State of Play for Select Open AI Models Globally

Chapeaux Note: 13 October 2025

Dr. Syed Muntasir Mamun

Abstract

In 2025, the global open AI model ecosystem has undergone a profound transformation, marked by the ascendancy of Chinese-origin models over their American counterparts in adoption, performance, and innovation. Chinese labs, led by prolific players like Qwen and DeepSeek, have released over 20 major models, capturing more than 50% of fine-tune shares and surpassing U.S. cumulative downloads by August, driven by platform-oriented strategies that prioritize market share, community engagement, and multimodal versatility. American models, including Llama 4, Gemma 3, and GPT-OSS, face declining trust due to inconsistent openness and fragmented investments, resulting in slower evolution and reduced researcher access. Other models, primarily from fading European efforts like Mistral, play a marginal role with limited scale. This comparative analysis highlights key parameters—releases, adoption, performance, strategy, and contributions to idea generation—revealing how open models fuel AI's evolution through democratized experimentation and original thought, while emphasizing the need for Western investments to maintain competitive edges in innovation, safety, and international influence. Without such commitments, academic and societal progress risks stagnation amid concentrated closed-model power.

Table of Contents

The Shifting Sands of AI: Chinese Models Dominate Open AI Landscape in 2025

The year 2025 marks a pivotal moment in the evolution of artificial intelligence, as the open AI landscape undergoes a dramatic transformation. Chinese-origin models have unequivocally emerged as the dominant force, outperforming their American counterparts across key metrics including adoption, performance, and innovation. This significant shift is not merely a technological phenomenon but a reflection of deeper geopolitical and economic currents that are reshaping the global AI ecosystem.The Rise of Chinese Open AI: A Confluence of Strategy and Execution

The ascendancy of Chinese models can be attributed to a combination of aggressive release strategies and a relentless focus on market penetration. By prioritizing rapid deployment and widespread accessibility, Chinese developers have successfully fostered global adoption, positioning their models as versatile platforms capable of supporting a diverse array of applications. This proactive approach has created a virtuous cycle of user engagement and iterative improvement, propelling these models to the forefront of AI development.

Furthermore, Chinese investment in open-source ecosystems has been instrumental in cultivating a vibrant community of researchers, developers, and startups. This collaborative environment has stimulated a surge of original ideas and novel applications, ranging from sophisticated agentic tools and multimodal integrations to highly specialized fine-tuned models. The collective intelligence of this burgeoning community has accelerated the pace of innovation, allowing Chinese open AI to continually push the boundaries of what is possible.The American Lag: Inconsistent Commitment and Fragmented Efforts

In stark contrast, American models, despite their historical influence and foundational contributions to AI, have experienced a relative decline in the open AI arena. This setback can be largely attributed to inconsistent commitments to openness, fragmented developmental efforts, and a slower adaptation to the evolving needs of the AI community. While individual breakthroughs have occurred, the lack of a unified and sustained strategy for open-source AI has hampered their ability to compete effectively on a global scale.

The consequences of this American lag are far-reaching. Academics and researchers in the U.S. risk being confined to studying outdated models, which stifles original thought and impedes long-term innovation. This not only diminishes their ability to contribute

cutting-edge research but also risks a concentration of power in a few closed labs, limiting the broader democratization of AI advancements.The Power of Openness: Driving Innovation and Democratizing Access

The current landscape underscores a fundamental truth: open models are indispensable for fostering original ideas and maintaining a competitive edge in AI evolution. By democratizing access to advanced AI capabilities, these models empower a wider range of stakeholders—from independent researchers to burgeoning startups—to experiment, iterate, and generate novel applications. This collaborative environment fosters a rapid diffusion of ideas, leading to breakthroughs that closed models, by their very nature, cannot match.

The applications stemming from open models are diverse and transformative. Agentic tools, capable of autonomous decision-making and task execution, are becoming increasingly sophisticated. Multimodal integrations, seamlessly blending various data types like text, images, and audio, are unlocking new levels of contextual understanding. Specialized fine-tunes are allowing for highly customized AI solutions tailored to specific industries and challenges. These advancements collectively propel the field forward at an unprecedented pace.Navigating the Challenges: Control, Safety, and National Strategy

While openness offers immense benefits, it also presents significant challenges, particularly concerning control and safety as models approach frontier capabilities. The widespread accessibility of powerful AI tools necessitates robust frameworks for ethical deployment, bias mitigation, and responsible governance. Addressing these concerns is paramount to ensuring that the benefits of open AI are realized without compromising societal well-being.

For nations and organizations to maintain a competitive edge in this rapidly evolving domain, strategic investment in open ecosystems is crucial. This includes cultivating researcher viability through funding and support, mitigating the concentration of power in closed labs, and actively influencing international markets. China's current success serves as a compelling testament to how sustained openness can build momentum, creating a powerful feedback loop of user engagement and iterative improvements that sustains leadership. The lessons from 2025 are clear: the future of AI belongs to those who embrace openness, collaboration, and a strategic vision for a democratized and innovative AI landscape.

Focus on Chinese-Origin Models

China's Ascendant AI Landscape: A Deep Dive into the 2025 Open Model Ecosystem

The year 2025 has marked a significant turning point in the global artificial intelligence landscape, with China emerging as a dominant force in the proliferation and adoption of open models. The sheer volume and diversity of contributions from over 20 notable Chinese organizations have led to an unprecedented surge in releases, spanning a wide array of model sizes, modalities, and crucial use cases. This dynamic ecosystem is rapidly reshaping the competitive landscape and setting new benchmarks for innovation and accessibility.

Key Players and Their Impact:

At the forefront of this expansion are key players such as Qwen, a formidable offering from Alibaba, and DeepSeek. Qwen, in particular, has demonstrated an astonishing rate of development and impact. Its frequent, high-quality releases, exemplified by the launch of Qwen3 in April and its subsequent specialized variants like Qwen3-Coder (tailored for advanced coding applications) and Qwen3-Omni (designed for broad, multimodal capabilities), have collectively matched the output of the entire American open model ecosystem. This rapid iteration and specialization underscore a strategic approach to comprehensive market coverage.

The "Many Models as a Platform" Strategy:

China's approach to AI development mirrors the successful "many models as a platform" paradigm, akin to the broad and diverse ecosystem fostered by Android in the mobile operating system space. This strategy prioritizes community engagement, recognizing that a vibrant developer community is crucial for sustained innovation. Rapid feedback integration is another cornerstone, allowing for quick improvements and adaptations based on real-world usage. The comprehensive coverage of this ecosystem is notable, extending from fundamental embedding models (essential for understanding and representing data) to advanced vision-language models (enabling AI to interpret both images and text) and sophisticated text-to-speech technologies.

Metrics of Dominance:

The adoption metrics unequivocally highlight China's growing dominance in the open-source AI arena. By mid-2025, Chinese models had captured over 50% share of fine-tuned models with more than five downloads, indicating a strong preference and active engagement from developers. Furthermore, cumulative Hugging Face downloads from Chinese laboratories surpassed those from the United States around August 2025, a critical milestone demonstrating a shift in global developer preference and access.

Performance-wise, independent evaluations, such as those conducted by ArtificialAnalysis, have consistently shown Chinese models surpassing global benchmarks, further solidifying their technical prowess.

Strategic Imperatives and Community Norms:

Strategically, China's aggressive open-source push is driven by a clear prioritization of market share over immediate financial profits. This long-term vision is reinforced by evolving community norms, which were significantly influenced by breakthroughs like DeepSeek R1 in January. This emphasis on open accessibility and collaboration has fostered an environment conducive to original idea generation. Examples include the development of advanced agentic coding agents, capable of independently writing and debugging code, and sophisticated multimodal reasoning capabilities that allow AI to process and understand information from various data types simultaneously. This open approach encourages global developers to build upon these accessible tools, thereby accelerating the overall evolution of AI.

Challenges and Considerations:

While the benefits of this rapid expansion are evident, potential risks and challenges remain. A significant concern is the potential for alignment with state values. For instance, DeepSeek R1's responses have been observed to promote "socialist harmony," which, while beneficial in regulated environments and for ensuring social cohesion, could potentially limit uncensored innovation. This raises questions about the long-term impact on the diversity of thought and the unfettered exploration of ideas within the AI development landscape. Balancing the promotion of national values with the need for unconstrained innovation will be a critical challenge for China's AI ecosystem moving forward.

