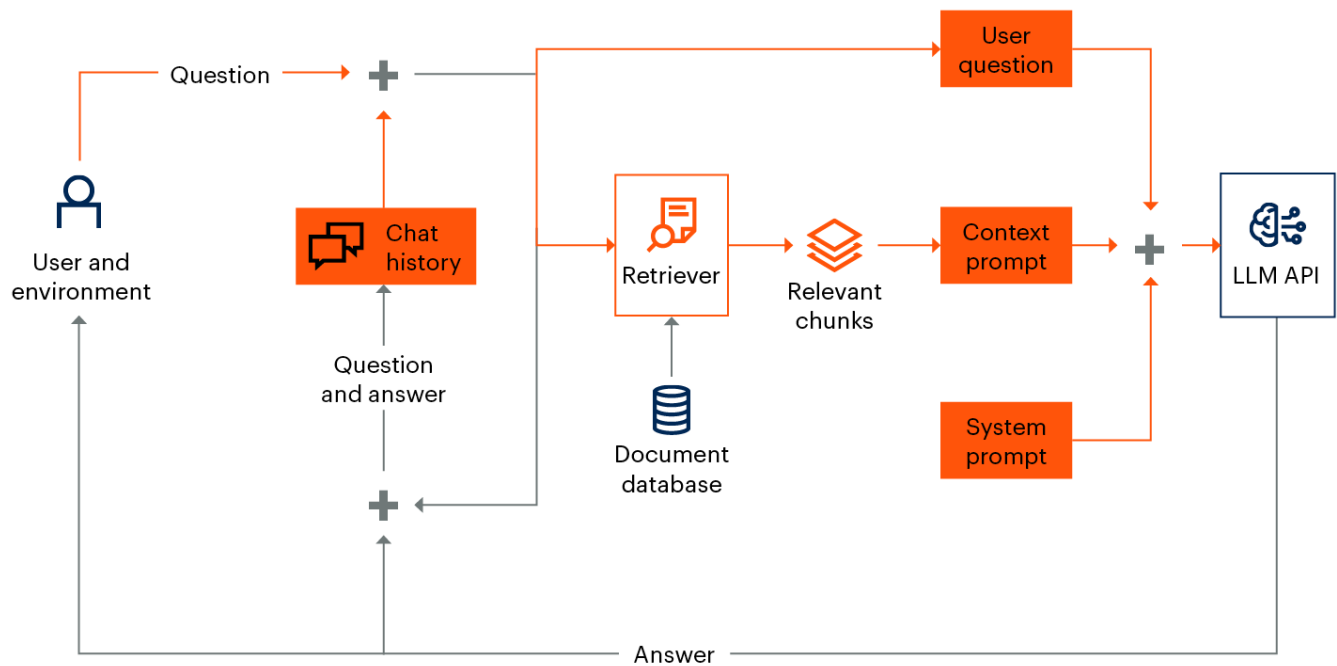
How SLMs Can Help?

SLMs require significantly fewer computational resources for deployment.

SLMs make it possible to have on-premises or on a private cloud deployment, eliminating major risks and concerns regarding privacy and security.

High-User-Interaction Volumes

High user interaction volumes



Source: Gartner
818216_C

Gartner

Implementation Challenge With LLMs

Enterprises may have large employee and customer bases.

Third-party LLMs use a token-based charging model. These costs can get significant (\$1M+/year) with high user interaction volumes.

