Enterprise GenAI as Generalized Search Intelligence

Why Performance Is Jagged and Where Humans Remain Essential

Yakov Pyotr Shkolnikov · Soundness

Generative AI (GenAI) has made substantial progress in performance and usability. Yet its adoption in enterprises has been uneven, and many applications have failed to scale from prototypes to production. To help explain this uneven progress, we suggest that current enterprise GenAI systems can be interpreted as a generalized search intelligence (GSI). Such a system combines knowledge learned from large volumes of text, training that improves the selection of useful outputs, and information retrieved from databases, web searches, and connected tools to enrich the model's context. This view yields practical heuristics for how organizations should integrate GenAI into their processes.

Knowledge engineering remains a critical bottleneck to success, even as AI accelerates the discovery of individual processes.

Full research paper: SSRN preprint forthcoming
yshkolni@gmail.com

Introduction

Large language models (LLMs) have transitioned from laboratory experiments into a large-scale industry. Their uneven performance in enterprise applications raises a practical question: what makes these systems useful when they work, and what limits their deployment?

HOW LLMS GENERATE OUTPUT

LLMs generate output through autoregressive token prediction. At each step, the LLM uses the preceding tokens as context to predict the next token, which is then added to the context. This process repeats until the output sequence is complete (Bengio et al., 2003; Vaswani et al., 2017). The generated tokens may form part of a direct answer to the user or specify a tool call executed by the surrounding harness (Schick et al., 2023).

Two complementary mechanisms have been used to increase the chances that model outputs are useful. The first has been improving the model's ability to predict the next token from the preceding context, including when and how to call tools. The second has been adding tools that enrich the model's context and enable interaction with external systems (Figure 1).

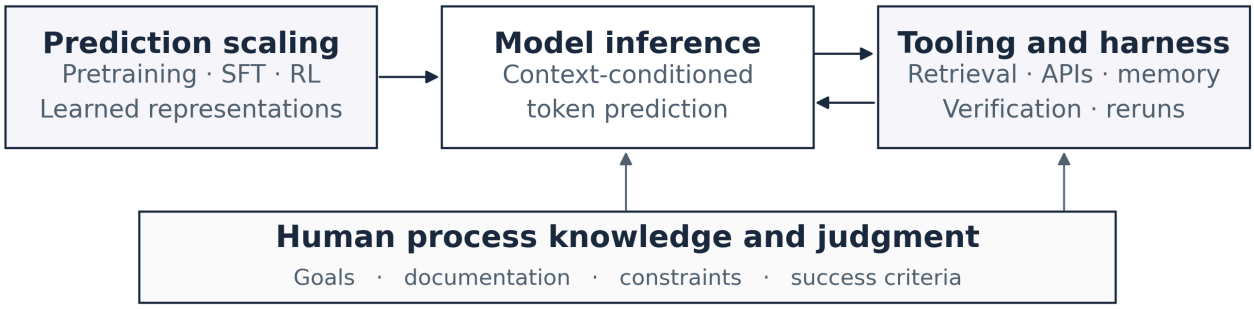


Figure 1. Prediction scaling, model inference, and the surrounding harness jointly support enterprise GenAI under the generalized search intelligence (GSI) interpretation. Human process knowledge and judgment supply goals, constraints, and criteria for success.

Interpreting enterprise GenAI as generalized search

In enterprise GenAI, combining the scaling of prediction and tooling has produced a system that can be interpreted as a generalized search intelligence (GSI). The user's request guides the model's retrieval and tool use, which augment the context from which the model generates output token sequences. The harness serves as a knowledge-engineering mechanism that enables this retrieval, as well as the verification and ranking of information. The information available for search includes documentation, data definitions, tool schemas, and decision rules that were missing or underrepresented during model training. Additional approaches such as Tree of Thoughts explicitly search multiple candidate reasoning trajectories before selecting a result (Yao et al., 2023).

SCALING PREDICTION

Prediction capability has improved through the invention of several techniques. In the original decoder models, the weights were optimized primarily to predict the next token in the training corpus. Subsequent work retained next-token prediction while changing which sequences the model was trained or rewarded to produce.

First, supervised fine-tuning increased the probability that a prompt would be followed by a complete answer. Reinforcement learning further optimized the model weights toward results that humans preferred (Ouyang et al., 2022). Chain-of-thought-based training increased the quality of the user-facing answer by having the model first generate intermediate reasoning tokens (Guo et al., 2025). Finally, models were trained or rewarded to generate tool-call tokens that request external operations whose returned information can help complete the task (Schick et al., 2023).