The Special Case of the Alibaba Qwen Model

The Alibaba Qwen series, launched by the Qwen team at Alibaba Cloud, has emerged as a powerhouse in the open AI model landscape since its inception with Qwen 1.0 in 2023, evolving through Qwen 2 (2024), Qwen 2.5 (late 2024 to early 2025), and the flagship Qwen 3 suite in April 2025, followed by advanced variants like Qwen 3-Max and Qwen 3-Coder by September 2025 (QwenLM, 2025a; Lambert, 2025a). Positioned as versatile, high-performance large language models (LLMs) with a focus on open-source accessibility, Qwen models span dense and mixture-of-experts (MoE) architectures, supporting multilingual capabilities, multimodal integrations (e.g., vision-language in Qwen 3-VL), and specialized applications like coding and reasoning. By October 2025, Qwen has achieved unprecedented adoption, surpassing Meta's Llama in cumulative downloads and research usage, driven by frequent releases, strong benchmarks, and

permissive Apache 2.0 licensing (The ATOM Project, 2025; Lambert, 2025a). However, the series faces scrutiny for potential privacy risks, ethical biases tied to its Chinese origins, inconsistent real-world performance, and high computational demands. This critique evaluates the series' strengths, weaknesses, and implications, informed by recent reviews, benchmarks, and community discourse.

Strengths

Qwen's primary strength lies in its exceptional performance across benchmarks, often rivaling or exceeding frontier models from Western labs. Qwen 3, for instance, features diverse model sizes (from 0.6B to 235B parameters) with reasoning toggles, enabling smaller models like the 4B dense variant to compete with GPT-4 levels through algorithmic and data enhancements, including over 30T pretraining tokens for larger models (Lambert, 2025a). Post-training techniques, such as supervised fine-tuning (SFT) for chain-of-thought (CoT) behaviors and reinforcement learning (RL) inspired by DeepSeek R1, yield significant gains in reasoning, coding, and multimodal tasks, with inference-time scaling boosting scores dramatically (Lambert, 2025a; TechPoint Africa, 2025). Qwen 3-Max ranks in the global top three on LMArena text leaderboards, surpassing GPT-5-Chat in some metrics, while Qwen 3-Coder excels in coding benchmarks like SWE-rebench, matching GPT-5-High's pass@5 rate (32.4%) on real GitHub tasks (Dev.to, 2025a; Reddit, 2025a). Adoption in research is unparalleled, with Qwen 2.5 and 3 cited extensively due to their potency in fine-tuning, accessibility across sizes, and integration with libraries like Hugging Face and vLLM (Reddit, 2025b; SiliconFlow, 2025). Community feedback highlights Qwen's "smart" logical consistency, multilingual prowess (e.g., superior in non-English languages like Polish), and real-world utility in tasks like math, writing, and image generation, making it a go-to for developers and researchers seeking cost-effective alternatives to proprietary APIs (Reddit, 2025c; Medium, 2025).

Weaknesses

Despite its accolades, Qwen grapples with notable limitations, particularly in robustness, ethics, and practical deployment. Real-world coding performance has been described as "unimpressive" in zero-shot scenarios, with models like Qwen 3 hallucinating or struggling with library-specific tasks, requiring RAG (retrieval-augmented generation) or additional prompting to shine—issues exacerbated in smaller variants like the 7B Coder, which sometimes underperforms larger predecessors (Reddit, 2025d; Reddit, 2025e). Qwen 3 lacks native multimodality in its base form, lagging behind competitors like Llama 4 or GPT-4o in agentic tool use and vision, and its reasoning chains (often starting with "[Okay]") feel derivative, lacking the "taste" or vibes of models like DeepSeek R1 or Gemini 2.5 Pro (Lambert, 2025a; Reddit, 2025f). Ethical and security concerns loom large due to its Chinese origins: privacy pitfalls include potential data exfiltration risks in

inference engines, biases aligned with state values (e.g., promoting harmony in responses), and vulnerabilities to backdoors in weights or code generation (The Firewall Blog, 2025; Reddit, 2025g). High computational demands for larger models (e.g., Qwen 2.5-Max requiring substantial hardware) pose barriers for consumer use, and some users report superficial outputs in creative tasks, with models deemed "rubbish" for advanced searches or error-prone code (DigitalDefynd, 2025; Reddit, 2025h). Licensing, while permissive, mandates attributions in derivatives, adding minor friction, and base models for the largest MoEs remain unreleased, limiting full openness (Lambert, 2025a).

Comparison to Competitors

Qwen's trajectory in 2025 positions it as a leader among Chinese models, outpacing DeepSeek in release frequency and variety while challenging Western frontrunners. Compared to Llama, Qwen 3 offers more accessible sizes (e.g., strong 32B models) and has dethroned it in research and subreddit discussions, with superior reasoning and multilingual support, though Llama edges out in knowledge depth and sycophancy resistance (Reddit, 2025i; Lambert, 2025a). DeepSeek R1 and V3 provide stiffer competition in pure reasoning and robustness, but Qwen's broader ecosystem (e.g., Qwen 3-Omni for multimodality) and market-share focus make it more versatile, akin to Android versus iOS (Medium, 2025; Dev.to, 2025b). Against proprietary models like OpenAI's GPT-5 or Google's Gemini 2.5 Pro, Qwen holds its own in benchmarks (e.g., Qwen 3-Max topping text leaderboards) but falls short in context length, zero-shot creativity, and isolation from corporate risks, with users noting Qwen's edge in open-source affordability despite occasional "dumber" outputs in logic tasks (Reddit, 2025f; TechPoint Africa, 2025). European models like Mistral have faded, but Qwen's soft power in global adoption highlights China's dominance, though privacy concerns deter some Western enterprises compared to U.S.-based alternatives (The Firewall Blog, 2025; Reddit, 2025g).

Verdict as in October 2025

The Qwen series exemplifies China's aggressive push in open AI, delivering high-performance, research-friendly models that have redefined accessibility and innovation in 2025, with strengths in benchmarks, adoption, and multimodal extensions outweighing peers in many metrics (QwenLM, 2025b; Lambert, 2025a). However, weaknesses in real-world robustness, ethical alignments, and deployment hurdles underscore the need for cautious integration, particularly amid geopolitical tensions. For Alibaba to solidify Qwen's legacy, future iterations should prioritize multimodal natives, enhanced security audits, and broader robustness testing—potentially cementing it as the open standard while addressing criticisms that could otherwise hinder global trust (DigitalDefynd, 2025; Dev.to, 2025a).

Focus on American-Origin Models

The landscape of American-developed large language models (LLMs), once a dominant force, particularly with Meta's Llama series, has experienced a notable decline in influence. This downturn is largely attributed to inconsistent commitment from key players and a reduced frequency of significant model releases. While Llama 4's introduction in April was initially perceived as a crucial event, the subsequent failure of rumored updates like Llama 4.1/4.2 and an anticipated 8-billion parameter variant to materialize significantly eroded community trust and enthusiasm.

Despite this broader trend, other American entities have made notable contributions. Google introduced Gemma 3 in March, and OpenAI, in a move towards greater transparency, released GPT-OSS in August. This open-source iteration of GPT demonstrated strong capabilities in tool-use and reasoning, marking a significant step in the direction of more accessible advanced models. Nvidia contributed with Nemotron Nano, and the Allen Institute for AI (Ai2) unveiled OLMo 2 32B, a model that, through advanced data and algorithmic techniques, remarkably approximates the performance of the original GPT-4.

However, these individual successes haven't stemmed a broader decline in adoption. U.S. models now lag behind in both cumulative downloads and market share for fine-tuning. A growing number of users are migrating to alternatives like Qwen, largely due to its perceived reliability and consistent performance. While American models still maintain competitive edges in specialized niches such as reasoning tasks, they generally fall short in terms of overall variety and the integration of multimodal extensions, which are becoming increasingly crucial for modern AI applications.

Strategically, the U.S. approach to AI development appears fragmented. There's a noticeable emphasis on developing "single-path, high-quality" models or platform-centric solutions, exemplified by initiatives like Gemma. However, investments in these areas, such as Ai2's $100 million grant from the NSF and Nvidia for OLMo, are often insufficient to meet the substantial scaling requirements of cutting-edge AI research and deployment. This fragmented and underfunded landscape severely hinders the generation of original thought, as academic institutions and independent researchers struggle to access state-of-the-art open models. This lack of access risks making their contributions increasingly irrelevant compared to advancements occurring behind the closed doors of proprietary AI development.

