Model Distillation and the Diffusion of AI Capability

How Smaller Models are Reshaping Inference Economics

Author: Phil Marshall, PhD

Publish Date: June 2026

Executive Summary

Model distillation allows selected capabilities from large frontier models to be transferred into smaller, cheaper deployment models. When applied to narrow and well-defined tasks, distilled models can materially reduce inference cost, latency, and infrastructure complexity. In some cases, modeled serving costs can fall by an order of magnitude when workloads are handled entirely by smaller models.

However, distillation is not a general replacement for frontier models. Smaller distilled models can perform well on selected benchmarks and routine workloads, but they often degrade on open-ended, unfamiliar, complex, or high-risk tasks. This is especially true when the distilled model is very small, or when the workload requires broad reasoning, domain expertise, coding capability, or robust handling of edge cases.

The strategic opportunity is therefore not simply to replace large models with small ones. It is to build tiered inference architectures, where routine work is handled by smaller models and frontier models are reserved for complex, ambiguous, novel, or high-value cases. In this architecture, model routing, evaluation, monitoring, and escalation become as important as model selection itself.

The result is a more layered AI deployment architecture. Frontier models remain critical for creating new capabilities, generating high-quality training data, handling difficult cases, and validating outputs where risk is high. Distilled models create value when they can serve defined workloads more cheaply and reliably. The advantage shifts to organizations that can match each workload to the right model tier while maintaining clear standards for quality, risk, and performance.

© Copyright Tolaga Research, 2026

Why Model Distillation Matters

Model distillation and inference economics

Model distillation improves inference efficiency by training a smaller student model to approximate selected behaviors of a larger and more capable teacher model. The objective is not to compress the full intelligence of the teacher into a smaller system. It is to transfer useful capabilities for a defined set of tasks.

This distinction is important. Distillation works best when the target workload is narrow, repeatable, and measurable. Examples include customer support classification, structured summarization, document extraction, coding assistance for bounded tasks, internal knowledge workflows, and domain-specific question answering. These are settings where the desired output can be specified, evaluated, and monitored.

The economic logic is straightforward. Smaller models require less memory, less compute, and simpler infrastructure. They can often run on fewer GPUs, or in some cases on a single accelerator. This reduces serving cost, improves latency, and makes deployment easier to scale across production workloads.

The limitation is that the student model does not inherit the full breadth or flexibility of the teacher. It may perform well on familiar and bounded tasks while failing on unfamiliar questions, subtle reasoning problems, or complex edge cases. Distillation therefore improves economics only when it is matched to the right workload.

The learning signal

Classical distillation trains the student model using the teacher model's probability distribution across possible outputs. These soft targets provide a richer training signal than simple correct-or-incorrect labels. The student learns not only the answer selected by the teacher, but also how the teacher ranked alternative outputs.

Modern large language model workflows have adapted this approach. Rather than relying only on direct access to the teacher model's internal probabilities, developers often use sequence-level distillation. In this approach, the teacher model generates large volumes of high-quality synthetic training data, such as explanations, worked examples, reasoning traces, summaries, or domain-specific responses. The student model is then fine-tuned on those outputs.

This has practical advantages. Sequence-level distillation does not necessarily require access to the teacher model's architecture, weights, or logits. In some cases, developers can approximate aspects of a teacher model's behavior using API-generated outputs. This makes distillation more accessible, although it also introduces risks if the generated data is narrow, noisy, biased, or insufficiently representative of the production workload.

© Copyright Tolaga Research, 2026

The quality of the training signal is therefore central. A smaller model can only learn what the distillation process exposes it to. If the training examples over-represent clean, standard, or benchmark-like tasks, the student model may appear strong in testing but degrade when deployed against messy real-world inputs.

The infrastructure impact

The infrastructure impact of distillation can be significant. A 70B parameter model often requires multi-GPU serving, tensor parallelism, and careful memory management. A distilled 7B parameter model may be able to run on a single GPU without tensor-level partitioning across multiple accelerators.

