Frontiers of AI Research in 2025

The entry of DeepSeek into a crowded AI model lineup has been nerve-racking for markets and businesses. Executives and investors are asking, “Have we bet on the right model? Should we switch to new models driven more by open source or open weights, like DeepSeek, that are cheaper, faster and more efficient?” But, all this handwringing may be missing a crucial point. Our discussions with researchers, AI labs, venture capitalists and startups at the frontiers of AI infrastructure development point to a growing reality in the breathtaking evolution of AI: The differences among the models may be growing so marginal that they will not remain a differentiator. What will matter now is how businesses create value with them.

The rapid proliferation of use cases and paradigm shifts in the application of AI discussed in the previous article have been fueled by the increasing performance and multi-modality of the advancing AI foundational models that have been released by large tech companies over the last five years. This second article in the series takes a deeper dive into the current state of AI model development, performance concerns, diverse and hybrid approaches in AI model development and new advances.

Diverse Approaches Build Upon Foundational AI Model Advances

In the decade since the advent of the transformer class of deep learning AI models — the foundation for large language models (“LLMs”) — and the diffusion class of models that underpin image and video generation AI, advances in the performance of these foundational models have relied on scaling increasing computational resources and data for training.[1]

There is early evidence to suggest we now may be nearing the higher reaches of the scaling S-curve for the current class of foundational models, based on traditional scaling methods that draw on increased computational resources and data available for training these models.[2] This maturity is evident from the results on yield that have been coming from research labs in the last few months.[3]

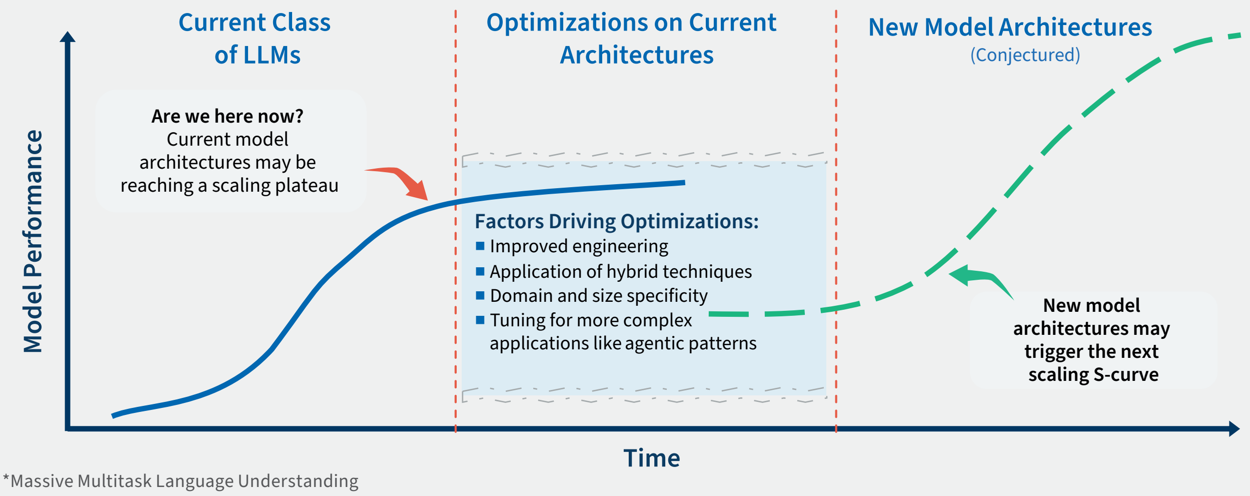
At the same time, we are seeing a combination of R&D activities that will push things forward: new engineering approaches to deliver mid-term increase in performance and/or to reduce cost; hybrid AI model approaches to serve increasingly complex applications; and continuing

> *“There is early evidence to suggest we now may be nearing the higher reaches of the scaling S-curve for the current class of foundational models.”*

fundamental research on new classes of models that may coexist or supersede the current foundational models in the longer term. The DeepSeek R1 release has been a remarkable validation of this state of affairs, where that team released multiple novel techniques within a single model release. Some specific illustrations of the major R&D activities include:

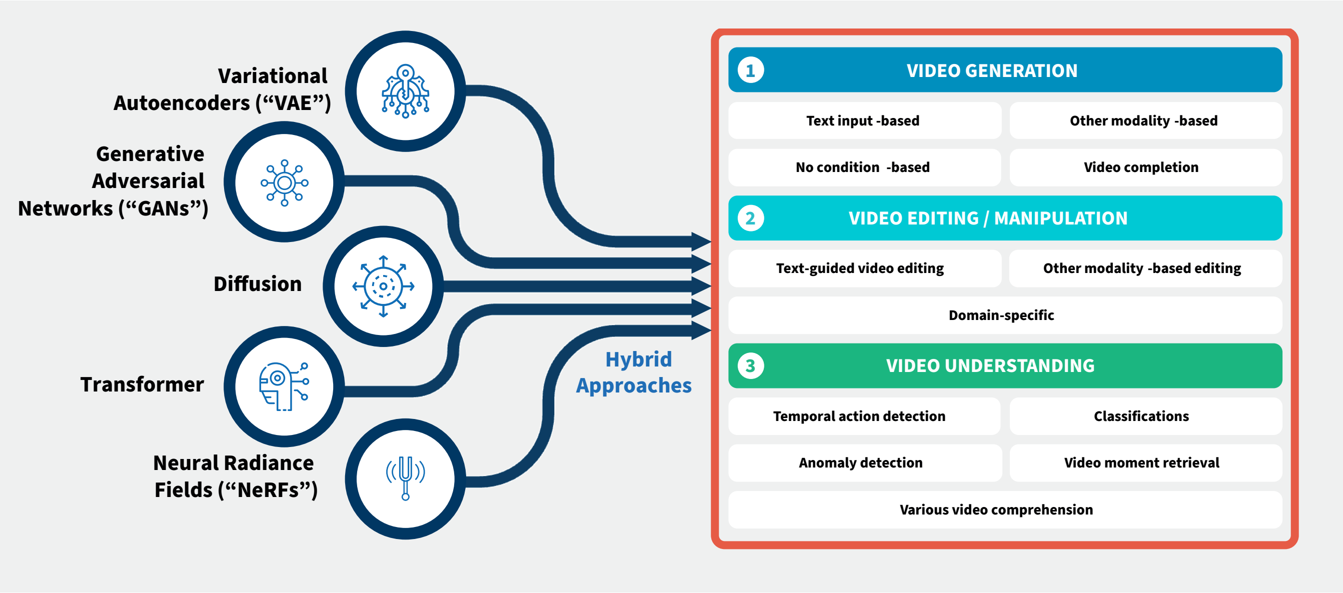
Performance of LLM Models against MMLU* Benchmarks

Benchmark evaluates an LLM's overall knowledge and ability to solve problems across 57 subjects, i.e. math, computer science, law, history, etc.


Current Class of LLMs
Optimizations on Current Architectures
New Model Architectures
(Conjectured)
Are we here now? Current model architectures may be reaching a scaling plateau
Factors Driving Optimizations:
■ Improved engineering
■ Application of hybrid techniques
■ Domain and size specificity
■ Tuning for more complex applications like agentic patterns
New model architectures may trigger the next scaling S-curve
Model Performance
Time