To reclaim its leadership in the global AI race, the United States must critically re-evaluate its strategy. A paramount step is to prioritize sustained and substantial funding for fully open models—encompassing data, code, and weights. Such an investment is crucial for fostering innovation in vital areas like AI safety, promoting societal education regarding AI, and reasserting international influence. This strategic shift would also significantly reduce the current reliance on closed and proprietary systems like those developed by OpenAI, thereby fostering a more collaborative, transparent, and ultimately more innovative AI ecosystem.

The Curious Case of the Llama

The Meta Llama series, starting with the original LLaMA in February 2023 and evolving through Llama 2 (July 2023), Llama 3 (April 2024), and most recently Llama 4 (April 2025), represents one of the most influential open-weight large language model (LLM) families in AI development. Designed to democratize access to advanced AI, Llama models have been praised for their openness, efficiency, and broad applicability, but they have also faced mounting criticism for inconsistent performance, strategic missteps, and failure to keep pace with competitors. As of October 2025, Llama 4—comprising variants like the 17B-parameter Scout (with 16 experts) and Maverick—marks a pivot toward multimodal capabilities, handling text, images, audio, and more (Lambert, 2025a; Meta, 2025a). However, this evolution has not quelled concerns about Meta's waning leadership in open models, as evidenced by declining adoption, underwhelming benchmarks, and community skepticism. This critique examines the series' strengths, weaknesses, and broader implications, drawing on recent analyses and user feedback to provide a balanced assessment.

Strengths

Llama's core appeal lies in its commitment to open-source principles, which has historically driven massive adoption and innovation. Pre-2025, Llama dominated the open model space, with billions of downloads and widespread use in research, fine-tuning, and applications ranging from chatbots to code generation. Its permissive licensing (though increasingly restrictive in later versions) allowed developers to build upon it freely, fostering an ecosystem of derivatives and integrations (Meta, 2024). Llama 4 builds on this by introducing native multimodality, enabling more personalized experiences in areas like video generation (e.g., via Meta Movie Gen) and cross-modal reasoning, which sets it apart from purely text-based predecessors (Meta, 2025b; Lambert, 2025b). Efficiency is another highlight: Models like Llama 4 Scout offer strong performance at smaller scales, making them viable for resource-constrained environments, and Meta's May 2025 "Llama for

Startups" program provides support to encourage enterprise adoption (Meta, 2025c; Hugging Face, 2025). In benchmarks, Llama excels in certain niches, such as handling diverse modalities, and its integration with platforms like Hugging Face and Amazon Bedrock has made it accessible for deployment (Lambert, 2025a; Amazon, 2025). These attributes have positioned Llama as a tool for reducing AI power concentration, empowering global developers to innovate without relying on closed APIs from giants like OpenAI or Google.

Weaknesses

Despite these merits, Llama has drawn sharp criticism for its performance shortcomings and strategic inconsistencies. Llama 4's reception has been notably lukewarm, with early assessments highlighting underwhelming results on key benchmarks like ArtificialAnalysis, where the 400B Maverick variant scores lower than smaller competitors such as Qwen's QwQ-32B or Gemma 3 27B (ArtificialAnalysis, 2025; Lambert, 2025c). Critics argue that Llama 4 lacks novel architectural innovations, relying on scaled-up mixtures of experts (MoE) without closing gaps in coding, reasoning, or "soft intelligence" compared to models like DeepSeek V3, Claude 4.5, or Gemini 2.5 Pro (Lambert, 2025d; Anthropic, 2025). Adoption has declined sharply in 2025, with users shifting to Chinese alternatives like Qwen and DeepSeek due to Meta's broken promises—rumors of Llama 4.1, 4.2, and an 8B variant never materialized, eroding community trust (Lambert, 2025e; Reddit, 2025a). Licensing changes, including restrictions on the 2T-parameter "Behemoth" preview, have introduced unnecessary friction, deterring open-source enthusiasts who prefer MIT-licensed options (Lambert, 2025f; GitHub, 2025). Bias issues have also surfaced, with Meta's efforts to mitigate "woke" AI perceived as politically motivated rather than neutral, potentially compromising model fairness (Ethics in AI Forum, 2025). Broader critiques point to Meta's talent and compute advantages not translating into sustained leadership, as the series falls behind in scalability and real-world utility, with some calling Llama 4 "objectively horrible" and "atrocious for its size" (Reddit, 2025b; X Community, 2025a).

Comparison to Competitors

In the global open model landscape, Llama's trajectory contrasts starkly with rising stars. Chinese models like Qwen and DeepSeek have overtaken Llama in cumulative downloads and fine-tune shares, offering superior performance, frequent updates, and unrestricted licenses that encourage broader experimentation (Lambert, 2025g; The ATOM Project, 2025). Google's Gemma 3 and Nvidia's Nemotron provide more consistent efficiency, while closed models from Anthropic (Claude) and Google (Gemini) outshine Llama in reasoning and multimodal tasks (Google, 2025; Lambert, 2025a). European efforts like Mistral have faded, but Llama's issues highlight a Western lag: Meta's shift toward "superintelligence" labs and cautious openness has ceded ground, as seen in Zuckerberg's

August 2025 comments prioritizing risk mitigation over rapid iteration (Meta, 2025d). User sentiment on platforms like Reddit and X echoes this, with developers lamenting Llama's "smell" of stagnation and calling for more innovative releases to regain momentum (Reddit, 2025c; X Community, 2025b; X Community, 2025c).

Verdict as in October 2025

The Llama series has undeniably advanced open AI by providing accessible, high-quality models that spurred global innovation, but its 2025 iteration reveals deepening flaws in execution, performance, and strategy. While multimodal enhancements in Llama 4 offer promise, the model's failure to innovate, coupled with trust-eroding delays and restrictive policies, has diminished its competitive edge in a field dominated by more agile rivals (Interconnects.ai, 2025; Lambert, 2025e). For Meta to reclaim relevance, it must prioritize transparent roadmaps, unrestricted openness, and benchmark-beating advancements—otherwise, Llama risks becoming a cautionary tale of squandered potential in the fast-evolving AI ecosystem (Reddit, 2025b).

Models from Other Regions

Beyond the duopoly of Chinese and American artificial intelligence, a diverse yet struggling landscape of open models from other regions persists. While these efforts represent a smaller share of the global AI ecosystem, they offer glimpses into alternative development paths and priorities.

European contributions, particularly from France's Mistral AI, showed early and significant promise. Their initial releases were met with enthusiasm, demonstrating strong performance and a focus on high-quality, efficient models. However, by 2025, Mistral AI's momentum had significantly waned. They struggled to maintain their competitive edge, lacking major updates and losing adoption to more prolific and better-resourced competitors, primarily from the US and China. Mistral's strategic focus on efficiency and quality, while admirable, ultimately proved insufficient to overcome the substantial resource constraints faced when competing against state-backed or venture-capital-rich giants. This highlights a critical challenge for non-dominant players: sustaining innovation and market presence in a rapidly evolving, capital-intensive field.

Other global players include scattered and often nascent efforts from labs outside the US and China. While there is potential for significant releases from countries like India or South Korea, the document explicitly states that no major breakthroughs outside the dominant duo materialized in 2025. These "other" models often adopt hybrid strategies,

focusing on specific research niches or specialized applications where they might find a competitive advantage. Examples include models from organizations like OpenBMB or BAAI, which, despite having Chinese affiliations, often operate with a more research-oriented mandate, blurring the lines between national ecosystems.

The primary struggle for these non-dominant models lies in achieving widespread adoption. Their limited scale, both in terms of computational resources and developer communities, severely hinders their ability to compete with the extensive support and rapid iteration cycles of larger players. While they contribute to the broader evolution of AI through specialized innovations—such as developing highly efficient small models that can run on less powerful hardware—their overall role remains marginal.

This marginalization underscores a critical need for broader international investments to foster a more competitive and diverse global AI landscape. Without such support, the current trajectory points towards a bipolar dominance by the US and China. Such a concentrated power structure could stifle diverse perspectives, limit the range of ethical and application-specific considerations in AI development, and ultimately hinder the global progress of artificial intelligence by reducing the variety of approaches and solutions being explored. Preventing this bipolar dominance is crucial for ensuring that the benefits of AI are distributed equitably and that the technology develops in a way that serves a wider array of societal needs and values.