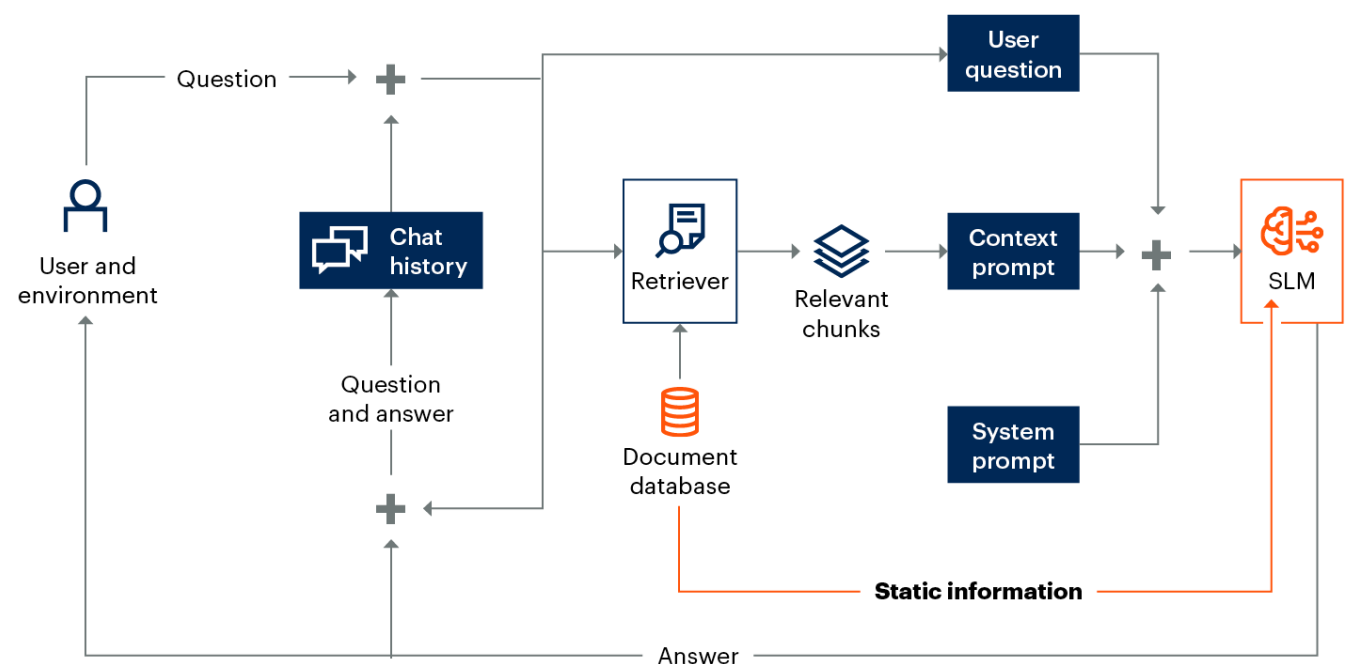
How SLMs Can Help?

SLMs may incur more initial investment and require more skills in development, as they benefit from task specialization.

However, due to their serving efficiency, SLMs require fewer recurring costs in deployment.