SCALING TOOLING

Tooling underwent a parallel development. First, retrieval-augmented generation (RAG) augmented the model context with retrieved external information, including information more current than the model's training data (Lewis et al., 2020). Tool calling then allowed models to interact with computer systems through application programming interfaces (APIs) and retrieve data from databases and other services.

Harnesses soon added the ability to store information in flat files and databases and return previous outputs to the model as context. They also allowed models to write scripts and other code that the surrounding system could execute to invoke software libraries. They now coordinate planning, decide when to call the LLM, and reprompt it or rerun generation when outputs fail verification. Standards such as the Model Context Protocol (MCP) let models call new tools without retraining (Model Context Protocol, 2025).

A historical analogy for generalized search is IBM's Deep Blue, which achieved strong results in chess, a domain with explicit rules and measurable outcomes, by utilizing search (Campbell et al., 2002). Its success on this well-defined problem did not show that the same approach could extend to activities valuable to enterprises, such as independently discovering processes and criteria for success in an organization. The same question now applies to enterprise GenAI. This article uses the GSI framework to compare how current systems handle well-defined problems with how they handle open-ended enterprise problems: can the combined system reliably discover undocumented processes and exercise sound judgment, or will it instead keep the burden of knowledge engineering on humans?

Why enterprise GenAI performance is jagged

Interpreting current enterprise GenAI systems as GSI may help explain the "jagged technological frontier": AI improves performance on some tasks but degrades it on others of seemingly similar difficulty (Dell'Acqua et al., 2026). Consider what happens when retrieval and tool access are removed. Without retraining to incorporate new knowledge, performance declines as the model's representation becomes outdated. Even tasks commonly considered to require reasoning alone, such as coding, suffer without access to current documentation (Zhou et al., 2023). The experiment is imperfect: because many LLMs are now trained to use tools, removing access to those tools can itself reduce performance. It is nevertheless informative. The inability to use APIs, access current information, or inspect data in organizational databases is not simply a failure of model reasoning. It reflects a lack of access to the information the enterprise system may need to produce a useful answer.

For many businesses, this experiment is already underway. When systems are disconnected and processes are undocumented, people may not know what to search for or how to verify the result, and even with a human in the loop, the system may produce results that do not meet the organization's requirements. Moreover, much of the information models encounter during training and inference consists of finalized artifacts produced by human and organizational decisions. These artifacts often omit the process by which the content was generated: the rejected alternatives, the influence of organizational incentives, and the negotiations that produced a common solution. Learning to retrieve and reproduce patterns of past decisions does not necessarily correspond to an ability to make those decisions.

Learning to retrieve and reproduce patterns of past decisions does not necessarily correspond to an ability to make those decisions.

These gaps in access and process knowledge may explain much of the falloff between prototypes and production. Organizations attempting to apply AI where systems are disconnected, processes are poorly defined, people cannot judge the quality of outputs, or tasks require a jump beyond the available representation are expecting autonomous process discovery and judgment that current enterprise deployments have not reliably demonstrated. The result can be high average performance on routine, constrained decisions alongside costly failures on undocumented exceptions and under shifting conditions, especially when deploying AI itself can cause such a shift.

Enterprise deployment evidence

THE DEPLOYMENT RECORD

50%

at least 50% of GenAI projects had been abandoned after proof of concept by the end of 2025

Gartner (2026a)

37%

only 37% of broader AI initiatives launched in the previous 12 months were live and delivering value

S&P Global (2026)

70%

of enterprises will abandon agentic AI systems built through vendor forward-deployed engineering by 2028

Gartner (2026b), forecast

The current enterprise deployment record provides observational evidence consistent with a GSI interpretation. Large numbers of AI prototypes fail to make it to production. Gartner (2026a) reported that at least 50% of GenAI projects had been abandoned after proof of concept by the end of 2025, while S&P Global (2026) found that only 37% of broader AI initiatives launched in the previous 12 months were live and delivering value. Based on interviews with AI practitioners, RAND identifies misunderstood problem definitions, including mismatched metrics and poor workflow fit, and missing data as leading root causes of AI/ML project failure (Ryseff et al., 2024). Under the GSI interpretation, these can reflect failures of knowledge discovery and problem formulation, as well as the costs of creating and maintaining an appro-

priate knowledge harness. While these measures use different populations and definitions, the outcomes are consistent with an architecture in which the harness provides essential process information and verification. They do not, by themselves, establish the cause of any particular project failure.