Comparative Table on Various Parameters

Parameter	Chinese-Origin Models	American-Origin Models	Other Models (e.g., European)
Major Releases in 2025	High volume: DeepSeek R1 (Jan), Qwen3 (Apr), MiniMax-M1/Baidu ERNIE 4.5 (Jun), Kimi K2/GLM 4.5/StepFun Step3 (Jul), Qwen Next/Omni/VL (Sep); ~20+ organizations.	Moderate: Gemma 3 (Mar), Llama 4 (Apr), GPT-OSS/Nvidia Nemotron Nano/Seed-OSS (Aug); Fewer, inconsistent follow-ups.	Low: Mistral fading with no major 2025 releases noted; sporadic from others.

Adoption (Downloads & Fine-Tunes)	Dominant: >50% fine-tune share; Cumulative HF downloads overtook U.S. ~Aug 2025; Qwen surpasses Llama as most used family.	Declining: Trailing in cumulative downloads; Users switching to Qwen for fine-tuning.	Minimal: Early adoption for Mistral, but now negligible.
Performance	Leading: Surpassed global benchmarks (ArtificialAnalysis scores); Strong in multimodal, agentic, and reasoning (e.g., Qwen3-Coder-480B).	Competitive in niches: OLMo 2 32B ~GPT-4; Lags in variety and scaling (e.g., MoE).	Variable: Mistral efficient but outdated; Others niche-specific.
Strategy & Variety	Platform-oriented (e.g., Qwen as "Android"): Many sizes/modalities; Single-path high-quality (e.g., DeepSeek); Market share focus.	Mixed: Platforms (Gemma/Llama) vs. high-quality (GPT-OSS); Wavering openness erodes trust.	Hybrid/research-focused: Efficient but limited scale and engagement.
Role in Idea Generation & AI Evolution	High: Enables global experimentation, feedback loops; Fosters original multimodal/agentic innovations; Community norms enforce openness.	Moderate: Supports researcher access but fragmented; Risks academic irrelevance without more investment.	Low: Contributes niche ideas but lacks momentum for broad evolution.

Competitive Edge Maintenance	Strong: Low-cost access drives international adoption/soft power; Sustained releases build ecosystem.	Weakening: Needs more funding (e.g., beyond Ai2's $100M) to reduce power concentration and influence markets.	Limited: Relies on alliances; Vulnerable to dominance by China/U.S.

Benchmark Analytics between LlAMA and DeepSeek

Comparative Statement on Llama and DeepSeek AI Models

In 2025, the open AI model landscape pits Meta's Llama series against China's DeepSeek models in a contest of innovation, accessibility, and performance, with DeepSeek emerging as a disruptive force challenging Llama's historical dominance. Llama, evolving to Llama 4 in April, emphasizes scalable, multimodal architectures with Mixture of Experts (MoE) designs for efficiency and broad applicability, but faces criticism for restrictive licensing, underwhelming benchmarks in reasoning, and eroding community trust due to unfulfilled promises. DeepSeek, highlighted by its R1 release in January, focuses on reasoning-heavy, cost-effective models with modular designs and MIT licensing, enabling superior performance in complex tasks like math and coding at lower inference costs, while fostering rapid adoption through transparency and customization. This rivalry underscores open models' role in AI evolution: Llama promotes researcher access and power decentralization through permissive (albeit increasingly limited) openness, while DeepSeek accelerates original idea generation via efficient, verifiable reasoning and synthetic data distillation, potentially shifting competitive edges toward resource-efficient innovation. However, DeepSeek's advantages in specialized domains and lower costs may outpace Llama unless Meta addresses strategic inconsistencies, highlighting the need for sustained Western investments to maintain balance in global AI development.

Focus on Llama-Origin Models

Meta's Llama series, rooted in American open AI efforts, has transitioned from pre-2025 dominance to a more contested position with Llama 4's April release, featuring MoE architectures for enhanced efficiency and multimodality. Key variants include Scout (17B active/109B total parameters, 10M context), Maverick (17B active/400B total, 1M context), and the previewed Behemoth (288B active/2T total), trained on trillions of tokens for tasks like video generation and cross-modal reasoning. Performance shines in multilingual and coding benchmarks, with Maverick achieving high ELO scores (1417 on LMArena) and strong non-reasoning evaluations on ArtificialAnalysis, though it lags in creative writing and reasoning compared to rivals. Adoption has waned due to restrictive licensing (requiring "Llama-" branding and EU vision bans), unmaterialized updates (e.g.,

Llama 4.1/4.2), and high memory demands, shifting users to more accessible alternatives. Strategically, Llama adopts a platform approach with varied sizes for research and startups, but its cautious openness and focus on risk mitigation have diminished its edge in fostering original thoughts, as academics grapple with outdated models relative to closed frontiers. Strengths include reducing AI power concentration via open weights, while weaknesses involve political bias mitigation and scalability issues.

Focus on DeepSeek-Origin Models

DeepSeek, a Chinese frontrunner, burst onto the scene with R1 in January 2025, a reasoning language model (RLM) using a 4-stage RL-heavy process on a V3 base, emphasizing verifiable domains, modular customization, and MIT licensing for unrestricted building. V3/R1 variants boast 671B total parameters (37B active via MoE), excelling in reasoning, math, and code generation on benchmarks like HumanEval, often outperforming Llama in complex tasks at 10x lower pricing ($0.55/M input tokens). Adoption surges due to efficiency on mid-tier hardware, synthetic data generation, and community contributions, with R1 distillations boosting smaller models and sparking progress in open RLMs. As a "single-path, high-quality" substitute for API models, DeepSeek prioritizes market share through transparency and low costs, enabling original innovations in creativity and technical domains, though it requires large compute for full potential. Strengths lie in structured outputs and multilingual support, with weaknesses in broad general knowledge compared to proprietary benchmarks.

Comparative Table on Various Parameters

Parameter	Llama Models (e.g., Llama 4)	DeepSeek Models (e.g., R1/V3)
Major Releases in 2025	Llama 4 (April): Scout, Maverick, Behemoth previews; Fewer, MoE-focused.	DeepSeek R1 (January), V3 updates; Prolific, reasoning-centric.
Adoption (Downloads & Fine-Tunes)	Declining: Restrictive license erodes trust; Strong in research/startups.	Surging: MIT license drives community builds; High in technical users.
Performance	Strong in multilingual/coding (ELO 1417); Lags in reasoning/creativity.	Leading in reasoning/math (outperforms on HumanEval); Efficient inference.

Strategy & Variety	Platform-oriented: Varied sizes/modalities; Cautious openness.	Single-path: Modular, customizable; Focus on verifiable, cost-effective RL.
Role in Idea Generation & AI Evolution	Moderate: Enables access but fragmented; Risks academic irrelevance.	High: Distillation/synthetic data spur innovations; Accelerates open RLMs.
Competitive Edge Maintenance	Weakening: Needs transparent roadmaps; EU restrictions limit.	Strong: Low costs/soft power; Challenges Western dominance.

Benchmark Analytics between Llama and Mistral

Comparative Statement on Llama and Mistral AI Models

In 2025, from the Western hemisphere, Meta's Llama series and Mistral AI's models represent contrasting approaches in the open AI ecosystem, with Llama emphasizing large-scale, multimodal capabilities and broad platform adoption, while Mistral prioritizes efficiency, reasoning-focused designs, and cost-effective performance for enterprise and specialized tasks. Llama 4, released in April, advances multimodality and Mixture of Experts (MoE) architectures, achieving strong benchmarks in coding and multilingual tasks but facing criticism for restrictive licensing and inconsistent innovation. Mistral, a French startup, has maintained momentum through releases like Medium 3 (May), Magistral family (June), Medium 3.1 (August), and Magistral Small 1.2 (September), excelling in compact models that rival larger ones in reasoning, math, and now vision analysis, often at lower costs. This rivalry highlights open models' pivotal role in AI evolution: Llama fosters global accessibility and power decentralization through varied sizes and integrations, enabling original idea generation in research and startups, whereas Mistral drives efficiency-driven innovations, such as transparent multi-step reasoning and high-throughput applications, helping maintain competitive edges in resource-constrained environments. However, Mistral's "fading" trajectory noted in broader analyses stems from fewer headline releases compared to Chinese peers, while Llama's trust issues risk ceding ground unless addressed through more agile strategies.

