Gartner

Gartner Insights

Use This Cheat Sheet to Manage Your LLM Portfolio

Ben Yan, Leinar Ramos, Tong Zhang



Use This Cheat Sheet to Manage Your LLM Portfolio

24 July 2026 - ID G00854598 - 10 min read

By: Ben Yan, Leinar Ramos, Tong Zhang

Initiatives: Technologies and Markets; Building AI and Agentic Foundations

Frontier LLMs now exceed many users' actual needs. The optimal strategy is not pursuing frontier models, but selecting the right models for the right purposes and upgrading at the right time. Leaders responsible for AI should use this cheat sheet to maintain a competitive and efficient LLM portfolio.

Insights at a Glance

As the large language model (LLM) market matures, the initial AI race on LLMs has turned into a structured, three-tier ecosystem:

- **Frontier models (the "trailblazers")**: Most competent models for complex, multistep reasoning, planning and problem solving, such as math or coding.
- **Commodity models (the "workhorses")**: "Good enough" models that balance performance, cost-efficiency and speed, utilized for general-purpose tasks and well-defined enterprise scenarios.
- **Compact models (the "helpers")**: Specialized, low-latency, low-cost models optimized for narrow tasks like **data classification**, PII masking, or supporting UI on devices with limited computing resources.

The traits of open-weight models tend to align with the commodity and compact categories. This alignment enhances their competitiveness and strategic value, transforming these open-weight models from "nice-to-haves" to "must-haves."

Recommendations:

- Leaders responsible for AI should curate a competitive and efficient LLM portfolio leveraging the diversity of LLMs to maximize business value. Don't pay for the intelligence you don't need.

- Beyond selecting the right models, it is also critical to determine the optimal LLM upgrade cadence. Migrating to a newer model as early as possible is not a default "win"; the decision must be a calculation of tangible ROI and transition friction.

Strategic Planning Assumption

- By 2028, 50% of enterprises will have a model portfolio of frontier, commodity and compact models, up from less than 10% in 2026.

Impact

The LLM frontier keeps moving, with newer models outperforming and replacing older ones. [1] For many applications, current models are sufficient, and the GenAI community often prefers fast and efficient LLMs over frontier models. [2]

LLMs can be categorized into three tiers:

- **Frontier models (the "trailblazers")**: Most competent models, reserved for complex, multistep reasoning, planning and problem solving.
- **Commodity models (the "workhorses")**: Serve as the primary enterprise workhorses, offering an optimal mix of performance, speed and cost-efficiency. Capable of handling a diverse set of tasks.
- **Compact models (the "helpers")**: Focus on simple or narrow tasks with extreme high efficiency. While no definitive threshold exists, compact models are generally characterized by a size under 32B, a factor that eases both deployment and fine-tuning.

Recommended models are consistently positioned near the Pareto curve, offering an optimal balance of performance and cost. Models that fall significantly below the curve are not recommended and are likely to become obsolete soon.

There are many models available in the market, but only a few models (on the Pareto curve) really matter. These offer the best trade-off between cost and capability (see Figure 1).

Figure 1: Maximizing Value in LLM Selection: Performance and Cost Trade-Offs



Source: Gartner
854598

Gartner

Based on Gartner clients' interaction data, 25% of organizations use frontier models, 95% use commodity models (most of which are proprietary) and 10% use compact models. Less than 10% of organizations utilize all three types. This evolving LLM landscape presents enterprises with opportunities to optimize their LLM portfolios and enhance the ROI of their AI investments.

Actions

Leaders responsible for AI should:

- **Curate and optimize the LLM portfolio**: Use the selection cheat sheet to maintain a model portfolio and consistently assess the suitability. Optimize your AI budget by reallocating the overspending on frontier models to commodity and compact models.
- **Reevaluate open-weight LLM options:** Review available open-weight LLMs and thoroughly assess their associated risks. Carefully select appropriate use cases and implement effective risk mitigation strategies to unlock significant value from these models.

- **Manage the upgrade cadence:** Regularly evaluate advancements in LLM technology, but manage upgrades at a reasonable cadence. Balance the potential benefits of new models against the costs and risks of transition to ensure sustainable and effective AI operations.

Cautions

- Before deciding to manage LLMs internally and build your AI apps, carefully evaluate whether it is more strategic and cost-effective for your organization to adopt an AI product with embedded, vendor-managed LLMs.
- Considerations such as sustainability, regulatory compliance and data sovereignty vary across enterprises and can significantly limit the range of available LLM options (see the more detailed discussion in How to Select the Right Large Language Model).
- The evolution of LLMs means that agentic workflows designed for earlier, less capable models may face obsolescence. To mitigate this risk, design your agentic workflows and architectures with modularity and flexibility to avoid significant modifications to the overarching solution architecture.