*Massive Multitask Language Understanding

- An increasing **focus on test time computation**, that is, on inferencing versus training. This approach allows scaling computation as part of results generation, which effectively allows models to "think longer" and produce better results. Multiple approaches are being applied toward this, including prompting models to explain their reasoning and then critiquing and correcting those intermediate steps, or asking models for multiple responses to the same query and then selecting a winner. These approaches are an important aspect of AI agents, which depend on this process of reasoning and self-critiquing to prepare effective plans of action.
- The rise of **small foundational models ("SFM")** and **specialized foundational models ("SpFMs")**. Many use cases do not require the sophistication, general reasoning or depth of the natural language processing capabilities of LLMs. Small models are being developed to train narrow problems and to be deployable on lower compute device including edge devices. Specialized models are being developed for highly specialized and complex functions, such as in the medical field, and although they do not require the general reasoning of LLMs, they do require highly specialized and domain-specific knowledge and reasoning for those fields.
- Research combining **stochastic AI** (deep learning models like GenAI) with **deterministic AI** involves traditional symbolic logic and rules-based systems that serve to apply constraints, guardrails and workflow structure, and that can provide more explanations for outcomes that are sensitive and subject to regulatory and compliance risks. Paradigms involving such a combination — like causal AI and neurosymbolic AI — are increasingly taking hold.
- A combination of model approaches will be applied to build **new model variants** to solve even more complex problems. In addition to combining rules-based systems with predictive LLM-based systems as discussed earlier, we are seeing different foundational model classes being commingled for specific purposes. Take video and 3D generation, extraction and understanding, for example, where autoregressive transformer architecture at the foundation of most LLMs with diffusion is being combined with neural radiance field-class ("NeRF") models to solve use cases for video manipulation or to form a deeper comprehension of footage through extraction of temporal action and character movements. There are scores of hybrid model variants in experimentation for tackling these areas alone.[4]


Variational Autoencoders ("VAE")
Generative Adversarial Networks ("GANs")
Diffusion
Transformer
Neural Radiance Fields ("NeRFs")
Hybrid Approaches
1 VIDEO GENERATION
Text input -based
Other modality -based
No condition -based
Video completion
2 VIDEO EDITING / MANIPULATION
Text-guided video editing
Other modality -based editing
Domain-specific
3 VIDEO UNDERSTANDING
Temporal action detection
Classifications
Anomaly detection
Video moment retrieval
Various video comprehension

- Another area of focus is in developing **large action models ("LAMs")** and, more recently, **large concept models ("LCMs")**, which are a modification of foundational models to support action-oriented responses and more complex reasoning in support of the agentic frameworks earlier discussed. We expect companies to invest in business and domain-specific LAMs and LCMs to increasingly support semi- and fully-autonomous workflows. For example, Salesforce has released xLAM to support Agentforce, their platform to deploy AI agents around sales and CRM functions. Meta has recently shared research on improving abstract reasoning through a new class of LCMs.[5]
- Models focusing on **machine-to-machine or agent-to-agent interaction** will become more prevalent as interactions between platforms with classical machine learning techniques accelerate due to the proliferation of use cases involving AI-to-AI agent workflows on both sides of a transaction. This inter-platform acceleration will be an important turning point for many reasons. First, a lot of training, compute and data for foundational models is expended on handling the human interface. With that interface out of consideration for large tracts of machine interaction, the models could become simpler, faster and cheaper. Second, inter-platform acceleration will change how B2C and B2B transactions occur among consumers and enterprises. Take search advertising, for example. Perplexity AI is exploring how AI agents that perform a task, say, purchasing a product on behalf of a consumer, may now consume advertising produced by another AI agent.[6] In other words, ad monetization of human attention could shift to ad monetization of agent attention.
- Lastly, there will be an emergence of **completely new model classes** that will continue to drive the next long-range advances in AI performance. For example, we have seen early-stage research where diffusion-like models are being evolved to support LLM-like behaviors with significant increase in inferencing quality for a fraction of the cost.

Model Economics and Explainability Drives Scaling Viability

A key positive development underlying innovation is the shifting model economics, which will bring the agentic workflow capabilities necessary for decision-making and human-AI collaboration closer to reality.

> *"LLM inferencing costs have declined faster in today's AI inflection than compute and network bandwidth costs did during the computer and internet revolutions."*

While much discussion has focused on model training costs and the level of infrastructure and power that will be needed to support this level of growth, a less discussed development is the significant drop in inferencing costs (i.e., the cost to execute a model in production runtime). LLM inferencing costs, for example, have declined faster in today's AI inflection than compute and network bandwidth costs did during the computer and internet revolutions. Taking constant performance levels, as measured by a metric called MMLU, the inferencing cost per million tokens (how throughput is measured) has decreased by a factor of 1000x since 2021. Then, GPT-3 cost an estimated $60 per million tokens; now, for the lowest cost open source models based on Llama, that cost is down to a few cents per million tokens, because of a combination of reduced GPU costs, software optimization, better model tuning, smaller models and more available open source models and tools.[7,8]

This deflation of model runtime costs is an extremely important development in relation to the move to agentic workflows. For comparison, human speech can be quantified as representing close to 10,000 tokens per hour. The math would imply, then, that deploying an AI agent attempting to replicate a human persona could cost less than two dollars at the current cost profile. This reduction in cost is significant because beyond deploying individual agents, it allows the unit economics to work where a coalition of multiple agents can collaborate to complete or augment increasingly complex enterprise workflows.