Focus on Llama-Origin Models

As detailed in prior assessments, Meta's Llama series, including Llama 4 (April 2025) and MoE variants like Scout, Maverick, and Behemoth preview, uses massive datasets for multimodal generation. While strong in ELO and non-reasoning tasks, it underperforms in

creative writing and advanced reasoning. Adoption has declined due to licensing and unfulfilled updates, pushing users to alternatives. Llama's platform model supports researchers and startups, but cautious openness and bias mitigations hinder original thought.

Focus on Mistral-Origin Models

Mistral AI, a European (French) player, has evolved from early successes like Mistral 7B and Mixtral 8x7B to a 2025 lineup emphasizing efficient, reasoning-centric models. Key releases include Medium 3 (May, frontier-class for efficiency), Magistral family (June, first reasoning-focused with Small and Large variants for multi-step transparency), Medium 3.1 (August, multimodal enhancements), Small 3.2 24B (September, compact with improved tone), and Magistral Small 1.2 (September, adding image analysis). Benchmarks show Mistral models punching above their weight: Medium 3.1 rivals larger models in performance-per-parameter, excelling in math, code generation, and high-throughput tasks on evaluations like HumanEval and ArtificialAnalysis, often at lower inference costs ($0.20/M tokens vs. competitors). Adoption focuses on enterprise and open-source communities, with permissive licensing (e.g., Apache 2.0 for many) enabling fine-tuning, though some variants remain proprietary. Strategically, Mistral pursues "agility over scale," targeting niches like multilingual support (128+ languages) and vision-reasoning, fostering original innovations in efficient AI for developers and businesses. Weaknesses include smaller context windows (e.g., 128K tokens) and less multimodality depth compared to Llama, contributing to its perceived fade amid prolific Chinese releases.

Comparative Table on Various Parameters

Parameter	Llama Models (e.g., Llama 4)	Mistral Models (e.g., Medium 3.1, Magistral)
Major Releases in 2025	Llama 4 (April): Scout, Maverick, Behemoth previews; MoE-multimodal focus.	Medium 3 (May), Magistral (June), Medium 3.1 (Aug), Small 3.2/Magistral 1.2 (Sep); Reasoning/efficiency emphasis.
Adoption (Downloads & Fine-Tunes)	Declining: Strong in research/startups but eroded by restrictions; Billions historically.	Steady: Enterprise-focused with open-source appeal; Growing in efficiency niches.

Performance	Strong in multilingual/coding (ELO 1417); Lags in reasoning/creativity; High in multimodality.	Leading in efficiency/math/code; Rivals larger models per parameter; Emerging vision-reasoning.
Strategy & Variety	Platform-oriented: Varied sizes (8B-2T), multimodal; Cautious openness.	Agility-focused: Compact reasoning models, some proprietary; Multilingual efficiency.
Role in Idea Generation & AI Evolution	Moderate: Broad access spurs research; Risks fragmentation without updates.	High: Efficient designs enable niche innovations; Transparent reasoning advances multi-step AI.
Competitive Edge Maintenance	Weakening: Scale advantages offset by trust issues; Needs agile iterations.	Strong in cost/performance: Challenges scale with efficiency; Vulnerable to volume from rivals.

Conclusion

The year 2025 marks a definitive turning point in the landscape of open AI models, characterized by a pronounced and, for the moment, seemingly irreversible shift towards Chinese preeminence. This ascendancy is not merely anecdotal but demonstrably driven by a confluence of factors: an unprecedented torrent of new model releases, consistent outperformance in critical benchmarks, and adoption rates that have comprehensively outstripped those of American and other global initiatives.

Quantitative data unequivocally supports this narrative. By mid-year, Chinese models had collectively captured over 50% of cumulative Hugging Face downloads and fine-tune shares. This dominance is spearheaded by influential organizations such as Qwen, DeepSeek, and, most recently, Ant Group, whose new trillion-parameter LLM (Large Language Model) has particularly excelled in complex mathematical reasoning. These entities have collectively forged a robust, interconnected ecosystem that, in its accessibility and versatility, strikingly mirrors the pervasive reach and community-driven innovation characteristic of the Android operating system. This fosters widespread community

engagement, encourages iterative development, and accelerates the overall evolution of AI.

The impact of this surge extends across a diverse spectrum of use cases. It encompasses advanced multimodal vision-language models capable of interpreting and generating content across different data types, to sophisticated agentic coding agents that automate and optimize software development. Crucially, this rapid progress is fueled by the democratization of access to powerful AI tools. By making cutting-edge models widely available, Chinese developers are empowering a global community of innovators, enabling them to generate novel ideas and applications that significantly push the boundaries in areas such as advanced reasoning, sophisticated tool use, and operational efficiency.

In stark contrast, American models, despite notable contributions like Gemma 3, Llama 4, and the ongoing advancements within Ai2's OLMo project, have struggled with a different set of challenges. Inconsistent openness, a perceived erosion of trust, and a slower pace of development have resulted in a fragmented ecosystem. This fragmentation poses a substantial risk, potentially leading to academic and innovative irrelevance when compared to the rapidly advancing and often closed-source frontiers of AI development elsewhere.

Meanwhile, other regions, such as Europe, find themselves playing an increasingly marginal role. The waning influence of models like Mistral underscores a emerging bipolar dynamic in the global AI landscape. China's market-share-oriented strategy, which skillfully leverages community norms and open-source principles, appears to be a sustainable path to maintaining its leadership position.

Ultimately, this evolving scenario powerfully underscores the critical and multifaceted role that open models play in maintaining a competitive edge in the global AI race. They serve as potent mechanisms for diffusing knowledge, thereby preventing the concentration of power in a few proprietary systems. Furthermore, through their international adoption, open models become instruments of soft power, fostering global collaboration and technological alignment. However, without deliberate and strategic interventions—particularly from Western nations—there is a palpable risk of a widening gap. This could lead to a significant ceding of ground not only in long-term AI innovation but also in critical areas such as AI safety research and the broader societal empowerment that robust, open AI ecosystems can facilitate.

Way Forwards

To secure, maintain, and enhance a leading position within the dynamic open AI model ecosystem, a comprehensive and multi-faceted strategy is imperative. This approach must strategically prioritize significant investment, robust collaboration, and cohesive policy alignment across various fronts.

1. Scale Domestic Investments in Foundational Open Models:

Governments and leading institutions, particularly within the United States and Europe, must substantially increase their financial commitments beyond existing initiatives. While programs such as Ai2's commendable $100 million NSF-Nvidia grant for the OLMo project are valuable starting points, they represent only a fraction of what is truly needed. The ambition should be to allocate at least $500 million annually to accelerate the development of truly open models. This means not only releasing the model weights but also making available the complete training data, the underlying code, and comprehensive documentation, ensuring full transparency and reproducibility. The goal is to develop open models that can genuinely rival frontier capabilities in terms of scale, sophistication, and multimodal understanding. Such an amplified investment is critical to bridge the current developmental gap, enabling projects like OLMo 2 to evolve into advanced Mixture-of-Experts (MoE) architectures that can compete directly with the powerful, often state-backed, models emerging from countries like China. This influx of capital would also foster innovation in novel architectural designs, efficient training methodologies, and ethical AI development, ensuring that Western open models are not merely catch-up efforts but global leaders.

2. Foster Collaborative Ecosystems through Public-Private Consortia:

Establishing robust public-private consortia is essential to harness collective intelligence and resources. Expanding successful models like the ATOM Project, these consortia should bring together a diverse array of stakeholders: leading academic institutions (e.g., Allen Institute for AI, top-tier universities), cutting-edge technology firms (e.g., Nvidia, Google, Microsoft), and vibrant open-source communities. The focus of these collaborations should be on defining and implementing "truly open" standards for AI development, moving beyond mere open-sourcing to embrace full transparency in training methodologies, dataset curation, and evaluation protocols. Critically, these consortia should facilitate the open sharing of training know-how, best practices, and high-quality, ethically sourced datasets. This will significantly lower the barrier to entry for researchers, encouraging wider adoption and stimulating the generation of novel ideas and innovative applications.

Furthermore, embedded within these collaborative efforts must be a strong emphasis on addressing AI safety and ethical concerns through transparent evaluation frameworks and shared research into alignment and responsible deployment.

