Article

A Review of Large Language Models (LLMs) Development: A Cross-Country Comparison of the US, China, Europe, UK, India, Japan, South Korea, and Canada

Wan Chong Choi [*] , Chi In Chang , Iek Chong Choi , Lai Chu Lam

Posted Date: 25 April 2025

doi: 10.20944/preprints202504.2136.v1

Keywords: country landscape; national AI strategies; large language models; LLM; artificial intelligence generated content; AIGC; artificial intelligence; chatbot; natural language processing; ChatGPT; Claude; Gemini; LLaMA; Deepseek; ERNIE; BLOOM; Luminous

Article

A Review of Large Language Models (LLMs) Development: A Cross-Country Comparison of the US, China, Europe, UK, India, Japan, South Korea, and Canada

Wan Chong Choi [1,2,*], Chi In Chang [2], Iek Chong Choi [3] and Lai Chu Lam [4]

[1] Department of Computer Science, Illinois Institute of Technology, USA
[2] Department of Psychology, Golden Gate University, USA
[3] School of Education, City University of Macau, Macao
[4] School of Informatics, Computing and Cyber Systems, Northern Arizona University, USA
* Correspondence: wchoi8@hawk.iit.edu

Abstract: This paper provided a comprehensive comparative analysis of national strategies in developing large language models (LLMs), drawing on literature from 2020 to April 2025. The study examined how different countries approached the design, training, and deployment of LLMs, revealing distinct trajectories shaped by computational infrastructure, policy orientation, and linguistic objectives. The United States maintained a leadership position through models such as ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), and LLaMA (Meta), supported by robust private-sector innovation and integrated cloud ecosystems. China, despite hardware constraints, advanced rapidly through efficient and open models including ERNIE (Baidu), Tongyi Qianwen (Alibaba), PanGu (Huawei), and DeepSeek, emphasizing cost-effective scaling and bilingual capabilities. The European Union pursued openness and multilingual inclusivity with models such as BLOOM (France), Luminous (Germany), and OpenGPT-X, aligned with emerging AI regulatory standards. Other regions demonstrated targeted advancements: India introduced BharatGPT and Airavata to support its linguistic diversity; the UAE developed Falcon and Jais to lead Arabic AI; South Korea produced HyperCLOVA; and Japan contributed models like Nekomata and EvoLLM-JP. These developments illustrated that LLMs have become strategic assets reflecting broader national priorities, from technological sovereignty to regulatory alignment. The study concluded that while model architecture remains central, the evolution of LLMs is increasingly determined by geopolitical, infrastructural, and socio-linguistic factors that shape their integration into national digital ecosystems.

Keywords: country landscape; national AI strategies; large language models; LLM; artificial intelligence generated content; AIGC; artificial intelligence; chatbot; natural language processing; ChatGPT; Claude; Gemini; LLaMA; Deepseek; ERNIE; BLOOM; Luminous

1. Introduction

1.1. Background

Large Language Models (LLMs) have rapidly advanced since the breakthrough of OpenAI's GPT-3 in 2020, a 175-billion-parameter Transformer that demonstrated unprecedented few-shot learning capabilities. This milestone sparked a global race in both academia and industry to develop ever more capable LLMs.

LLMs are also increasingly applied across a wide range of fields such as healthcare [1], education [2–6], business [7], law [8], scientific research [9], customer service, software development, journalism, and public administration, where their ability to process and generate language is being adapted to domain-specific needs.

Nations around the world now view LLMs as strategic assets for economic growth, scientific leadership, and even national security. As a result, a diverse landscape of LLM development has emerged, with different countries and regions pursuing their own models, often tailored to local languages or specific domains. Each country's approach reflects its unique mix of research institutions, industrial players, government policies, and available computing infrastructure.