This changes the deployment profile. Smaller models can reduce memory requirements, lower per-query cost, simplify serving architecture, improve responsiveness, and make it easier to host models closer to the application. They can also support higher throughput for routine workloads where frontier-level capability is unnecessary.

The trade-off is reduced capacity. Smaller models may perform well when the task is familiar and bounded, but they typically underperform on open-ended, unfamiliar, or highly complex tasks. This makes workload selection critical.

The healthcare analogy

Distillation is structurally similar to using specialized clinics rather than sending every patient to a major hospital. Routine and well-understood cases can be handled by smaller, more focused services. Complex, unusual, or high-risk cases are escalated to a major hospital with broader expertise and deeper resources.

The analogy is useful because it highlights the economic purpose of tiering. Not every task requires the most capable system. A well-designed system reserves scarce and expensive capacity for the cases where it is most valuable.

The analogy also exposes a limitation. Unlike human specialists, small AI models cannot reliably judge when a task exceeds their competence. A distilled model may produce a confident answer even when it lacks the reasoning depth or domain knowledge required. Escalation therefore, cannot depend on the small model's self-awareness alone. It must be supported by programmatic routing logic, confidence checks, validation rules, workload classification, and monitoring.

There is also an economic trade-off. If every output from the smaller model must be reviewed by a frontier model, some of the infrastructure savings are lost. The goal is to design routing and validation systems that preserve quality without turning every task into a two-model inference process.

© Copyright Tolaga Research, 2026

Engineering realities

Distillation is a rigorous engineering discipline, not simple compression. Matching a teacher model's behavior in production requires careful data selection, evaluation design, quality testing, and iteration.

One practical challenge is superficial mimicry. Student models can learn the teacher model's tone, structure, formatting, and confident style without preserving the same depth of reasoning or factual reliability. These surface features may satisfy automated evaluations while masking brittle logic, incomplete generalization, or factual errors.

This creates deployment risk. A distilled model may look strong in a controlled test environment but fail under edge cases, ambiguous instructions, adversarial prompts, or domain-specific exceptions. Standard benchmarks can help, but they are not enough. Enterprises need workload-specific evaluations that reflect the actual distribution of production tasks.

Good distillation therefore requires representative data, error analysis, human review, benchmark design, red-teaming, monitoring, and fallback mechanisms. The deployment system must be able to detect where the smaller model is likely to degrade and route those cases to a stronger model or human reviewer.

Distillation is not MoE

Distillation and MoE solve different problems

Distillation should not be confused with Mixture-of-Experts. The two approaches address different problems.

Mixture-of-Experts is primarily a model-scaling technique. It increases total model capacity by adding expert sub-networks, while activating only a subset of those experts for each token. The goal is to increase capability and parameter scale without activating the full model for every token.

Distillation operates at a different level. It trains a smaller standalone model to approximate selected behaviors of a larger model. The teacher model may itself be dense, MoE-based, or part of a broader model workflow. The student model is then deployed as a cheaper and simpler system for defined workloads.

The key difference is that MoE is about scaling large models more efficiently, while distillation is about transferring selected capabilities into smaller models that are easier to deploy.

© Copyright Tolaga Research, 2026

Complementary, rather than substitutes

Distillation and MoE are complementary rather than direct substitutes. MoE helps frontier and near-frontier models scale capacity. Distillation helps downstream systems deploy selected capabilities more economically.

A large MoE model can act as a teacher. It can generate synthetic training data, reasoning examples, domain-specific outputs, or evaluation cases. Those outputs can then be used to train smaller distilled models for execution in lower-cost environments.

This means the model ecosystem is likely to contain both approaches. Frontier and near-frontier models, including dense and MoE systems, will continue to push the capability boundary. Distilled models will serve more targeted workloads where cost, latency, control, and deployment simplicity matter.

The practical implication is that enterprises should not ask whether distillation replaces MoE. They should ask where each approach fits in the inference stack.

The Diffusion of Frontier Capability

Separating capability creation from deployment

Distillation changes the economics of AI by partially separating the creation of frontier capability from its deployment.