3. Enhance Community Engagement and Rebuild Trust in Western Models:

To regain and sustain community trust and enthusiasm, Western open model initiatives ought to draw inspiration from successful strategies employed by models like Qwen. This involves prioritizing user feedback loops, ensuring frequent and incremental model releases, and striving for broad multimodal coverage (e.g., text, image, audio, video). These actions are vital to rebuild trust and generate excitement around Western-developed open models. Practical initiatives could include: organizing global hackathons to encourage creative applications and fine-tuning; offering significant incentives for community members to contribute to fine-tuning, bug fixes, and feature development; and developing user-friendly platforms that facilitate the creation and sharing of derivative models. The aim is to re-ignite the "tuning buzz" and collaborative spirit that characterized earlier eras of open-source software development, fostering a dynamic and self-sustaining ecosystem of innovation.

4. Policy Advocacy and International Influence for Open AI:

Effective policy and strategic international influence are paramount. Advocacy efforts should focus on promoting openness in AI development without inadvertently restricting the flow of critical information and research. This means carefully balancing measures like export controls with strong incentives for allied nations to collaboratively develop and deploy open models. Leveraging soft power will be crucial, particularly by actively exporting Western open models to emerging markets. This strategy serves a dual purpose: countering the growing influence of closed or less transparent models originating from other geopolitical spheres and, equally important, investing in local training infrastructure within these emerging markets. Such investments are vital to narrow the "local-closed model gap," empowering local developers and researchers and ensuring that the benefits of open AI are globally distributed and democratically accessible.

5. Continuous Monitoring, Adaptation, and Risk Mitigation:

The landscape of AI is rapidly evolving, necessitating a commitment to continuous monitoring and adaptive strategies. Regular and rigorous assessments of geopolitical implications are essential, including a close examination of potential state alignments and

biases embedded within models from various regions, particularly those from Chinese developers. Parallel to this, sustained investment in AI alignment research is critical to ensure that open models are developed and deployed in a manner that genuinely advances societal good and minimizes unintended negative consequences. Tracking key performance metrics and ecosystem health indicators via established platforms like ArtificialAnalysis and Hugging Face will provide invaluable data. This data should then be iteratively used to refine strategies, identify emerging challenges, and capitalize on new opportunities, ensuring that the approach to open AI remains agile, responsive, and forward-looking.

Any player wishing to augment its its open model ecosystem, ensuring sustained innovation and a balanced global AI landscape.

References

1. Allen Institute for AI (2025) 'NSF-NVIDIA Grant for OLMo', AllenAI.org. Available at: https://allenai.org/blog/nsf-nvidia (Accessed: 13 October 2025).
2. Amazon (2025) Amazon Bedrock Integration with Llama. Available at: https://aws.amazon.com/bedrock/ (Accessed: 13 October 2025).
3. Anthropic (2025) Claude 4.5 Model Overview. Available at: https://anthropic.com/claude (Accessed: 13 October 2025).
4. ArtificialAnalysis (2025) LLM Benchmark Scores 2025. Available at: https://artificialanalysis.ai/ (Accessed: 13 October 2025).
5. Autonomous (2024) 'Llama vs. Mistral: Which Performs Better and Why?', Autonomous.ai, 26 September. Available at: https://www.autonomous.ai/ourblog/llama-vs-mistral-which-performs-better (Accessed: 13 October 2025).
6. Civo (2025) 'DeepSeek vs Llama vs GPT-4 | Open-Source AI Models Compared', Civo.com, 7 May. Available at: https://www.civo.com/blog/deepseek-vs-llama-vs-gpt4-ai-models (Accessed: 13 October 2025).
7. Codersera (2025) 'Llama 4 vs Mistral 7B: A Comprehensive Comparison of AI Models', Codersera.com. Available at: https://codersera.com/blog/llama-4-vs-mistral-7b-a-comprehensive-comparison-of-ai-models (Accessed: 13 October 2025).
8. Collabnix (2025) 'AI Models Comparison 2025: Claude, Grok, GPT & More', Collabnix.com, 1 July. Available at:

https://collabnix.com/comparing-top-ai-models-in-2025-claude-grok-gpt-llama-gemini-and-deepseek-the-ultimate-guide/ (Accessed: 13 October 2025).

9. DeepSeek AI (2025) 'Llama 4 vs DeepSeek AI: Complete Model Comparison', DeepSeek.ai, 10 April. Available at: https://deepseek.ai/blog/llama-4-vs-deepseek (Accessed: 13 October 2025).
10. Dev.to (2025a) 'Qwen3-Max 2025 Complete Release Analysis: In-Depth Review of Alibaba's Most Powerful AI Model',
11. Dev.to, 24 September. Available at: https://dev.to/czmilo/qwen3-max-2025-complete-release-analysis-in-depth-review-of-alibabas-most-powerful-ai-model-3j7l (Accessed: 13 October 2025).
12. Dev.to (2025b) 'The Great AI Battle of 2025: OpenAI vs DeepSeek vs Qwen - Who's Actually Winning?',
13. Dev.to, 31 August. Available at: https://dev.to/shiva_shanker_k/the-great-ai-battle-of-2025-openai-vs-deepseek-vs-qwen-whos-actually-winning-55j3 (Accessed: 13 October 2025).
14. DigitalDefynd (2025) '15 Pros & Cons of Qwen AI [2025]', DigitalDefynd.com. Available at: https://digitaldefynd.com/IQ/qwen-ai-pros-cons/ (Accessed: 13 October 2025).
15. DocsBot AI (2025a) 'Llama 4 Scout vs Mistral Large 2', Docsbot.ai. Available at: https://docsbot.ai/models/compare/llama-4-scout/mistral-large-2 (Accessed: 13 October 2025).
16. DocsBot AI (2025b) 'Llama 4 Maverick vs Mistral Large 2', Docsbot.ai. Available at: https://docsbot.ai/models/compare/llama-4-maverick/mistral-large-2 (Accessed: 13 October 2025).
17. Eden AI (2025) 'Llama 3.3 vs DeepSeek-R1', Edenai.co. Available at: https://www.edenai.co/post/llama-3-3-vs-deepseek-r1 (Accessed: 13 October 2025).
18. Eesel AI (2025) 'What I Learned After Testing Mistral AI's New Models', Eesel.ai, 12 September. Available at: https://www.eesel.ai/blog/mistral-ai-new-models (Accessed: 13 October 2025).
19. Elephas (2025) 'DeepSeek vs Llama (2025 Comparison): Which Local AI Model is Best?', Elephas.app, 28 February. Available at: https://elephas.app/blog/deepseek-vs-llama-2025-comparison-which-local-ai-model-is-best-cm7kddany00ekip0lidqqoeq3 (Accessed: 13 October 2025).

20. Epoch AI (2025) 'Consumer GPU Model Gap', Epoch.ai. Available at: https://epoch.ai/data-insights/consumer-gpu-model-gap (Accessed: 13 October 2025).
21. Ethics in AI Forum (2025) Bias in Open Models: A 2025 Review. Available at: https://ethicsinaiforum.org/reports/2025-bias (Accessed: 13 October 2025).
22. GitHub (2025) Llama Licensing Discussions. Available at: https://github.com/meta-llama/llama/discussions (Accessed: 13 October 2025).
23. Google (2025) Gemini 2.5 Pro Documentation. Available at: https://deepmind.google/technologies/gemini/ (Accessed: 13 October 2025).
24. Hugging Face (2025) Llama Model Hub. Available at: https://huggingface.co/meta-llama (Accessed: 13 October 2025).
25. InfoQ (2025) 'Mistral AI Releases Magistral, Its First Reasoning-Focused Model Family', Infoq.com, 16 June. Available at: https://www.infoq.com/news/2025/06/mistral-ai-magistral/ (Accessed: 13 October 2025).
26. Interconnects.ai (2025) Llama 4 Review. Available at: https://www.interconnects.ai/p/llama-4 (Accessed: 13 October 2025).
27. Lambert, N. (2025) 'China's Top 19 Open Model Labs', Interconnects.ai, 5 October. Available at: https://www.interconnects.ai/p/chinas-top-19-open-model-labs (Accessed: 13 October 2025).
28. Lambert, N. (2025) 'DeepSeek R1: Recipe for o1', Interconnects.ai, 5 October. Available at: https://www.interconnects.ai/p/deepseek-r1-recipe-for-o1 (Accessed: 13 October 2025).
29. Lambert, N. (2025) 'GPT-OSS: OpenAI Validates the Open', Interconnects.ai, 5 October. Available at: https://www.interconnects.ai/p/gpt-oss-openai-validates-the-open (Accessed: 13 October 2025).
30. Lambert, N. (2025) 'Gemma 3, OLMo 2 32B, and the Growing', Interconnects.ai, 5 October. Available at: https://www.interconnects.ai/p/gemma-3-olmo-2-32b-and-the-growing (Accessed: 13 October 2025).
31. Lambert, N. (2025) 'Llama 4', Interconnects.ai, 5 October. Available at: https://www.interconnects.ai/p/llama-4 (Accessed: 13 October 2025).