The responses of OpenAI, Anthropic, and Microsoft are also informative, with growing investments in forward-deployed engineers (FDEs) and consultancies to help with workflow redesign, data integration, and organizational adoption (OpenAI, 2026; Althoff, 2026). Anthropic (2026) also announced a $100 million commitment to train 10,000 “Frontier Deployed Engineers” from its customers and partners by the end of 2027. Given the code-writing capabilities already demonstrated by these models, the continued need for specialized deployment labor is consistent with a remaining bottleneck in discovering process requirements and performing the knowledge engineering required for reliable integration.

Shallow and deep search

Applied to enterprise tasks, the GSI interpretation also raises a conundrum.

When a task is well documented, with a clear representation of its processes, easily located data, and easily verified results, when are extensive reasoning capabilities economically justified? In software, if a task is repeated frequently and is well defined, writing a library once and reusing it can provide cheaper and more reliable code, and once documented, the library can be found and called by simple autocomplete tools. For business reporting, creating a reusable dashboard may be more economical than a natural-language interface for exploring it.

Both examples are shallow searches that often do not require frontier GenAI (Figure 2). The required information is known, easily available, and current; the output is easily checked; and the task ends with a reusable artifact whose value exceeds the value of the search itself.

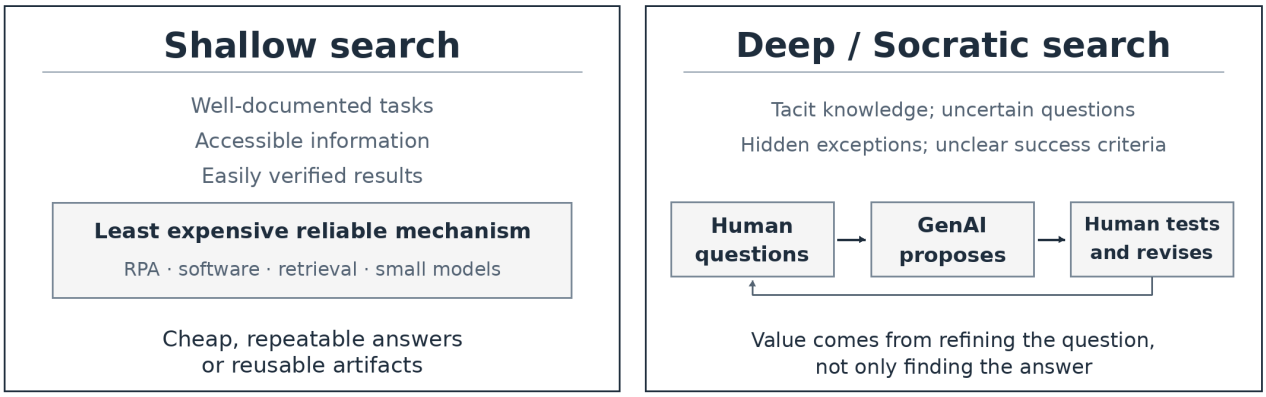


Figure 2. Shallow searches favor inexpensive, reliable mechanisms for well-defined tasks. Deep searches use an iterative Socratic loop in which people refine questions and verify results.

But for applications without these required components, the burden of making the system work, defining its goals, and verifying its outputs often falls on people. This is something that many people using AI have observed. A user asks a chatbot for an answer. The result misinterprets the request. The user reasons about

the failure and changes the question to be more precise, but the model returns another wrong answer. The user figures out what the model misunderstood and corrects it while fixing the data. This continues until the model finally provides an acceptable answer.

This example demonstrates a deep search in which the initial question may be wrong, important knowledge is tacit, success criteria are undefined, edge cases dominate, errors can be hidden by fluency, and value comes from the search for an answer rather than the answer found. Through this questioning, the user supplies information about the underlying processes that was omitted from the presentation of work used in model training, while also helping define the knowledge and tools the harness must provide.

In these deep searches, human expertise remains an important part of the current system, and users gain value from refining the questions themselves. The GenAI system can thus function as a Socratic engine that challenges users to examine their assumptions, much like an AI sparring partner (Herrmann, 2025) or a source of productive friction (Su et al., 2026).

Heuristics for using GenAI in enterprises

The accidental Socratic function therefore provides a simple set of heuristics for deciding where to apply AI within a company. For shallow search tasks, robotic process automation, conventional software, and inexpensive models are often sufficient.