Frontier models remain expensive to train and important for pushing the boundary of what models can do. They are also valuable for difficult reasoning, unfamiliar tasks, synthetic data generation, tool-use planning, evaluation, and quality assurance. However, once new capabilities are demonstrated, some of them can be transferred into smaller systems that are cheaper and faster to operate.

This creates a diffusion effect. Capabilities that first appear in frontier systems can migrate into smaller models through distillation, fine-tuning, quantization, synthetic data generation, and application-specific training. These smaller models do not fully replace frontier systems, but they can reduce the need to invoke them for every task.

The result is a more layered AI market. Frontier models create and refresh capability. Smaller models execute selected workloads more efficiently. The value shifts from using the largest model everywhere to designing an inference architecture that matches model capability to task requirements.

© Copyright Tolaga Research, 2026

The emerging model ecosystem

The model ecosystem is increasingly structured around a capability ladder. Frontier and near-frontier models push the capability boundary and may serve as teachers.

Open and accessible model families provide base models that can be fine-tuned, distilled, compressed, and deployed in lower-cost settings.

DeepSeek provides one of the clearest public examples. Its R1 release included several dense distilled models, ranging from 1.5B to 70B parameters, based on Qwen and Llama model families. More broadly, Qwen, Gemma, Llama, and Mistral are becoming important base-model ecosystems that can be adapted for narrower workloads.

Cloud platforms are also beginning to offer distillation and model tuning as enterprise capabilities. Google, for example, offers Gemini tuning through Vertex AI and has introduced a Gemini Distillation Service for training smaller student models from larger teacher models. This points toward a market in which organizations do not rely only on off-the-shelf frontier models. Instead, they combine frontier APIs, open models, synthetic training data, fine-tuning, distillation, evaluation, and routing into customized inference systems.

Enterprise deployment implications

For enterprise deployment, the relevant student-model range will often be in the 7B to 32B parameter class, depending on workload complexity, latency requirements, cost targets, and quality thresholds.

Models in this range can be easier to host, monitor, tune, and route than frontier-scale systems. They may retain enough capability for defined tasks such as customer support, document summarization, legal review, coding assistance, research synthesis, and internal knowledge workflows. However, the right model size depends on the workload. A 7B model may be sufficient for classification or structured extraction, while a 32B or 70B distilled model may be required for harder reasoning or coding tasks.

This reinforces the importance of workload-specific evaluation. Enterprises should avoid assuming that smaller models are good enough simply because they perform well on selected benchmarks. They need to measure performance against their own data, users, edge cases, and quality thresholds.