For example, by mid-2023, China alone had over 130 large-scale models, leading officials to declare "a war of a hundred models" [10] in AI. Meanwhile, open collaborations like the BigScience project in the EU produced BLOOM [11], a 176-billion-parameter multilingual model, illustrating a commitment to open science and shared resources. The U.S. continues to dominate in cutting-edge model development (e.g. OpenAI's ChatGPT [12]), but other regions are quickly closing the gap by leveraging national strengths – be it Korea's focus on Korean-language models, Japan's use of its world-class supercomputers, or the UAE's investment in open-source models.

This review provides a comprehensive country-by-country overview of LLM development up to April 2025, examining how different nations contribute to and shape the LLM landscape.

1.2. Research Questions

This review addressed the following key question to understand the international landscape of LLM development:

RQ: What were the key characteristics of large language models developed in different countries or regions?

By examining the questions, we aimed to map out the country landscape of LLM development and drew insights into how geopolitics, resources, and research cultures drove the evolution of large language models.

2. Methodology

This review is based on a structured survey of publicly available sources documenting the development of LLMs worldwide. Our analysis covers 2020 to April 2025, focusing on major model releases, institutional roles, and national-level strategies.

We conducted a structured review of scientific publications, technical reports, and preprints that document LLMs' design, training, and deployment. To complement this, we also systematically analyzed non-academic sources such as press releases, corporate blogs, and government policy documents. These sources were selected based on relevance and credibility, prioritizing direct statements from model developers and official institutions.

For each country or region, we identified representative models and associated organizations (academic, industrial, or governmental), and contextualized their development within local technological, policy, and linguistic environments. In comparing countries, we examined how differences in infrastructure, institutional ecosystems, and language priorities influence the direction of LLM development.

The synthesis integrates findings from these sources to address the two central research questions. We do not aim for exhaustive coverage of all LLMs, but rather for a representative analysis that captures key national and regional patterns in the evolving global LLM landscape.

3. United States

The United States has maintained a leading position in the development of large language models (LLMs) through a combination of early technical leadership, access to large-scale computational infrastructure, and a highly active ecosystem of private companies, academic institutions, and cloud providers. U.S.-based organizations have defined many central advances in scaling laws, alignment strategies, training optimization techniques, and deployment modalities. The emergence of general-purpose models such as ChatGPT catalyzed interest in LLMs across scientific and commercial domains, with U.S. models often setting performance benchmarks. The dominance of the private sector—driven by firms such as OpenAI, Google, Meta, and Microsoft—has resulted in fast iteration cycles and extensive commercialization, including the integration of LLMs into APIs and enterprise tools. Cloud infrastructure from providers like Azure, Google Cloud, and AWS has further enabled experimentation at an unprecedented scale. We highlighted the findings as Table 1.

Table 1. Comparison of Major U.S.-Developed LLMs and Their Characteristics**.**

Company	**LLM Series**	**Representative Models**	**Open-Source or Commercial Use**	**Characteristics**
OpenAI	ChatGPT	GPT-3, GPT-3.5, GPT-4, GPT-o1, GPT-o3	Commercial Use	One of the earliest and leading LLM developers; GPT-3 introduced scaling laws; GPT-3.5 added RLHF; GPT-4 is multimodal; from GPT-o1 onward, strong image recognition and generation capabilities.
Anthropic	Claude	Claude 1, Claude 2, Claude 3	Commercial Use	Focused on safety and alignment via Constitutional AI; partial transparency; delivered via API.
Google DeepMind	Gemini	Palm, Gemini 1, Gemini 1.5	Commercial Use	Built on Pathways architecture; optimized for reasoning and multimodal tasks; deployed via Bard and Vertex AI.
Meta	LLaMA	LLaMA 1, LLaMA 2	Open-Source	LLaMA 2 released under a permissive license; emphasizes transparency; used in research and industry.
Microsoft	Copilot	GPT-based (via OpenAI)	Commercial Use	Integrates OpenAI models into Office and enterprise tools; also acts as cloud and platform provider.
xAI	Grok	Grok 1, Grok 1.5	Commercial Use	Focuses on real-time information retrieval; integrated with X (formerly Twitter); emphasizes responsiveness and open discourse.
Amazon AWS	Titan / Nova	Titan Text G1 – Express, Titan Embeddings, Nova Pro, Nova Sonic	Commercial Use	Delivered through Bedrock; designed for enterprise needs; supports text generation, embeddings, image/video generation, and speech-based tasks; emphasizes modularity, privacy, and flexibility.
Nvidia	Nemotron Ultra	Nemotron-4 340B, Nemotron-3 43B	Open-Access (with constraints)	Reference models for GPU optimization; supports alignment and fine-tuning; delivered via NeMo framework.
Perplexity AI	Perplexity	Perplexity 2, Perplexity 3	Commercial Use	Retrieval-augmented generation (RAG); supports live citation and source tracking; optimized for factual consistency.