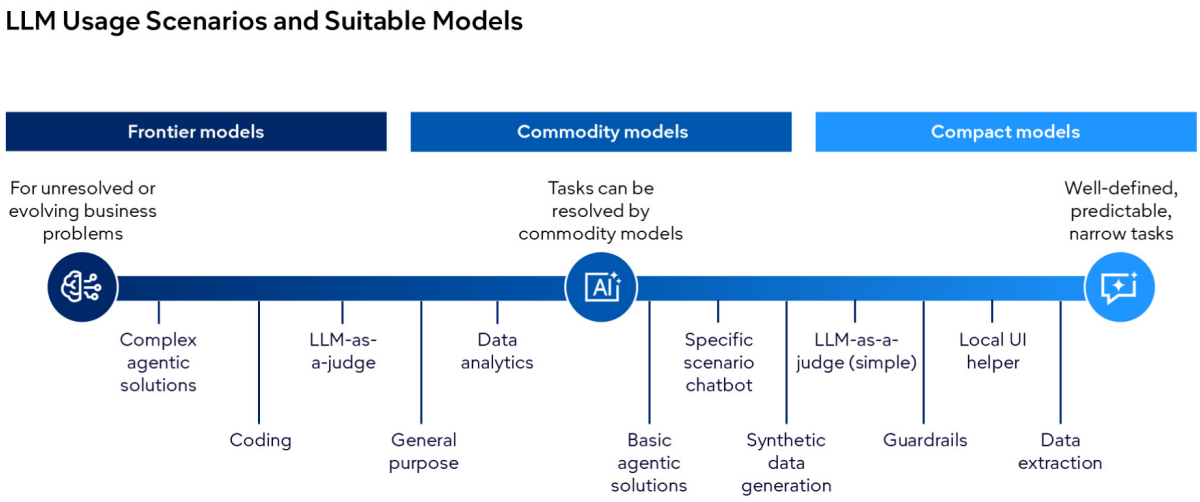
How to Execute

Use This Cheat Sheet to Select the Best-Fit LLM Model For Your Scenario

The attached Powerpoint deck helps you choose the LLM model most appropriate to your requirements. It gives details on the three most common scenarios and their rationales, and representative LLM models for each.

Download the LLM Selection Cheat Sheet
For enterprise LLM adoption, usage typically falls into three primary scenarios. Each corresponds to a distinct category of LLMs (see Figure 2).

Figure 2: LLM Usage Scenarios and Suitable Models



Source: Gartner
854598

Gartner

Scenario 1: Unsolved or Evolving Business Problems

When AI solutions have not fully addressed business challenges, or there is room for improvement, higher-capability LLMs can deliver better results. In these cases, frontier models are worth exploring. However, adopting frontier models does not guarantee immediate ROI; success depends on the organization's specific context.

Representative frontier models include the latest models from Anthropic, OpenAI and Google.

Scenario 2: Tasks Resolved by Commodity Models

Some tasks are already well-handled by commodity models, making frontier models unnecessary. While these tasks are complicated and require good LLM capabilities, efficiency and cost-effectiveness are also important. If performance still falls short of user expectations, the bottlenecks are usually not the models themselves, but other factors such as enterprise data quality or application integration readiness.

Representative models in this category include flash or mini versions of frontier models, highly cost-efficient models, or open-weight options.

Scenario 3: Well-Defined, Narrow Tasks

For tasks that are highly specific, predictable and narrowly scoped, even commodity models are overkill. Compact models, sometimes enhanced with lightweight fine-tuning, can deliver excellent results with maximum cost efficiency and speed.

Representative models here are smaller-sized or distilled versions of frontier and commodity models.

Please note that certain LLMs may exhibit greater verbosity when addressing the same issue, a phenomenon referred to as the “overthinking tax.” This should be also measured, particularly in reasoning tasks, to ensure high cost-efficiency. [3]

Evolving Models and Model Routers

Over time, new frontier models will emerge, old frontier model pricing typically declines and they transition into commodity models, go deprecated or obsolete. Meanwhile certain frontier and commodity models are distilled into more efficient, compact models, such as small language models (SLMs).

Ideally, model routers can be deployed to select proper models in a model pool based on users’ requests. They function as smart traffic controllers that direct each request to the best-suited model, ensuring an optimal balance between performance and efficiency (for more, see Optimize AI Cost and Reliability Using AI Gateways and Model Routers).