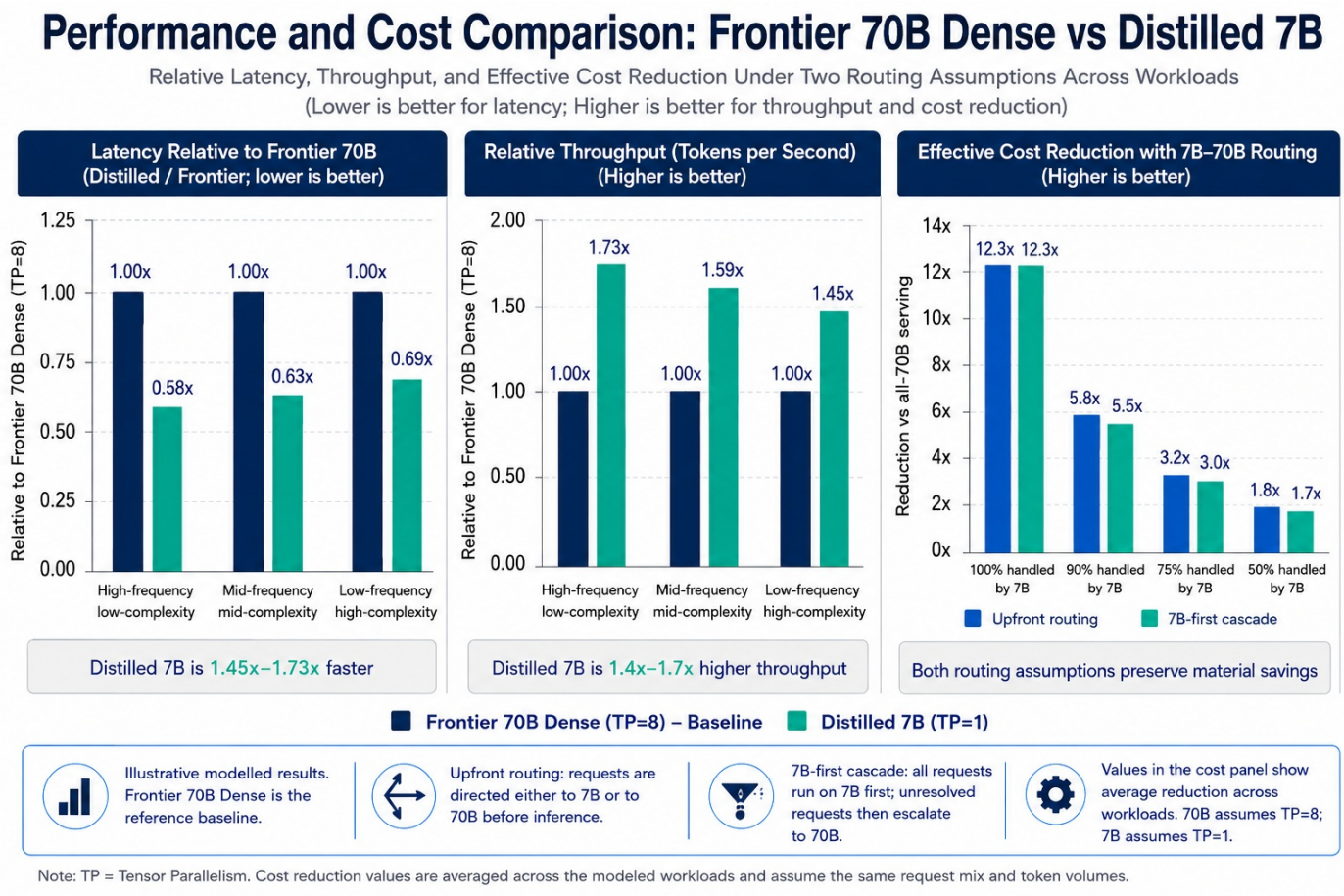
© Copyright Tolaga Research, 2026

Illustrative Economics of Tiered Inference

The infrastructure benefits described above become more meaningful when translated into serving economics. Using Tolaga's AI Infrastructure Simulator, the following modeled example compares an all-70B architecture with alternatives that route eligible requests to a distilled 7B model. The results are illustrative, but they show that the value of distillation depends on routing accuracy, escalation rates, and the share of workload that can be handled safely by the smaller model.

Modeled serving architecture

The baseline scenario uses a 70B dense model with tensor parallelism across eight GPUs. The distilled 7B scenario assumes tensor parallelism of one, allowing the model to run without tensor-level partitioning across multiple accelerators. This difference in deployment footprint is a major source of the modeled latency, throughput, and cost advantages.



Across the workloads shown, the distilled 7B model delivers lower latency and higher throughput than the 70B baseline. Latency falls to approximately 0.58x to 0.69x of the 70B baseline, making the distilled model roughly 1.45x to 1.73x faster. Throughput improves by a similar margin.

© Copyright Tolaga Research, 2026

The cost impact is more significant. If all requests can be handled by the distilled 7B model, the modeled cost reduction is approximately 12.3x relative to all-70B serving. This reflects the lower compute and memory requirements of the smaller model, as well as the simpler serving architecture.

However, the full savings are only available when the smaller model can safely handle the workload. If a meaningful share of requests still needs to be routed to the 70B model, the effective cost reduction depends on the routing rate.

Routing assumptions: upfront routing versus 7B-first cascade

There are two broad routing approaches.