3.1. ChatGPT (OpenAI)

OpenAI's GPT series [12] has established the modern LLM paradigm. The sequence began with GPT-1 and GPT-2, scaling further to GPT-3 in 2020, a 175-billion parameter model demonstrating strong few-shot performance without task-specific fine-tuning. GPT-3 introduced a turning point by validating that performance scales with parameter count and dataset size influenced subsequent model designs globally.

In 2022, OpenAI fine-tuned GPT-3.5 using reinforcement learning from human feedback (RLHF) to create InstructGPT [13], and subsequently released ChatGPT, which brought instruction-following capabilities into public use [14]. These models were deployed through OpenAI's web-based interface and API access, enabling widespread integration into downstream applications.

In March 2023, OpenAI introduced GPT-4, a multimodal model capable of processing both text and images, though core architectural details such as parameter count remain undisclosed [14]. GPT-4 achieved competitive or "human-level" scores on standardized academic and professional benchmarks, according to OpenAI's internal evaluations.

Notably, OpenAI shifted from an open research model—originally exemplified by the publication of GPT-2 and GPT-3 papers—to a closed deployment paradigm, where access is mediated via API and integrated into Microsoft's Azure platform. This marked a strategic shift toward commercial control and enterprise licensing.

3.2. Claude (Anthropic)

Anthropic [15] was founded in 2021 by former OpenAI researchers focusing on AI alignment and safety, particularly in response to concerns about model controllability and robustness. The Claude series [16], beginning with Claude 1 and progressing to Claude 2 and Claude 3, has aimed to incorporate safety constraints more deeply into model training. A defining feature of Anthropic's approach is Constitutional AI, which substitutes traditional reinforcement learning methods with rule-based feedback designed to make models behave consistently with a set of ethical principles. This framework enables a degree of self-supervision that may reduce dependence on human annotators while improving alignment. While full technical reports for Claude models are not always released, Anthropic provides partial documentation, reinforcing its emphasis on selective transparency. Claude models have been deployed through APIs and integrated into commercial platforms, often in partnership with enterprise customers. Funding for Anthropic has come from several sources, including a substantial investment by Google, which has also collaborated on infrastructure provisioning via Google Cloud.

3.3. Gemini (Google DeepMind)

Google's contributions to LLMs span multiple entities, including Google Research and DeepMind [17]. The original PaLM (Pathways Language Model), introduced in 2022, had 540 billion parameters and demonstrated state-of-the-art performance on many few-shot NLP benchmarks [18]. PaLM was notable for its use of the Pathways system, which enabled efficient distributed training across thousands of TPUv4 chips.

In 2023, Google followed up with PaLM 2, a smaller but more efficient model family trained on multilingual and mathematical data, with significantly improved reasoning capabilities [19]. PaLM 2 reportedly includes models with parameter sizes of around 340 billion, although Google has not released full model specifications. The Gemini project [20], initiated through Google DeepMind in late 2023, represents Google's next-generation LLM series, combining deep neural architectures with multimodal processing abilities, including text, code, and image understanding. Gemini is a direct competitor to OpenAI's ChatGPT, with closed access and deployment via the Bard and Vertex AI platforms.