32. Lambert, N. (2025) 'Qwen 3: The New Open Standard', Interconnects.ai, 5 October. Available at: https://www.interconnects.ai/p/qwen-3-the-new-open-standard (Accessed: 13 October 2025).
33. Lambert, N. (2025) Open Models in 2025. Interconnects // Ai2, The Curve, 5 October 2025. Available at: Open_Models_in_2025_1760347230.pdf (Accessed: 13 October 2025).
34. Lambert, N. (2025) Open Models in 2025. Interconnects // Ai2, The Curve, 5 October.
35. Lambert, N. (2025a) 'DeepSeek R1's Recipe to Replicate o1 and the Future of Reasoning LMs', Interconnects.ai, 20 January. Available at: https://www.interconnects.ai/p/deepseek-r1-recipe-for-o1 (Accessed: 13 October 2025).
36. Lambert, N. (2025a) 'Qwen 3: The New Open Standard', Interconnects.ai, 28 April. Available at: https://www.interconnects.ai/p/qwen-3-the-new-open-standard (Accessed: 13 October 2025).
37. Lambert, N. (2025a) Open Models in 2025. Interconnects // Ai2, The Curve, 5 October.
38. Lambert, N. (2025b) 'Llama 4: Did Meta Just Push the Panic Button?', Interconnects.ai, 5 April. Available at: https://www.interconnects.ai/p/llama-4 (Accessed: 13 October 2025).
39. Lambert, N. (2025b) Multimodal Advances in Llama 4. Available at: https://www.interconnects.ai/p/llama-4-multimodal (Accessed: 13 October 2025).
40. Lambert, N. (2025c) Benchmarking Llama 4 vs Competitors. Available at: https://www.interconnects.ai/p/llama-4-benchmarks (Accessed: 13 October 2025).
41. Lambert, N. (2025d) Architectural Shortcomings in Llama. Available at: https://www.interconnects.ai/p/llama-architecture (Accessed: 13 October 2025).
42. Lambert, N. (2025e) Declining Trust in Meta's Open Models. Available at: https://www.interconnects.ai/p/meta-trust-issues (Accessed: 13 October 2025).
43. Lambert, N. (2025f) Licensing Changes in Llama Series. Available at: https://www.interconnects.ai/p/llama-licensing (Accessed: 13 October 2025).
44. Lambert, N. (2025g) Chinese Models Overtaking Llama. Available at: https://www.interconnects.ai/p/chinese-dominance (Accessed: 13 October 2025).
45. Leanware (2025) 'ChatGPT vs Mistral: Full AI Comparison Guide for 2025', Leanware.co, 1 October. Available at: https://www.leanware.co/insights/chatgpt-vs-mistral (Accessed: 13 October 2025).

46. Machine Translation (2025) 'Mistral vs LLaMA: A 2025 Comparison of Performance, Cost, and Capabilities', Machinetranslation.com, 23 July. Available at: https://www.machinetranslation.com/blog/mistral-vs-llama (Accessed: 13 October 2025).
47. Medium (2025) 'A Detailed Comparison of all LLMs in 2025 - GPT vs Gemini vs DeepSeek vs Llama vs Claude and More', Medium.com, 21 April. Available at: https://medium.com/@aryadav.2810/a-detailed-comparison-of-all-llms-in-2025-gpt-vs-gemini-vs-deepseek-vs-llama-vs-claude-and-more-f54b576c77d4 (Accessed: 13 October 2025).
48. Medium (2025) 'The AI Language Model Landscape in 2025: Qwen, DeepSeek, and Hunyuan Lead the Pack', Medium.com, 10 June. Available at: https://medium.com/@cognidownunder/the-ai-language-model-landscape-in-2025-qwen-deepseek-and-hunyuan-lead-the-pack-662d65db066a (Accessed: 13 October 2025).
49. Meta (2024) 'Open Source AI is the Path Forward', About.fb.com, July. Available at: https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/ (Accessed: 13 October 2025).
50. Meta (2024) Open Source AI is the Path Forward. Available at: https://about.fb.com/news/2024/07/open-source-ai-is-the-path-forward/ (Accessed: 13 October 2025).
51. Meta (2025) 'Superintelligence', Meta.com. Available at: https://www.meta.com/superintelligence/ (Accessed: 13 October 2025).
52. Meta (2025a) Llama 4 Release Notes. Available at: https://ai.meta.com/llama/ (Accessed: 13 October 2025).
53. Meta (2025b) Meta Movie Gen Integration. Available at: https://ai.meta.com/blog/movie-gen/ (Accessed: 13 October 2025).
54. Meta (2025c) Llama for Startups Program. Available at: https://ai.meta.com/startups/ (Accessed: 13 October 2025).
55. Meta (2025d) Superintelligence Labs Update. Available at: https://www.meta.com/superintelligence/ (Accessed: 13 October 2025).
56. Meta AI (2025) 'The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal Intelligence', Ai.meta.com, 5 April. Available at: https://ai.meta.com/blog/llama-4-multimodal-intelligence/ (Accessed: 13 October 2025).

57. Mistral AI (2025a) 'Models Overview', Docs.mistral.ai. Available at: https://docs.mistral.ai/getting-started/models/models_overview/ (Accessed: 13 October 2025).
58. Mistral AI (2025b) 'Models Benchmarks', Docs.mistral.ai. Available at: https://docs.mistral.ai/getting-started/models/benchmark/ (Accessed: 13 October 2025).
59. Mistral AI (2025c) 'Medium is the New Large', Mistral.ai, 7 May. Available at: https://mistral.ai/news/mistral-medium-3 (Accessed: 13 October 2025).
60. NBC News (2025) 'Why DeepSeek is Different, in Three Charts', NBCNews.com, 28 January. Available at: https://www.nbcnews.com/data-graphics/deepseek-ai-comparison-openai-chatgpt-google-gemini-meta-llama-rcna189568 (Accessed: 13 October 2025).
61. Nexgen Compute (2025) 'Mistral AI in 2025: A Deep Dive into Mistral Small 3.2, Magistral, and the Battle with OpenAI & Google', Nexgen-compute.com. Available at: https://www.nexgen-compute.com/blog/mistral-ai-in-2025-a-deep-dive-into-mistral-small-3-2-magistral-and-the-battle-with-openai-google (Accessed: 13 October 2025).
62. Openxcell (2025) 'Mistral vs Llama 3: Key Differences & Best Use Cases', Openxcell.com, 10 April. Available at: https://www.openxcell.com/blog/mistral-vs-llama-3/ (Accessed: 13 October 2025).
63. PageOn AI (2025) 'Comparing GPT, Llama, and Mistral Models in 2025', Blogs.pageon.ai. Available at: https://blogs.pageon.ai/foundation-model-selection-comparing-gpt-llama-mistral-models-2025 (Accessed: 13 October 2025).
64. Prompt Hackers (2025) 'Compare Llama 4 Behemoth vs Mistral Large 2', Prompthackers.co. Available at: https://www.prompthackers.co/compare/llama-4-behemoth/mistral-large-2 (Accessed: 13 October 2025).
65. Pubby (2025) 'Choosing the Right LLM: Llama vs Mistral vs DeepSeek', Ai.gopubby.com, 24 June. Available at: https://ai.gopubby.com/choosing-the-right-llm-llama-vs-mistral-vs-deepseek-6577136a895b (Accessed: 13 October 2025).
66. QwenLM (2025a) 'Qwen2.5-Max: Exploring the Intelligence of Large-scale MoE Model', QwenLM.github.io, 28 January. Available at: https://qwenlm.github.io/blog/qwen2.5-max/ (Accessed: 13 October 2025).