For deep search tasks, frontier GenAI is particularly useful in the following areas:

1. Apply it in research and development and early prototyping, where Socratic search can provide uneven but meaningful gains through exploration and innovation.

2. Apply it when mapping and standardizing company processes, where the Socratic method can accelerate discovery. Here, however, AI provides only part of the function. AI offers persistence, breadth, and a willingness to revisit contradictions indefinitely. It does not reliably supply missing organizational experience, negotiate organizational politics, or assume responsibility for difficult recommendations.

3. Apply it in employee training, where an engine can be intentionally configured as a Socratic interrogator that asks employees questions so they learn actively and through introspection rather than reading or memorization, as found with university programming students (Sun et al., 2026).

4. Apply it in large and complex software development tasks, where a combination of many requirements with existing code and systems makes it difficult for developers to locate relevant information and satisfy every constraint simultaneously, particularly when coding in an unfamiliar language or codebase.

Additionally, consider how frequently decisions fall outside established patterns represented in the model or made accessible through the harness, whether exceptions can be easily detected, and the severity of the consequences of errors. Processes that are stable and have reversible errors are better candidates for au-

tomation. Processes with costly errors, especially those with delayed consequences, are more likely to require frequent human intervention and burdensome review.

Organizations must also consider how frequently processes change due to markets and regulations, requiring updates to the harness. This includes changes due to the introduction of AI itself. These changes create ongoing harness maintenance costs.

Finally, organizations must retain sufficient staff with process expertise to manage changes, evaluate exceptions, and recognize when established decision rules no longer apply. The economic value of GenAI-enabled automation should therefore account for the additional costs of maintaining the harness, verifying outcomes, and retaining human expertise.

Enterprise value and cost of the Socratic engine

The organizational Socratic method is valuable but prolonged, expensive, and uncomfortable. The valuable questions are likely to overlap with the uncomfortable ones: whether the processes followed match the documentation, whether KPIs are measuring desired outcomes, who has authority to make decisions, and whether the evidence supports the strategy. An AI system used this way does not merely retrieve information. It forces the organization to account for itself. Its durable output may nevertheless be a company that understands itself better.

An AI system used this way does not merely retrieve information. It forces the organization to account for itself.

But the cost of using AI through a Socratic method can be extremely high inside an organization. Organizations often rely on outside consultants to solve this problem because consulting firms perform the same function while bearing the brunt of the discomfort. The cost is not simply model inference, but also the labor required for discovery, documentation, implementation, architecture, security, and verification. For processes that change frequently or are executed infrequently, these costs may be difficult to amortize, particularly given ongoing maintenance requirements. The result can be negative net value compared with continuing to rely on existing processes supported by human expertise and judgment. Consistent with this risk, Gartner (2026b) forecasts that 70% of enterprises will abandon agentic AI systems built through vendor forward-deployed engineering by 2028, due to escalating costs and the inability to evolve these systems independently.

Conclusion

Viewing enterprise GenAI through the GSI framework does not diminish its practical achievements or imply that transformer inference is a conventional search algorithm. Rather, it clarifies the conditions under which current enterprise architectures are useful and provides practical heuristics.

Shallow searches are the best candidates for the least expensive and most reliable implementation, such as with deterministic software, standard retrieval, and potentially with the assistance of small models. Deep searches, on the other hand, can benefit from the use of frontier GenAI models with humans in the loop to refine the questions, expose contradictions, and map out what the organization does not yet understand. In these cases, the system functions as a Socratic engine, and humans remain part of the intelligence, supplying missing context and judgments of success.

These patterns are not merely theoretical predictions. The observed outcomes of enterprise AI deployments have not demonstrated that even access to all company data is sufficient to independently reconstruct undocumented processes and automate them reliably. Knowledge engineering remains a critical bottleneck to success, even as AI accelerates the discovery of individual processes. Until reliable autonomous models and algorithms for process discovery and judgment are demonstrated, organizations should deploy enterprise GenAI where the decision space and evaluation criteria are already explicit, or budget for and retain the human expertise required to discover, define, and maintain the processes on which GenAI depends.

Prioritizing well-defined and verifiable use cases, choosing the least expensive reliable mechanisms, and retaining human expertise for ambiguous or changing processes can improve the likelihood of positive ROI as enterprise GenAI and its supporting systems continue to evolve.

AI Use Disclosure

Generative AI tools were used to create the figures and assist with minor editorial revisions to the text. The author developed the conceptual framework, analysis, arguments, and conclusions and reviewed the final manuscript.