3.4. LLaMA (Meta)

Meta's approach to LLM development has emphasized openness and reproducibility [21]. In 2022, Meta released the OPT-175B model, replicating GPT-3's architecture and making it available to researchers with full training logs and code. This was followed in 2023 by the LLaMA (Large Language Model Meta AI) series, culminating in LLaMA 2, a family of models with sizes up to 70 billion parameters [22].

Unlike other U.S. models that remained proprietary, LLaMA 2 was open-sourced under a license permitting both research and commercial use, marking one of the largest publicly accessible LLMs at the time. Meta trained LLaMA 2 on a mixture of publicly available and licensed datasets, emphasizing transparency in both data composition and model evaluation. The company has continued to position LLaMA models as foundational tools for academic and industrial applications, in contrast to the API-based commercialization seen in other firms. Meta's strategy aligns with broader trends toward democratizing LLM access while still participating in competitive model scaling.

3.5. Azure and Copilot Integration (Microsoft)

Microsoft [23] has played a central role in the U.S. LLM ecosystem, primarily through its partnership with OpenAI [12] and its efforts to integrate LLMs into productivity tools. Azure [24] has become the principal cloud provider for OpenAI's models, hosting GPT-based APIs and enabling enterprise deployments of ChatGPT through Microsoft 365 applications such as Word and Excel under the Copilot [25] branding.

While Microsoft has not developed its own LLMs from scratch at the scale of ChatGPT or PaLM, it has contributed to fine-tuning, deployment optimization, and developing safety layers for downstream use. Microsoft also participates in research collaborations involving data governance, system interpretability, and model robustness. As of 2024, Microsoft's role is best characterized as a systems integrator and infrastructure partner, transforming LLM capabilities into widely distributed applications for commercial and institutional users.

3.6. Grok (xAI)

xAI [26], founded by Elon Musk in 2023, represents a distinctive entry into the U.S. LLM ecosystem, positioned at the intersection of conversational AI and real-time information retrieval. Its Grok [27] models are designed to integrate directly with the X platform (formerly Twitter), where they function not only as generative agents but also as systems for navigating live user-generated content. The technical direction of xAI emphasizes access to up-to-date information streams, which diverges from the conventional LLM design centered on static pretraining corpora. Rather than aiming solely for benchmark performance or closed-domain reasoning, Grok models prioritize responsiveness to current events, contextual engagement, and user-aligned behavior in a dynamic social media environment. xAI has also positioned itself in opposition to existing AI governance trends, framing its approach as more open to politically and socially sensitive discourse.

3.7. Titan and Nova Models (Amazon AWS)

Amazon, through its AWS division [28], has developed and deployed a suite of foundation models under the Titan [29] brand, with the Nova [30] series representing more recent, performance-optimized variants. Unlike some of its competitors, AWS has concentrated on enabling scalable and modular LLM deployment for enterprise customers via Bedrock, its managed AI service. The Titan and Nova models are not framed as public-facing chatbots but rather as foundational components for integration into vertical applications in sectors such as finance, retail, logistics, and healthcare. A defining aspect of AWS's strategy is abstraction: it provides model access through APIs and low-code interfaces while decoupling users from underlying model complexity. Technical emphasis is placed on data privacy, deployment flexibility, and customizable fine-tuning pipelines. The Nova models, trained in collaboration with Anthropic and other partners, offer general-purpose reasoning capabilities with safety mechanisms and guardrails suited for sensitive business environments. AWS's position is less focused on releasing standout frontier models and more on delivering stable, customizable models for infrastructure-scale applications.