The first is upfront routing. In this model, requests are classified before inference and directed either to the 7B model or to the 70B model. Routine, low-risk, or well-understood requests go to the smaller model. Complex, ambiguous, or high-risk requests go directly to the larger model.

The second is a 7B-first cascade. In this model, all requests are initially run on the 7B model. Requests that cannot be resolved, or that trigger confidence or quality thresholds, are then escalated to the 70B model.

Both approaches can preserve material savings, but they have different trade-offs. Upfront routing avoids unnecessary double inference, but it requires strong workload classification before generation. A cascade is simpler to conceptualize and can allow the small model to resolve many requests first, but escalated cases incur additional overhead because they require a second inference pass.

The modeled results suggest that savings remain meaningful even when only a portion of requests can be handled by the smaller model. If 90% of requests are handled by the 7B model, the effective cost reduction is approximately 5.5x to 5.8x. At 75%, the reduction is approximately 3.0x to 3.2x. Even at 50%, the modeled reduction remains approximately 1.7x to 1.8x.

Economic takeaway

The economic logic is clear. Smaller distilled models can serve many defined workloads more efficiently, especially where task scope is known and quality thresholds can be measured. The strongest gains occur when routine requests can be confidently handled by the smaller model.

The cost saving is not just a function of model size. It depends on how reliably the system can determine which model should handle each task. Enterprises that route well can reduce cost without degrading quality. Enterprises that route poorly may save money in the short term but create hidden quality, reliability, and governance risks.

© Copyright Tolaga Research, 2026

Benchmark Evidence: The Limits of Distillation

Benchmark overview

Benchmark results demonstrate why distillation should not be treated as a universal substitute for larger models. The DeepSeek-R1 distilled model family provides a useful public example because it includes multiple student model sizes and results across different task types.

The comparison covers four evaluation tasks:

MATH-500 measures mathematical problem-solving. AIME 2024 is a more demanding competition-style mathematics benchmark. GPQA Diamond tests advanced graduate-level scientific reasoning. LiveCodeBench evaluates coding performance.

Distillation Quality Benchmarks

Smaller models retain a large share of math reasoning performance, but broader reasoning and coding performance degrade at 7B–8B scale.

Model	Size	MATH-500 (↑ higher)	AIME 2024 (↑ higher)	GPQA Diamond (↑ higher)	LiveCodeBench (↑ higher)
DeepSeek-R1 (teacher)	671B total / 37B active	97.3%	79.8%	71.5%	65.9%
R1-Distill-Llama-70B	70B	94.5%	86.7%	65.2%	57.5%
R1-Distill-Qwen-7B	7B	~83–93% (strong on math)	~55% (weaker)	— (not reported)	~37–38% (much lower)
R1-Distill-Llama-8B	8B	89.1% (strong on math)	— (not reported)	49.0% (weaker)	39.6% (much lower)

Key Takeaway

- Distillation preserves a large share of **math reasoning** capability, especially on MATH-500, even in 7B–8B models.
- **Broader reasoning** (AIME, GPQA) and **coding performance** drop substantially at 7B–8B scale.
- The **70B distilled model** shows the strongest cross-benchmark retention, approaching the teacher across all tasks.

Notes:

1. B = billion parameters. DeepSeek-R1 uses 671B total parameters with 37B active per token.

Source: DeepSeek-AI Technical Report & Official Model Cards

These benchmarks are useful because they show that distillation does not transfer capability evenly across domains. Performance can remain strong in one task category while degrading sharply in another.

© Copyright Tolaga Research, 2026

Strong math retention, weaker generalization

The results suggest that distillation can preserve a substantial share of mathematical reasoning performance, particularly when the student model remains relatively large. DeepSeek-R1-Distill-Llama-70B performs closest to the teacher model overall, with strong results on MATH-500 and AIME 2024.

However, performance is less consistent across broader reasoning and coding tasks. The 70B distilled model still trails the teacher model on GPQA Diamond and LiveCodeBench. This shows that even a large distilled model does not fully preserve frontier-model capability across all domains.