Organizational Language Models

Embedding Static Knowledge Into the Language Model



Source: Gartner
818216_C

Gartner

Implementation Challenge With LLMs

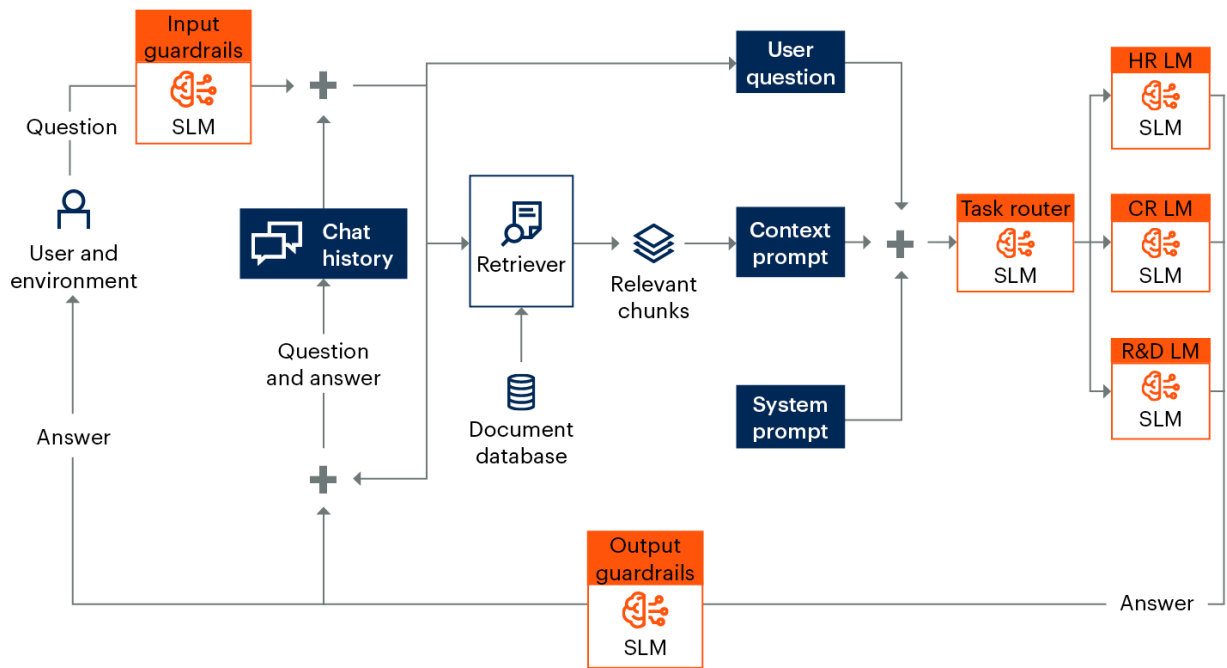
When using third-party LLMs, organizational context needs to be provided with *each* interaction so that the LLM can answer the user's question with correct organizational information.

How SLMs Can Help?

SLMs make customizations both technically and economically feasible for most organizations. Static organizational information can directly be embedded into SLMs, creating language models that are better aligned for the enterprises.

Multiple Task-Specialized Models

Multiple Task-Specialized Models



Source: Gartner
HR LM = human resources language model; CR LM = customer relationship language model; R&D LM = research and development language model
818216_C

Implementation Challenge With LLMs

Using a generic LLM for multiple-tasks is costly to serve and may create more hallucinations in deployment. Avoiding hallucinations makes using the GenAI solution more complex for users.

How SLMs Can Help?

Task specialization has the benefit of achieving high accuracy and robustness with low inference costs. SLMs allow for building and deploying large numbers of task-specialized language models.