3.8. Nemotron Ultra (Nvidia and Core Partners)

Nvidia, in collaboration with various ecosystem partners, has developed the Nemotron Ultra [31] series as part of its broader push to supply optimized LLMs tailored to its GPU and cloud platforms. Nemotron Ultra is designed primarily as a research and deployment reference for developers building on Nvidia hardware and software stacks. These models are intended to be highly efficient regarding inference cost and fine-tuning flexibility, making them suitable for deployment across various industry use cases. The architecture emphasizes supporting alignment workflows, controllable generation, and domain adaptation, often in conjunction with Nvidia's NeMo framework.

Unlike other proprietary systems, Nemotron Ultra is made available with documentation to facilitate experimentation. Nvidia's broader strategy does not rely on making its models dominant as stand-alone products but instead focuses on enabling other institutions to build performant models using Nvidia-optimized frameworks. This approach reinforces Nvidia's role as a critical infrastructure provider within the U.S. LLM landscape.

3.9. Perplexity AI: Retrieval-Augmented Generation (RAG) at Scale

Perplexity AI [32] is a fast-growing U.S. startup that takes a retrieval-augmented generation (RAG) approach as central to its product architecture. Rather than focusing on closed-ended generative models, Perplexity integrates LLMs with high-performance search and citation systems to ensure factual consistency and traceability. The core offering is a conversational interface that cites sources in real-time, allowing users to verify the origins of each generated response. This hybrid design is well-suited for knowledge-intensive tasks, making Perplexity a strong candidate for research, education, and expert systems deployment.

Perplexity's models are fine-tuned to operate efficiently within this retrieval paradigm, with latency and citation fidelity optimized as key metrics. The company's infrastructure supports live query expansion, document ranking, and multi-hop reasoning through combined generation and retrieval. In contrast to pure text-generation platforms, Perplexity frames itself as a transparent and source-grounded AI assistant, and its architecture reflects this commitment to factual reliability over creative fluency.

4. China

China's AI community has rapidly advanced large language models (LLMs) through both tech giants and new startups, producing numerous models that rival Western counterparts in capability [33]. A wave of open-source releases has lowered barriers to adoption, enabling widespread experimentation and application development. Major Chinese firms have integrated LLMs into search engines, enterprise cloud services, and consumer apps, while well-funded startups push the envelope in model efficiency, multilingual support, and ultra-long context handling. This section surveys the leading Chinese LLM initiatives – reorganized by their prominence in the ecosystem – detailing each model's history, technical design, openness, language scope, and deployment strategy. We highlighted the findings as Table 2.

Table 2. Comparison of Major China-Developed LLMs and Their Characteristics**.**

Company	LLM Series	Representative Models	Open-Source or Commercial Use	Characteristics
DeepSeek AI	DeepSeek	DeepSeek-V3, DeepSeek-R1	Open-Source	High performance with low training cost; bilingual (Chinese and English); Mixture-of-Experts architecture; top-ranked in Chinese benchmarks
Moonshot AI	Kimi	Kimi 1.0, Kimi k1.5	Commercial Use (Partially Open)	Ultra-long context (up to 2M characters); designed for enterprise use; supports multimodal inputs
Baidu	ERNIE	ERNIE 3.5, ERNIE 4.0	Commercial Use	Knowledge-enhanced pretraining; strong Chinese focus; integrated in search, cloud, and mobile apps