The smaller 7B and 8B distilled models remain relatively strong on MATH-500, but their broader reasoning and coding performance falls sharply. This is most visible in LiveCodeBench, where the smaller distilled models score around 37% to 40%, compared with 57.5% for the 70B distilled model and 65.9% for the teacher model.

The implication is not that small distilled models are useless. Rather, it is that their strengths are narrower. A small model may perform well on certain math benchmarks while still being weak on coding, scientific reasoning, or open-ended problem solving.

Why smaller distilled models need guardrails

These results should not be read as a permanent ceiling on distilled model performance. Modern LLM distillation techniques continue to improve through higher-quality synthetic data, reasoning traces, curriculum-style training, rejection sampling, reinforcement learning, and workload-specific fine-tuning. Future distilled models may perform better, especially when training data and evaluation processes are tightly matched to the target workload.

The main takeaway is that distillation transfers capability unevenly. It works best when the student model remains large enough, or when the target workload is narrow and well matched to the training signal. Small distilled models should not be assumed to retain general reasoning, scientific reasoning, or coding capability simply because they perform well on selected math benchmarks.

This is why guardrails matter. Smaller distilled models need clear task boundaries, confidence checks, escalation rules, and production monitoring. They should be deployed where their performance is measurable and where failure modes can be managed.

© Copyright Tolaga Research, 2026

Distillation and the Rise of Tiered Inference

The logic of tiered inference

Distillation makes tiered inference more practical by creating lower-cost model tiers for routine, bounded, and well-understood tasks. The architecture is not designed to replace frontier models. It is designed to reserve them for ambiguous, novel, high-risk, or complex work where their additional capability is justified.

This changes the deployment model for AI capability. Intelligence becomes stratified across model tiers, with each tier matched to a different class of workload, cost target, latency requirement, and risk profile.

Routing becomes a strategic capability

Tiered inference only works if routing decisions are reliable. The system needs workload classification to identify task type and complexity. It needs evaluation benchmarks that reflect production conditions. It needs confidence and uncertainty signals, even if those signals are imperfect. It needs fallback mechanisms for unresolved or high-risk cases. It also needs monitoring to detect quality degradation over time.

A poorly designed routing layer can undermine the economics of distillation. If too many requests are routed to the frontier model, savings shrink. If too many difficult requests are routed to the smaller model, quality degrades. The optimal system balances cost, latency, quality, and risk.

Routing therefore becomes one of the most important components of the AI inference stack. The question is not simply which model is best. The question is which model is best for this request, under these constraints, and at this level of risk.

Competitive advantage moves to the inference system

The advantage lies in the full inference system, not in model size alone. Effective deployment requires proprietary data for local context, workload-specific evaluation to detect degradation, runtime controls for cost and latency, fallback mechanisms for escalation, and governance processes that make deployment safe at the application level.

This shifts the source of competitive advantage. Frontier labs retain an important role in creating new capabilities and pushing the boundary of model performance. But enterprises and platform providers can create value by adapting those capabilities into efficient, reliable, and well-governed deployment systems.

The winners will not necessarily be those that use the largest model for every task. They will be those that deploy the right level of intelligence for each workload, measure performance rigorously, and manage escalation when smaller models are not good enough.