Funding and Independence

This research was conducted independently and was self-funded. This brief presents the author's independent research on generalized search intelligence.

References

Althoff, J. (2026, July 2). "Microsoft Frontier Company: AI Engineering That Amplifies and Protects Your Intelligence." The Official Microsoft Blog. https://blogs.microsoft.com/blog/2026/07/02/microsoft-frontier-company-ai-engineering-that-amplifies-and-protects-your-intelligence/

Anthropic. (2026, October 2). "Claude Frontier Academy: $100M to Train 10,000 Engineers." https://www.anthropic.com/news/claude-frontier-academy

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003). "A Neural Probabilistic Language Model." Journal of Machine Learning Research 3:1137–1155. https://www.jmlr.org/papers/v3/bengio03a.html

Campbell, M., Hoane, A. J., Jr., and Hsu, F. (2002). "Deep Blue." Artificial Intelligence 134(1–2):57–83. https://doi.org/10.1016/S0004-3702(01)00129-1

Dell'Acqua, F., et al. (2026). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of Artificial Intelligence on Knowledge Worker Productivity and Quality." Organization Science 37(2):403–423. https://doi.org/10.1287/orsc.2025.21838

Gartner. (2026a, January 26). "Why Half of GenAI Projects Fail: Avoid These 5 Common Mistakes." https://www.gartner.com/en/articles/genai-project-failure

Gartner. (2026b, September 29). "Gartner Predicts 70% of Enterprises Will Abandon Agentic AI Built by Vendor Forward-Deployed Engineering by 2028." https://www.gartner.com/en/newsroom/press-releases/2026-09-29-gartner-predicts-70-percent-of-enterprises-will-abandon-agentic-ai-built-by-vendor-forward-deployed-engineering-by-2028

Guo, D., et al. (2025). "DeepSeek-R1 Incentivizes Reasoning in LLMs Through Reinforcement Learning." Nature 645:633–638. https://doi.org/10.1038/s41586-025-09422-z

Herrmann, T. (2025). "AI as a Sparring Partner – An HCAI Approach to Promote Human Capabilities." In H. Degen and S. Ntoa (Eds.), Artificial Intelligence in HCI, Lecture Notes in Computer Science 15820:162–176. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-93415-5_10

Lewis, P., et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks." NeurIPS 2020. https://arxiv.org/abs/2005.11401

Model Context Protocol. (2025). "Specification (June 18, 2025)." https://modelcontextprotocol.io/specification/2025-06-18

OpenAI. (2026, February 5). "Introducing OpenAI Frontier." https://openai.com/index/introducing-openai-frontier/

Ouyang, L., et al. (2022). "Training Language Models to Follow Instructions with Human Feedback." NeurIPS 2022. https://arxiv.org/abs/2203.02155

Ryseff, J., De Bruhl, B. F., and Newberry, S. J. (2024). "The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed: Avoiding the Anti-Patterns of AI." RAND. https://www.rand.org/pubs/research_reports/RRA2680-1.html

S&P Global. (2026). "The AI and Labor Landscape 2026: Increased Investment, Persistent Productivity Gains and a Recalibrated Employment Outlook." https://www.spglobal.com/en/research-insights/special-reports/ai-impact-on-employment-2026

Schick, T., et al. (2023). "Toolformer: Language Models Can Teach Themselves to Use Tools." NeurIPS 2023. https://arxiv.org/abs/2302.04761

Su, X., et al. (2026). "Who Thinks First? Designing Productive Friction with Engage-to-Unlock GenAI." https://arxiv.org/abs/2610.01518

Sun, D., Zheng, Y., Xu, J., and Yang, Z. (2026). "When Generative AI Meets Socratic Method: Investigating Programming Learning Dynamics Through Behaviours, Interaction Qualities and Perceptions." Journal of Computer Assisted Learning 42(2):e70210. https://doi.org/10.1002/jcal.70210

Vaswani, A., et al. (2017). "Attention Is All You Need." NeurIPS 2017. https://arxiv.org/abs/1706.03762

Yao, S., et al. (2023). "Tree of Thoughts: Deliberate Problem Solving with Large Language Models." NeurIPS 2023. https://arxiv.org/abs/2305.10601

Zhou, S., Alon, U., Xu, F. F., Wang, Z., Jiang, Z., and Neubig, G. (2023). "DocPrompting: Generating Code by Retrieving the Docs." ICLR 2023. https://arxiv.org/abs/2207.05987