Alibaba	Tongyi Qianwen	Tongyi 1.0, 2.0; Qwen-7B/14B	Mixed (Open + Commercial)	Bilingual models; large-scale proprietary + open-source Qwen; widely deployed in Alibaba's ecosystem; frequently used in performance comparisons
ByteDance	Doubao	Doubao-Pro, Doubao-Lite	Commercial Use	Long context (128k tokens); aggressive pricing; integrated in Douyin and Feishu (Lark); strong in image generation
Huawei	PanGu	PanGu-Alpha, PanGu-Σ, PanGu 5.0	Commercial Use	Trillion-scale parameters; MoE structure; tailored for verticals (weather, finance, manufacturing); cloud integration
Tencent	Hunyuan	Hunyuan-VL	Commercial Use (Partially Open)	Dense Transformer; strong performance on Chinese tasks; integrated in WeChat, QQ, and fintech tools
iFlytek	SparkDesk	Spark v2.0–v4.0+	Commercial Use	NLP + speech capabilities; strong in education; supports text/audio
Baichuan Intelligence	Baichuan	Baichuan-7B, 13B, 53B, Baichuan 3/4	Open-Source	Apache license; bilingual; strong community adoption; high performance in Chinese; widely used in research
Zhipu AI	ChatGLM	ChatGLM-6B, ChatGLM2, ChatGLM-130B	Open-Source + API	Open-source and API accessible; bilingual; strong Chinese NLP; widely used in industry and academia
CAS	Zi Dong Tai Chu (ZDTC)	ZDTC 1.0, 2.0	Research Access Only	Multimodal model; supports text, vision, audio, and video; deployed in robotics and legal AI
SenseTime & Shanghai AI Lab	InternLM	InternLM-123B, InternLM2.0, InternLM-7B/20B	Mixed (Open + Commercial)	Long-context (300k chars); strong reasoning; multilingual; used in SenseChat and enterprise APIs
MiniMax AI	ABAB / MiniMax	ABAB-6.5, MiniMax-Text-01, MiniMax-VL-01	Mixed (Open + Commercial)	MoE-based; multimodal and multilingual; competitive open models and aggressive deployment

4.1. DeepSeek (DeepSeek AI)

DeepSeek [34] is a startup-led project that has upended the Chinese AI landscape with its fast-paced releases of high-performing, open LLMs. Founded in mid-2023 in Hangzhou, DeepSeek focused on cost-effective training innovations [35,36].

By late 2023, the company had released a series of models (V1, V2, V2.5) and in December unveiled DeepSeek-V3, matching the capabilities of much larger Western models at a fraction of the training cost [33,35] . Its flagship 2024 model, DeepSeek-R1, demonstrated reasoning performance on par with OpenAI's best systems while sharing its weights openly under an MIT License [37].

Technically, DeepSeek models leverage Transformer architectures augmented with innovations like Mixture-of-Experts layers to scale up parameters efficiently [35,36]. For example, DeepSeek-V3 reportedly uses around 70 billion parameters distributed across expert modules, allowing the model to attain frontier performance with significantly lower hardware requirements. The team also navigated U.S. chip export restrictions by using export-compliant GPUs and low-bit precision training, cutting total training expense to only $5 million[38]. DeepSeek's openness and cost breakthroughs have made it a cornerstone of China's open-source AI movement – its models (including code and weights) are published on platforms like HuggingFace [39], spurring a proliferation of downstream applications [40].

The startup offers an API and a chatbot (DeepSeek Chat) based on its latest model, but encourages integration of its models into other products. By early 2025, DeepSeek-V3 and R1 had rapidly become notable players, intensifying competition with Western firms by offering GPT-4-level reasoning and coding skills at much lower operational cost [40]. In summary, DeepSeek exemplifies the "fast, cheap, and open" approach – leveraging algorithmic efficiency to deliver an advanced bilingual model (Chinese and English) that is freely available for both research and commercial use. Its success has been described as "reenergizing" China's AI sector and proving that domestic teams can achieve cutting-edge results despite limited access to top-tier chips [33].

4.2. Kimi (Moonshot AI)

Kimi [41] is the LLM developed by Moonshot AI, a Beijing-based startup founded in 2023 that quickly rose to unicorn status [42]. Moonshot's Kimi model distinguishes itself with an extremely large context window and a focus on practical assistive capabilities.

By October 2023, the company launched Kimi 1.0, and in early 2024 it announced that Kimi could handle input prompts up to 2 million Chinese characters in length – a tenfold leap from its previous 200k character context limit [43]. This astonishing context size (equivalent to roughly 1.2–1.5 million tokens) is enabled by a "context caching" mechanism that stores and efficiently retrieves intermediate representations for long documents [44]. In practice, Kimi can ingest entire books or multi-hour meeting transcripts and perform summarization, question-answering, or code analysis across the whole content in a single session [42].