© Copyright Tolaga Research, 2026

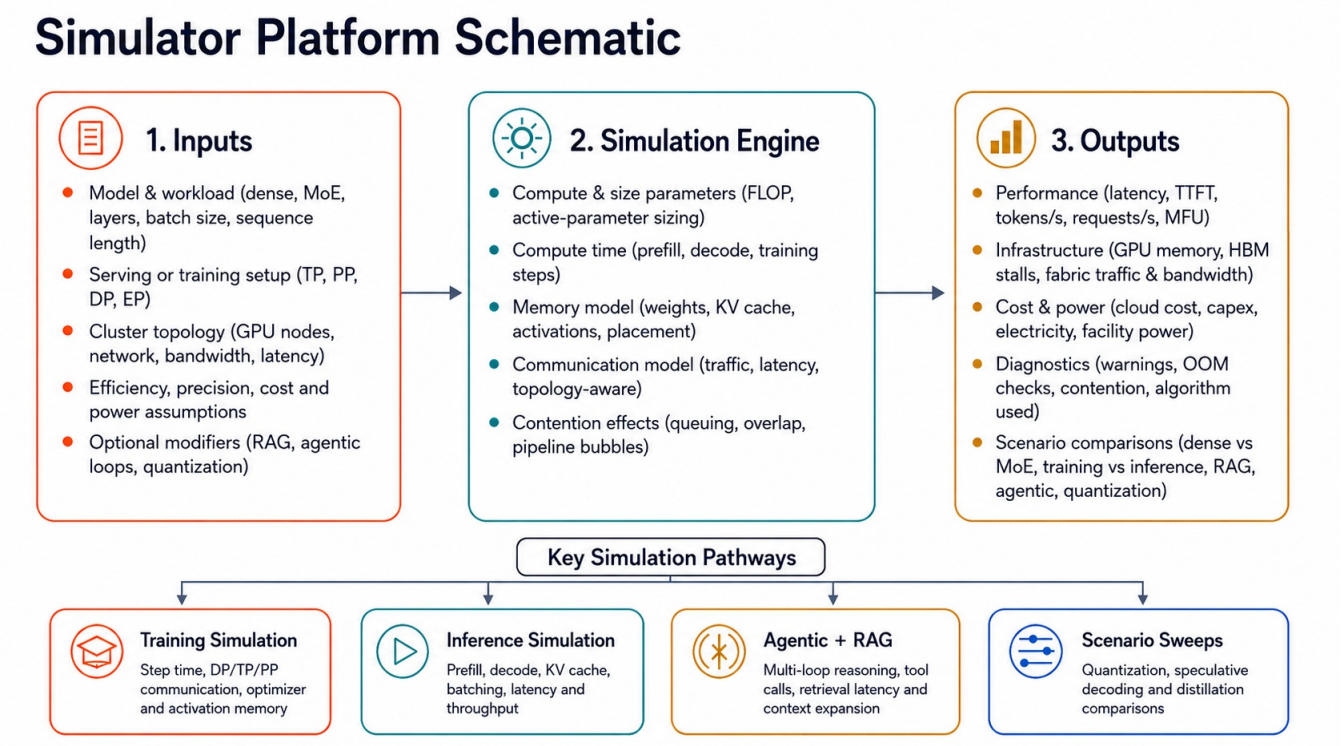
Conclusion

Model distillation is best understood as an inference economics tool. It does not eliminate the need for frontier models, but it can reduce unnecessary frontier-model usage when workloads are narrow, measurable, and well matched to the smaller model.

The larger implication is the rise of tiered inference. Frontier models remain essential for new capability creation, difficult cases, quality assurance, and high-risk reasoning. Distilled models handle efficient execution where their performance is sufficient. The critical layer between them is routing.

As AI deployment matures, the advantage will lie less in using the largest model everywhere and more in building the data, evaluation, routing, monitoring, fallback, and governance systems that determine how intelligence is applied safely and economically.

Tolaga AI Infrastructure Simulator



LEARN MORE

© Copyright Tolaga Research, 2026

About the Author: Phil Marshall, PhD

Phil Marshall's work in AI began more than 30 years ago with Kohonen maps, discrete neural networks, machine learning, and fuzzy logic. More recently, his work has targeted AI-enabled solutions for the automotive, healthcare, manufacturing, and smart city sectors. Over the past decade, he has applied natural language processing and large language models to analyze online content and generate distinctive, data-driven industry research insights.

Marshall also brings deep experience in telecom, wireless, mobile, and networking technologies, fields he has worked in since the early 1990s. More recently, he has combined this infrastructure background with advanced software and modeling capabilities to develop Tolaga's AI Infrastructure Simulator, which is used to analyze the performance, cost, and scalability of emerging AI architectures.

Marshall is a dual citizen of the United States and New Zealand and currently resides in Auckland, New Zealand.

Contact: philip.marshall@tolaga.com

© Copyright Tolaga Research, 2026