Key Findings and Recommendations

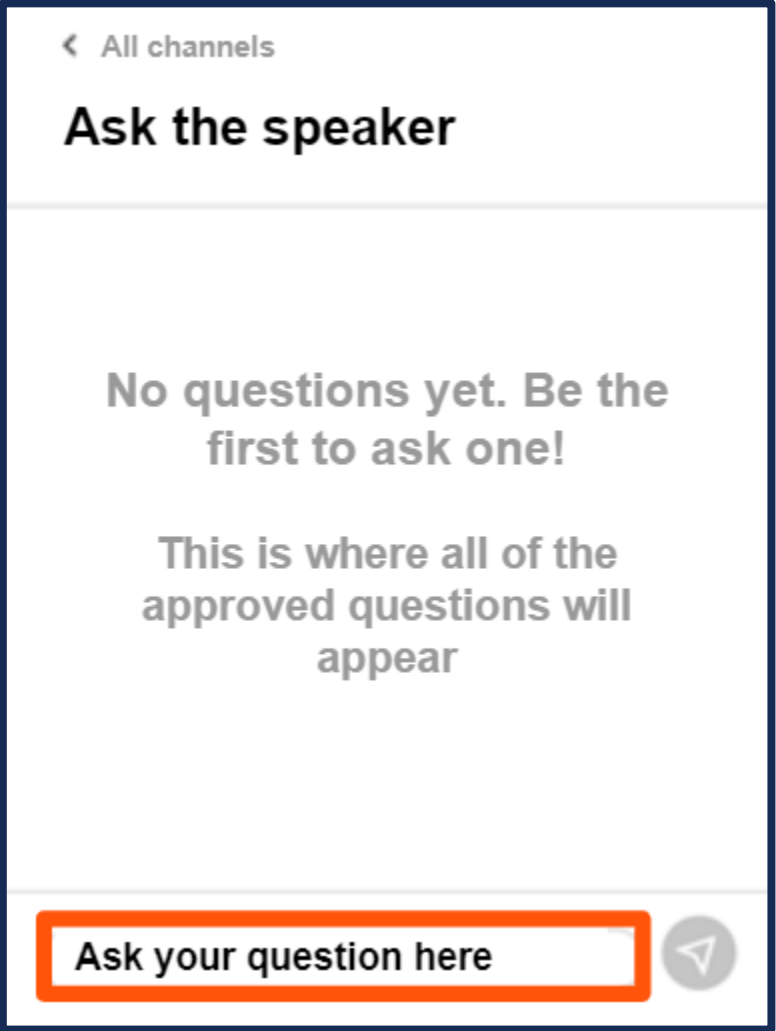
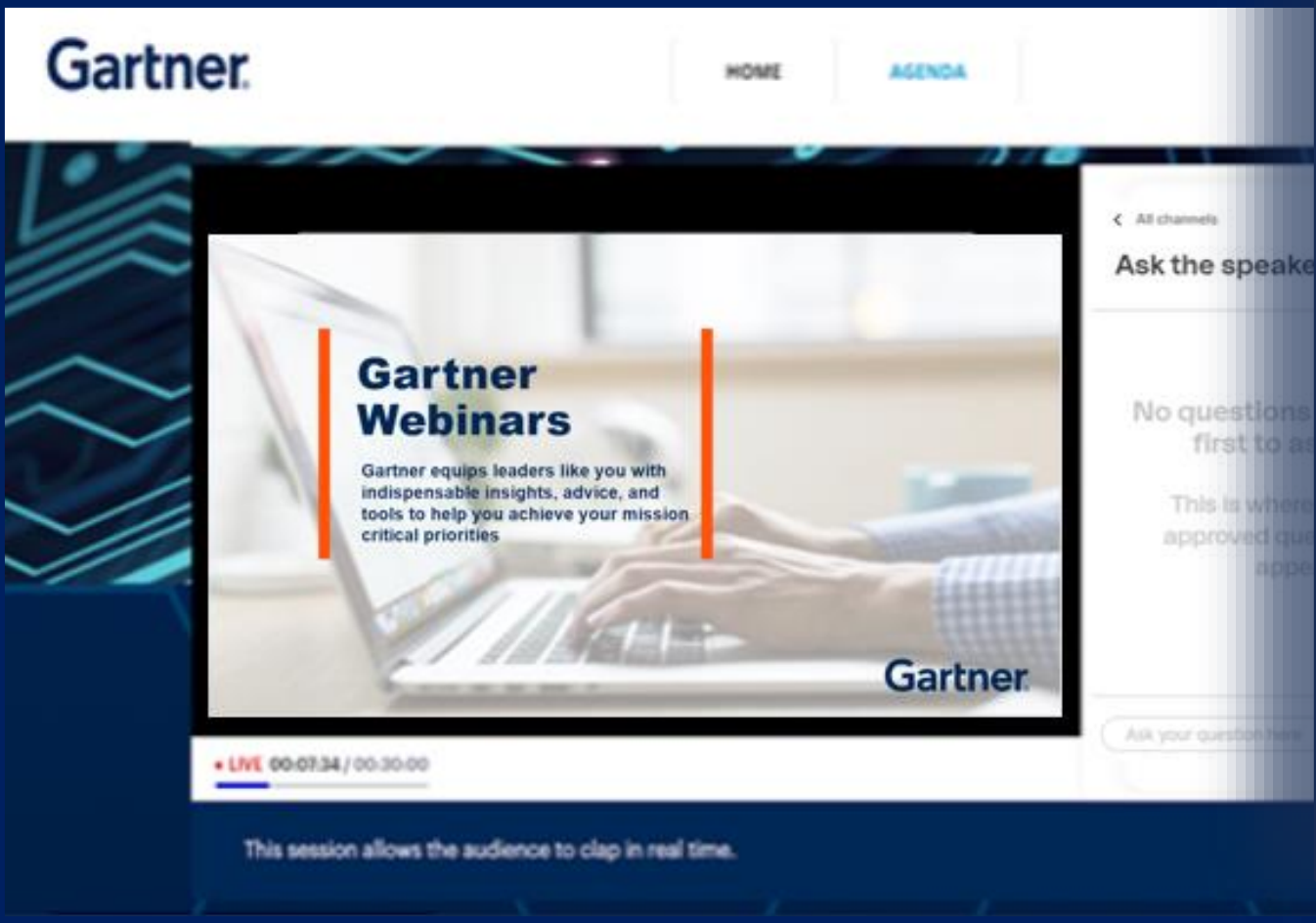
Key Findings