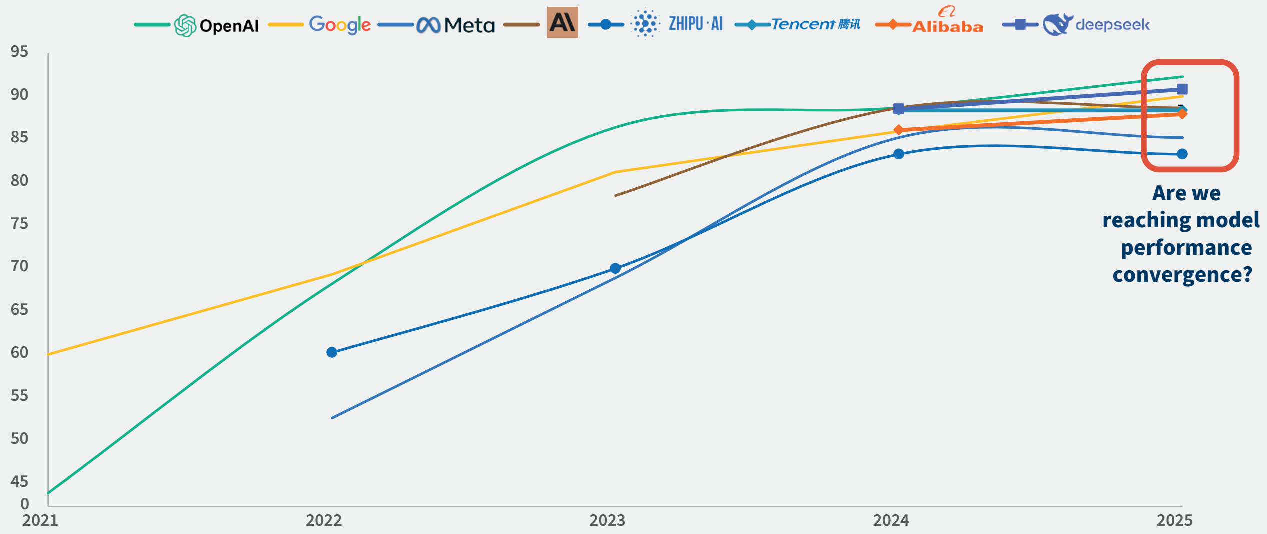
This deflation is also important because, as referenced earlier, as scaling laws for training potentially approach diminishing returns, there is an increasing focus to scale computation at inferencing, which allows models to think longer. Lower inferencing costs allow scaling applications where models can spend more time reasoning, self-critiquing and correcting themselves to deliver better quality responses and actions.

Explainability frameworks represent another important area of model development; this refers to the ability to explain how models predict or take the actions they do. Deep learning models, like ones the current class of foundational models are based on, are not currently explainable due to their intrinsic design, complexity, scale and vast training data, and a key impediment for deploying AI in regulated industries like healthcare and financial services is the unexplainable emergent behavior from these models. Extensive experimentation, research and development is now underway to develop explainable frameworks, like chain-of-thought ("CoT") prompting, to enable the models to explain how they reason and carry out their actions.

Are AI Models the New Electricity?

We are seeing decreasing costs, increasing complexity and corresponding use cases, as well as a proliferation of models and continued innovation into new areas. So, are we in fact seeing the commoditization of AI models? Those businesses looking for large-scale early investment are trying to pick a winner with the potential for huge market share and a strategic moat. But, current model classes are largely all based on openly available fundamental deep learning research, increasingly homogeneous and similar datasets and levels of compute. Indeed, if we look at model leaderboards, new models

Highest MMLU* Scores Achieved by Major U.S. and Asian Companies Over Time



Are we reaching model performance convergence?