Unlock the Value of Open-Weight Models

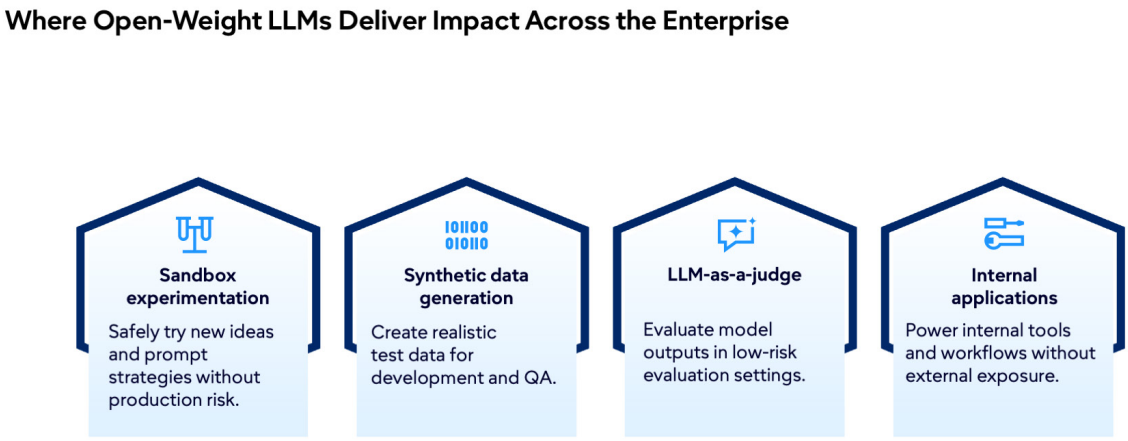
As highlighted in the LLM Selection Cheat Sheet, many open-weight models are featured as representative options. For many enterprises, the value of open-weight models remains underestimated:

- While open-weight LLMs may not be frontier models, they excel as “workhorses” or “helpers” for a variety of tasks, such as DeepSeek V4 (First Take: What’s New From DeepSeek V4), gpt-oss and Qwen models.

- Concerns regarding open-weight security are often overstated:
  - **Data sovereignty**: By hosting models on-premises, or using your preferred cloud service provider (CSP) infrastructure or APIs, you maintain higher control of your data than by using direct model APIs from the model maker. It means using an open-weight LLM does not equate to sending data to another country where the model was trained.
  - Safety: No model is entirely free from bias, as bias inevitably results from the selection of training data or training process, which is often related to regional, cultural or other norms. For use cases where this matters, like in customer service, a specific model may not be suitable for global use. These biases can be mitigated through additional custom guardrails, system-level moderations or model fine-tuning, although requiring additional effort and expertise. This challenge applies to both open-weight and commercial models.

- Open-weight models can deliver value across a range of scenarios, not just in production environments. For example, they are well-suited for the following (see Figure 3):
  - Experimenting new ideas in a sandbox environment
  - **Synthetic data** generation such as creating test cases
  - LLM-as-a-judge (for lower-risk scenarios than those typically addressed by frontier models) to evaluate model outputs during evaluation phase or regression test
  - Internal-use applications

Figure 3: Where Open-Weight LLMs Deliver Impact Across the Enterprise



Source: Gartner
854598

Gartner

These models also provide redundancy if proprietary “workhorse” models are deprecated. Vendors may retire their previous models to conserve computing resources or for other commercial reasons. Leveraging open-weight models helps reduce vendor lock-in and the pressure to prematurely upgrade to new models or overspend. However, it’s important to note that open-weight models can also be deprecated and may lose official support over time.

Given these advantages, leaders responsible for AI should reconsider the adoption of open-weight models as a strategic option in their enterprise LLM portfolio. One important consideration is that self-serving LLMs come with additional operational efforts (see Architects Should Use This Checklist for Evaluating Private LLM Deployments). Using open-weight model APIs through your CSP is a viable way to reduce this operational burden.

Create a Cadence of Upgrading LLMs

It is not necessary to upgrade models at the first opportunity for every task or scenario.

For Unsolved or Evolving Tasks

If the task or scenario remains open or not completely resolved by current frontier models, and higher capabilities could yield significant marginal returns, exploring newer frontier models is worthwhile (e.g., for coding, deep research, or exploratory agentic experiments). Begin with a group of power users at a reasonable cost, and base further scaling decisions on direct user feedback, short-term ROI calculations, or alignment with long-term AI strategy.