- Large language models (LLMs) provide impressive generic natural language understanding and generation capabilities. However, most organizations do not require these broad capabilities in their generative AI (GenAI) solutions, as their task scopes are more constrained and well-defined.
- Fine-tuning LLMs is costly in terms of the required computation and the quantity of data needed. Deploying LLMs is also costly in terms of computational requirements and carries risks related to output controls.
- Small language models (SLMs) offer a potentially cost-effective alternative for GenAI development and deployment for most organizations and use cases. They are easier to fine-tune, more efficient to serve and more straightforward to control.

Recommendations

- Use on-premises or private cloud SLMs for scenarios involving sensitive information. This approach reduces privacy and safety concerns, limits third-party dependencies regarding trust, and eases regulatory compliance.
- Customize SLMs to increase accuracy, robustness and reliability. Task specialization creates better-aligned models, which are easier to control in terms of input and output.
- Embed your static organizational knowledge into SLMs for reduced costs and increased efficiency. Dynamic information may still be provided in context. This hybrid approach is both efficient and effective.

Ask the speaker



Gartner IT Symposium | XPO

9 - 11 September 2024 | Gold Coast, Australia
21 – 24 October 2024 | Orlando, FL
28 – 30 October 2024 | Tokyo, Japan
4 – 7 November 2024 | Barcelona, Spain
11 - 13 November 2024 | Kochi, India

In an era of continual disruption, the role of CIO is evolving rapidly. CIOs must amplify their impact to lead IT beyond the function, partnering with C-suite peers to accelerate digital business models, enable the future of work and drive business growth. Join us at our CIO conferences to discover world-class insights to help you drive your mission-critical priorities.

Learn more: gartner.com/conf/cio

#GartnerSYM

The World's Most Important Gathering of CIOs and IT Executives™

At this year's conference, you'll learn:



Discover tools and techniques to enhance your IT and business strategies



Examine the opportunities and risks in adopting emerging and innovative technologies



Challenge how you think about leadership and discover new approaches to lead

Five Ways Artificial Intelligence and Machine Learning Deliver Business Impacts

Ensure that your artificial intelligence innovations are deployed effectively.

[Learn More](#)