The model architecture is a Transformer with proprietary enhancements for long-sequence handling and multi-modal inputs – Moonshot has showcased a Kimi-Explorer version using chain-of-thought prompting and image understanding, indicating some multimodal ability. Kimi is bilingual but has a strong Chinese NLP proficiency; its creators optimized it for Chinese academic and business domains.

Moonshot AI emphasizes alignment and reliability – the updated Kimi k1.5 model [45] (with ~15 billion parameters) was benchmarked to be nearly as strong as ChatGPT on complex reasoning while maintaining low hallucination rates. In internal tests, Kimi's math and problem-solving even edged out some GPT-4 results, thanks to rigorous fine-tuning [46].

Notably, Moonshot has pursued an open-platform strategy: it offers Kimi's base models and an API for free (in beta) to invited developers. It plans to open-source a version of its model for the research community. The company also monetizes unique features like the long-context "memory" through its cloud service – initially charging usage-based fees before cutting prices by half amid competition.

Its commercial strategy centers on deploying Kimi as a productivity assistant (for coding, document analysis, etc.) within enterprises and education. In sum, Kimi's development highlights pushing the limits of context length and real-world usability. By combining ultra-long context processing with solid reasoning performance, Moonshot's Kimi has positioned itself as an enterprise-friendly chatbot that can "remember" and act on far more information than conventional models, a feature critical for tasks like lengthy legal analyses or reviewing entire codebases.

4.3. ERNIE Bot (Baidu)

Baidu's ERNIE Bot (Wenxin Yiyan) [47] is the earliest major LLM release by a Chinese tech giant and remains one of the most widely used in China. Building on Baidu's years of research in its ERNIE (Enhanced Representation through Knowledge Integration) [48] series, the ERNIE Bot was publicly launched in March 2023. The model's backbone is a Transformer with ~10 billion base parameters initially, trained on a massive 4 TB text corpus augmented by a large-scale knowledge graph for knowledge-enhanced pre-training [47]. This fusion of neural and symbolic data endowed ERNIE Bot with strong language understanding and factual recall, especially in Chinese. Baidu iterated quickly:

by June 2023, it had upgraded to ERNIE 3.5, and in October 2023, it announced ERNIE 4.0 with improvements in creativity, reasoning, and multimodal handling (including text-to-image generation) [49].

Baidu claims ERNIE 4.0 has reached parity with GPT-4 in many tasks and has integrated a long-memory module targeting 2–5 million character contexts in future versions. The ERNIE Bot is predominantly Chinese-centric (reflecting Baidu's search engine data), but it also possesses considerable English ability and can code in several programming languages. Alignment with Chinese content regulations was a key development focus – Baidu implemented stringent filters and fine-tuned ERNIE Bot to refuse disallowed content and adhere to social norms, which facilitated its regulatory approval as one of China's first batch of compliant LLM services in August 2023 [47].

Regarding deployment, Baidu deeply integrated ERNIE Bot into its products: it powers Baidu Search's chat mode, serves as an AI assistant in the Baidu App, and is offered via Baidu Cloud APIs to enterprise customers. Within weeks of launch, ERNIE Bot had attracted hundreds of business partners and by the end of 2023 it amassed over 100 million users in China [47]. By April 2024, Baidu reported the user base exceeded 200 million [47] – a reflection of its broad accessibility on the web, mobile, and as a plugin for office software.

Technically, Baidu continues to leverage its PaddlePaddle AI framework and has increased ERNIE's parameter count to "hundreds of billions" in the latest version, placing it among the world's largest dense LLMs [50]. The company also launched domain-specific fine-tunes (e.g. an ERNIE Finance and ERNIE Health model). Baidu's commercial strategy for ERNIE Bot is twofold: driving engagement and ad revenue via its consumer-facing applications, and selling cloud-platform access for enterprise solutions.