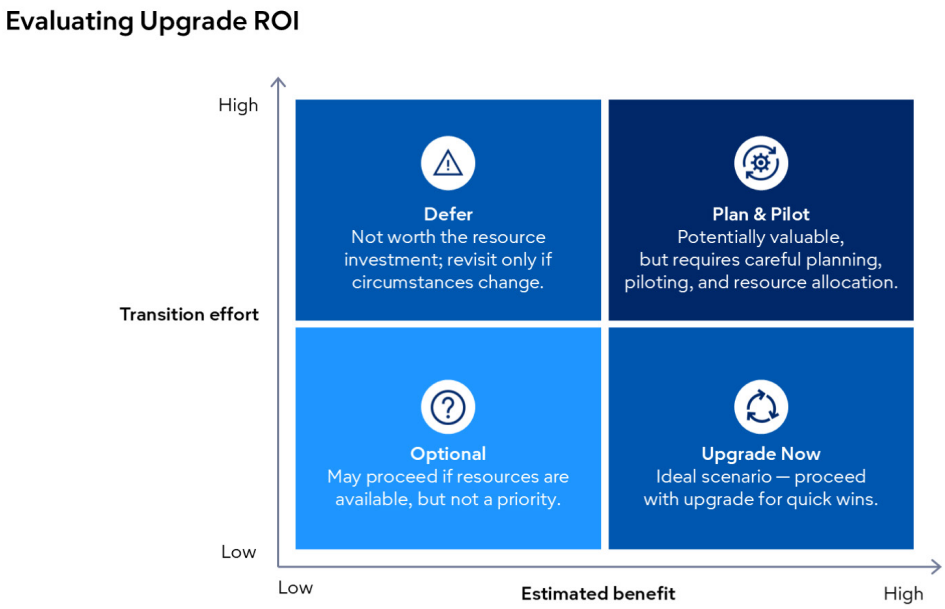
For Cost-Efficiency and Speed Improvements

When more cost-effective and faster models become available, those making replacement decisions should consider:

- The estimated total cost savings or speed benefits over the expected life cycle of the new model (until the next upgrade cycle).
- The transition effort required, including development, model fine-tuning, evaluation, regression testing and deployment (especially for **edge AI**). Note that models may exhibit different biases or behaviors with the same prompts or contexts, and regression testing is particularly critical when switching to a different vendor's model series. The testing efforts increase for specialized or customer-facing applications.

Consequently, the choice of when to perform a model upgrade must be based on a balanced evaluation of the estimated benefit against the potential transition efforts and risks (see Figure 4).

Figure 4: Evaluating Upgrade ROI



Source: Gartner
854598

Gartner

Some model transitions, such as the retirement of a version or the release of a forced update, are dictated entirely by vendors. Consequently, organizations have limited agency in these scenarios and must adapt as their providers' mandates change.

To ensure a smooth and informed transition, consider using a controlled sandbox environment to observe new model behavior in shadow mode. Implementing a probation period where both the existing and new models run in parallel allows for a thorough evaluation of performance and the collection of direct user feedback before a full rollout.

Success Measures

- Cost savings achieved by reallocating budget from high-cost frontier models to suitable commodity models, or specialized compact models for narrow tasks. If organizations choose to self-serve the models, then operational and management expenses need to be included in the total cost calculation.
- Speed improvements from using faster models, such as higher throughput on high-volume tasks or better **customer experience**.
- Reduction in vendor lock-in by successfully integrating open-weight models as redundant "workhorses" in the enterprise model pool.

Contributors

Mike Fang, Arnold Gao, Wilco van Ginkel, Pieter den Hamer, Julian Sun, Birgi Tamersoy

Evidence

[1] LLM Leaderboard, Arena

[2] LLM Rankings, OpenRouter

[3] OckBench: Measuring the Efficiency of LLM Reasoning, Arxiv

---

Recommended by the Authors

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

How to Select the Right Large Language Model

Boost AI ROI With a Portfolio Approach in Response to Chinese Models

AI Tokenomics: When to Cut AI Tokens — And When to Double Down

---

© 2026 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 Business and Technology Insights 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 insights may address legal and financial issues, Gartner does not provide legal or investment advice and its insights 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 insights is produced independently by its Business and Technology Insights Organization without input or influence from any third party. For further information, see " Guiding Principles on Independence and Objectivity." Gartner insights may not be used as input into or for the training or development of generative artificial intelligence, machine learning, algorithms, software, or related technologies.

Actionable, objective insights

Position your organization for success. Explore these additional complimentary resources and tools.



Report

The CHRO Report

Quarterly insights for HR leaders on challenges, risks and actions.

Download Now



Webinar

The Top CHRO Trends and Priorities Shaping 2027

Expert insights and recommendations to shape HR's strategy for 2027.

Watch Now



Roadmap

AI Roadmap for HR

Discover key steps to build an AI roadmap that delivers value in HR.

Download Now



Guide

CHRO Guide for HR Transformation

View practical strategies to reinvent HR and deliver value in the AI era.

Download Now

Already a client? Get access to even more resources in your client portal.
Log In ↗

Connect with us

Get actionable, objective business and technology insights that drive smarter decisions and stronger performance on your mission-critical priorities. Contact us to become a client:

U.S.: 1 855 322 5484
International: +44 (0) 3300 296 946

Become a Client

Learn more about Gartner for CHROs
gartner.com/en/human-resources

Stay connected to the latest insights

Attend a Gartner conference
View Conference

© 2026 Gartner, Inc. and/or its affiliates. All rights reserved. CM_GBS_5497850

Gartner®