67. QwenLM (2025b) 'Did Qwen Just Revolutionize AI with These New Model Releases?', Blog.devgenius.io, 23 September. Available at: https://blog.devgenius.io/did-qwen-just-revolutionize-ai-with-these-new-model-releases-a87c7883a49f (Accessed: 13 October 2025).
68. RedBlink (2025) 'Llama 4 vs DeepSeek V3: Comprehensive AI Model Comparison', RedBlink.com, 7 April. Available at: https://redblink.com/llama-4-vs-deepseek-v3/ (Accessed: 13 October 2025).
69. Reddit (2025a) 'How Better is DeepSeek R1 Compared to Llama3? Both are Open Source Right?', Reddit.com, 26 January. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1iadr5g/how_better_is_deepseek_r1_compared_to_llama3_both/ (Accessed: 13 October 2025).
70. Reddit (2025a) 'Mistral vs LLaMA 3 (2025) — Which Open-Source Model Should You Choose?', Reddit.com, 22 August. Available at: https://www.reddit.com/r/MistralAI/comments/1mwwc2g/mistral_vs_llama_3_2025_which_opensource_model/ (Accessed: 13 October 2025).
71. Reddit (2025a) 'We Tested Qwen3-Coder, GPT-5 and Other 30+ Models on New SWE-Bench Like Tasks from July 2025', Reddit.com/r/LocalLLaMA, 12 August. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1moakv3/we_tested_qwen3coder_gpt5_and_other_30_models_on/ (Accessed: 13 October 2025).
72. Reddit (2025a) r/MachineLearning Thread on Llama 4 Delays. Available at: https://www.reddit.com/r/MachineLearning/comments/llama4delays (Accessed: 13 October 2025).
73. Reddit (2025b) 'GPT-4o Mini vs Llama 3.1 405B vs Mistral Large 2 vs Claude Sonnet 3.5', Reddit.com, 26 July. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1ecs2dv/gpt4o_mini_vs_llama_31_405_b_vs_mistral_large_2/ (Accessed: 13 October 2025).
74. Reddit (2025b) 'Why is Qwen 2.5 the Most Used Models in Research?', Reddit.com/r/LocalLLaMA, 29 May. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1kyrhr7/why_is_qwen_25_the_most_used_models_in_research/ (Accessed: 13 October 2025).
75. Reddit (2025b) r/LocalLLaMA Critique of Llama 4. Available at: https://www.reddit.com/r/LocalLLaMA/comments/llama4critique (Accessed: 13 October 2025).
76. Reddit (2025c) 'The Latest Qwen Models are a Big Deal in More Ways Than You Think', Reddit.com/r/singularity, 12 November 2024. Available at:

https://www.reddit.com/r/singularity/comments/1gq525n/the_latest_qwen_models_are_a_big_deal_in_more/ (Accessed: 13 October 2025).

77. Reddit (2025c) User Sentiment on Llama Stagnation. Available at: https://www.reddit.com/r/AI/comments/llamastagnation (Accessed: 13 October 2025).
78. Reddit (2025d) 'Qwen 3: Unimpressive Coding Performance So Far', Reddit.com/r/LocalLLaMA, 28 April. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1ka8ban/qwen_3_unimpressive_coding_performance_so_far/ (Accessed: 13 October 2025).
79. Reddit (2025e) 'A Summary of Qwen Models!', Reddit.com/r/LocalLLaMA, 19 January. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1i4w47k/a_summary_of_qwen_models/ (Accessed: 13 October 2025).
80. Reddit (2025f) 'Qwen 2.5 vs Qwen 3 vs Gemma 3: Real World Base Model Comparison?', Reddit.com/r/LocalLLaMA, 15 May. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1kn6mic/qwen_25_vs_qwen_3_vs_gemma_3_real_world_base/ (Accessed: 13 October 2025).
81. Reddit (2025g) 'Qwen 2.5 = China = Bad', Reddit.com/r/LocalLLaMA, 3 October 2024. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1fv37i1/qwen_25_china_bad/ (Accessed: 13 October 2025).
82. Reddit (2025h) 'Llama 3.3 vs Qwen 2.5', Reddit.com/r/LocalLLaMA, 7 December 2024. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1h91e4h/llama_33_vs_qwen_25/ (Accessed: 13 October 2025).
83. Reddit (2025i) 'Which is Best Among These 3 Qwen Models', Reddit.com/r/LocalLLaMA, 28 April. Available at: https://www.reddit.com/r/LocalLLaMA/comments/1kafrae/which_is_best_among_these_3_qwen_models/ (Accessed: 13 October 2025).
84. Rival Tips (2025) 'Mistral Large 2 vs Llama 4 Behemoth Comparison', Rival.tips. Available at: https://www.rival.tips/compare/mistral-large-2/llama-4-behemoth (Accessed: 13 October 2025).
85. Shakudo (2025) 'Top 9 Large Language Models as of October 2025', Shakudo.io. Available at: https://www.shakudo.io/blog/top-9-large-language-models (Accessed: 13 October 2025).

86. SiliconFlow (2025) 'Ultimate Guide - The Best Qwen Models in 2025', SiliconFlow.com. Available at: https://www.siliconflow.com/articles/en/the-best-qwen-models-in-2025 (Accessed: 13 October 2025).
87. TechCrunch (2025) 'Mistral Claims Its Newest AI Model Delivers Leading Performance for the Price', Techcrunch.com, 7 May. Available at: https://techcrunch.com/2025/05/07/mistral-claims-its-newest-ai-model-delivers-leading-performance-for-the-price/ (Accessed: 13 October 2025).
88. TechCrunch (2025) 'OpenAI Calls DeepSeek State-Controlled, Calls for Bans on PRC-Produced Models', TechCrunch.com, 13 March. Available at: https://techcrunch.com/2025/03/13/openai-calls-deepseek-state-controlled-calls-for-bans-on-prc-produced-models/ (Accessed: 13 October 2025).
89. TechPoint Africa (2025) 'Qwen 3 AI Review: I Tested the Latest Model - Here's What I Found', TechPoint.africa, 20 August. Available at: https://techpoint.africa/guide/qwen-3-ai-review/ (Accessed: 13 October 2025).
90. The ATOM Project (2025) 'Adoption Metrics for Open Models', Atomproject.ai. Available at: https://atomproject.ai/ (Accessed: 13 October 2025).
91. The ATOM Project (2025) ATOM Project. Available at: https://atomproject.ai/ (Accessed: 13 October 2025).
92. The ATOM Project (2025) Adoption Metrics for Open Models. Available at: https://atomproject.ai/ (Accessed: 13 October 2025).
93. The Firewall Blog (2025) 'Privacy Pitfalls in AI: A Closer Look at DeepSeek and Qwen', Thefirewall-blog.com, 7 March. Available at: https://www.thefirewall-blog.com/2025/03/privacy-pitfalls-in-ai-a-closer-look-at-deepseek-and-qwen/ (Accessed: 13 October 2025).
94. VentureBeat (2025) 'Meta's Answer to DeepSeek is Here: Llama 4 Launches with Long Context Scout and Maverick Models, and 2T Parameter Behemoth on the Way!', VentureBeat.com, 5 April. Available at: https://venturebeat.com/ai/metas-answer-to-deepseek-is-here-llama-4-launches-with-long-context-scout-and-maverick-models-and-2t-parameter-behemoth-on-the-way (Accessed: 13 October 2025).
95. VentureBeat (2025a) 'Meta's Answer to DeepSeek is Here: Llama 4 Launches with Long Context Scout and Maverick Models, and 2T Parameter Behemoth on the Way!', Venturebeat.com, 5 April. Available at: https://venturebeat.com/ai/metas-answer-to-deepseek-is-here-llama-4-launches-wit

h-long-context-scout-and-maverick-models-and-2t-parameter-behemoth-on-the-way (Accessed: 13 October 2025).

96. VentureBeat (2025b) 'Mistral's Updated Magistral Small 1.2 Reasoning Model Can Analyze Images and More', Venturebeat.com, 18 September. Available at: https://venturebeat.com/ai/mistrals-updated-magistral-small-1-2-reasoning-model-can-analyze-images-and (Accessed: 13 October 2025).
97. X Community (2025a) Tweets on Llama 4 Performance. Available at: https://x.com/search?q=llama4performance (Accessed: 13 October 2025).
98. X Community (2025b) Developer Feedback on Llama. Available at: https://x.com/search?q=llamadeveloperfeedback (Accessed: 13 October 2025).
    X Community (2025c) Discussions on Llama Innovation. Available at: https://x.com/search?q=llamainnovation (Accessed: 13 October 2025).