Gartner for IT on social media

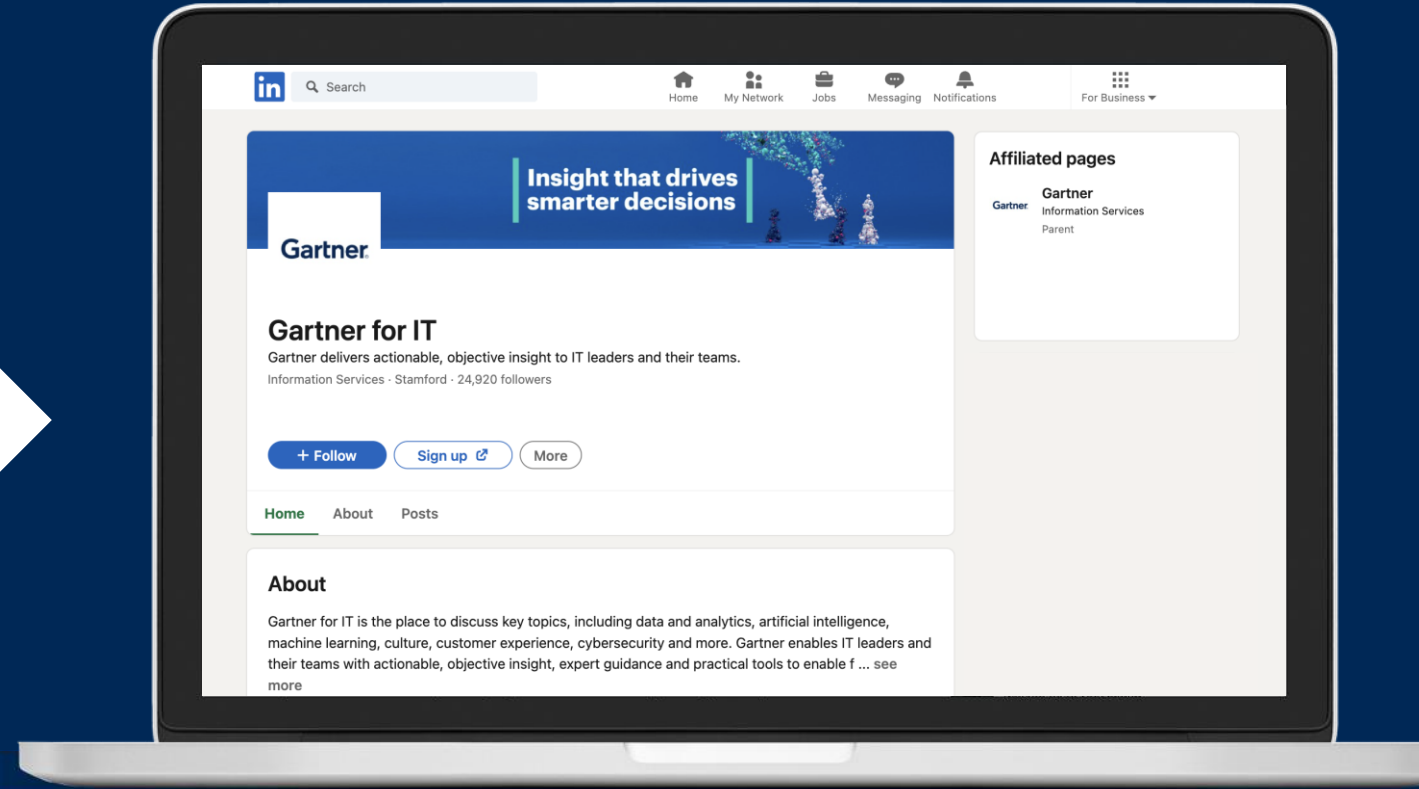
Want to stay in-the-know? Connect with us on LinkedIn and Twitter to receive the latest Gartner IT insights and updates across research, events and more. It's all curated specifically for IT leaders and decision-makers.

Follow us on



Looking for insights delivered to your inbox?

Subscribe to our bi-weekly newsletter



Become a Client

Clients receive 24/7 access to proven management and technology research, expert advice, benchmarks, diagnostics and more.

Fill out the form to connect with a representative and learn more.

[Learn More](#)

Or give us a call: **+441784614280 | +1 855 637 0291**

8 a.m. – 7 p.m. ET

8 a.m. – 5 p.m. GMT

Monday through Friday

Get more Gartner insights



Download the research slides



**View upcoming and on-demand Gartner webinars
at gartner.com/webinars**



Rate this session

Rate Session

RATE THIS SESSION (1-5)

LEAVE FURTHER FEEDBACK

I loved the speaker's insight on this topic!

Cancel

SEND