*Massive Multitask Language Understanding have emerged rapidly, closing with the most advanced OpenAI models. This includes severe competition from around the world, particularly from Asia, where they are showing increasing parity for significantly lower factors of compute, as exemplified by the DeekSeek R1 release.

Winner-like patterns, if they emerge at all, will likely result from an early mover advantage, the ability to integrate into enterprise and consumer products and the deployment scaffolding built around it, as well as the core model innovation itself.

A key question to ask is: Are models a new and exciting but increasingly commoditized infrastructure? Are they the new internal combustion engine, the new electricity, the new network pipes? If so, what will your business do with it?

Of course, if a model achieved artificial general intelligence ("AGI"), that is, human-level reasoning, then it would be a game-changer. But, how close are we? Our discussions with a wide range of researchers indicate that there is currently no consensus, partly because there are competing frameworks for measuring AI maturity, but more importantly because there is not a definitive marker for what constitutes general human intelligence and therefore what AGI must achieve.

> *"A key question to ask is: Are models a new and exciting but increasingly commoditized infrastructure? Are they the new internal combustion engine, the new electricity, the new network pipes? If so, what will your business do with it?"*

So, what are markets and the investment community to do in this evolving world and with the commoditization of AI models a real possibility? Will the focus move, as we believe it will, from models to use cases? Our next article will examine how markets and the investment community are reacting and adapting to technical advances in AI and the application of AI to business transformation.

Endnotes

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei, "Scaling Laws for Neural Language Models", OpenAI, (January 23, 2020), http://arxiv.org/pdf/2001.08361.

[2] Rachel Metz, Shirin Ghaffary, Dina Bass, and Luia Love, "Google and Anthropic Are Struggling to Build More Advanced AI", Bloomberg, (November 13, 2024), https://www.bloomberg.com/news/articles/2024-11-13/openai-google-and-anthropic-are-struggling-to-build-more-advanced-ai.

[3] Charles Luo, "Has LLM Reached the Scaling Ceiling Yet? Unified Insights into LLM Regularities and Constraints", "This maturity is evident from the results on yield that have been coming from research labs in the last few months", Cornell University, (December 21, 2024), https://arxiv.org/abs/2412.16443.

[4] Andrew Melnik, Michael Ljubljanac, Cong Lu, Qi Yan, Weiming Ren, Helge Ritter, "Video Diffusion Models: A Survey", Cornell University, (May 6, 2024), https://arxiv.org/abs/2405.03150.

[5] Loic Barrault et. al., "Large Concept Models: Language Modeling in a Sentence Representation Space", Meta, (December 11, 2024), https://ai.meta.com/research/publications/large-concept-models-language-modeling-in-a-sentence-representation-space/.

[6] Luís Rijo, "A new model for AI advertising emerges as agents could replace human attention", PPC Land, (January 1, 2025), https://ppc.land/a-new-model-for-ai-advertising-emerges-as-agents-could-replace-human-attention/.

[7] OpenAI, "Deprecations", OpenAI, accessed on Feb 20, 2025, https://platform.openai.com/docs/deprecations/instructgpt-models.

[8] Together AI, https://www.together.ai/pricing.

SUMEET GUPTA
Senior Managing Director
Leader of Digital Transformation, Leader of AI Transformation
sumeet.gupta@fticonsulting.com

The views expressed herein are those of the author(s) and not necessarily the views of FTI Consulting, Inc., its management, its subsidiaries, its affiliates, or its other professionals. FTI Consulting, Inc., including its subsidiaries and affiliates, is a consulting firm and is not a certified public accounting firm or a law firm.

FTI Consulting is an independent global business advisory firm dedicated to helping organisations manage change, mitigate risk and resolve disputes: financial, legal, operational, political & regulatory, reputational and transactional. FTI Consulting professionals, located in all major business centres throughout the world, work closely with clients to anticipate, illuminate and overcome complex business challenges and opportunities. **fticonsulting.com**

FTI™ CONSULTING