In summary, as China's earliest ChatGPT counterpart, ERNIE Bot has achieved wide adoption and continual technical enhancement. Its strengths lie in extensive training data, integration of factual knowledge, and Baidu's ecosystem leverage. These have solidified ERNIE's position as a foundational model in China's AI industry [50].

4.4. Tongyi Qianwen (Alibaba)

Alibaba's Tongyi Qianwen [51] (meaning "Truth from a Thousand Questions") is a family of LLMs developed by Alibaba Cloud that underscores the company's dual emphasis on model scale and openness. Debuted in April 2023, Tongyi Qianwen 1.0 was introduced as a 10-billion-parameter-class model powering Alibaba's enterprise applications and consumer products, such as the Tingwu assistant in DingTalk and Tmall Genie smart speakers [52,53].

By October 2023, Alibaba unveiled Tongyi Qianwen 2.0, announcing it had "hundreds of billions" of parameters—placing it among the largest LLMs globally [50]. This version came with fine-tuned variants; at its 2023 Cloud Summit, Alibaba released eight industry-specific versions tailored to finance, law, healthcare, and entertainment.

Technically, Tongyi Qianwen demonstrates strong bilingual proficiency (Chinese and English) and excels at tasks like instruction following, content generation, and code completion. Tongyi 2.0 incorporated enhancements in logical reasoning and factuality via techniques like supervised fine-tuning and reinforcement learning from human feedback (RLHF), built on top of a massive 20-trillion-token pretraining corpus [54].

In a move to democratize access, Alibaba also released Qwen-7B and Qwen-14B in August 2023. These smaller, open-source models were published under an Apache 2.0 license with training code and weights. The Qwen series performed strongly on Chinese NLP benchmarks like CLUE and provided a permissively licensed alternative to closed models like ChatGPT and partially open ones like Meta's LLaMA.

Deployment-wise, Tongyi Qianwen is deeply integrated into Alibaba's cloud ecosystem. The model is accessible via API through Alibaba Cloud, and by late 2023, the company claimed that over half of all LLMs in China were hosted on its infrastructure. Alibaba leverages its ModelScope hub to facilitate hosting third-party models alongside its own models.

Tongyi has been embedded in a range of Alibaba applications—e.g., office suites, e-commerce tools, customer service bots, and IoT devices—supporting capabilities like automatic summarization, product description generation, and real-time Q&A. The model adheres to Chinese regulatory standards and was one of the first LLMs officially approved for public deployment.

In summary, Tongyi Qianwen reflects Alibaba's full-stack approach to foundation models: combining high-scale proprietary models with accessible open-source releases, tailored deployment services, and competitive pricing.

4.5. Doubao & Cloud Lark (ByteDance)

ByteDance [55] – the owner of TikTok and a rising player in cloud AI – entered the LLM arena with its Doubao model family (formerly known internally as "Cloud Lark") [56,57].

In May 2024, ByteDance's cloud division (Volcano Engine) commercially launched the Doubao LLM suite [58], which comprises at least eight models [59]. These include Doubao-Pro (a general-purpose model with up to 128k token context length) and Doubao-Lite (a smaller base model), as well as specialized variants for speech recognition, speech synthesis, vision, and virtual character generation [59].

The Doubao models are built on transformer architectures comparable to GPT-3.5/4 and trained on multilingual data emphasizing Chinese. ByteDance has highlighted Doubao-Pro's long-form capability (its 128,000-token window is on par with GPT-4's extended context) and balanced bilingual skills. The company's approach heavily focuses on efficiency and affordability in deployment. At launch, ByteDance announced an aggressive pricing strategy: Doubao-Pro usage was priced at only ¥0.0008 per 1,000 tokens, much cheaper than OpenAI's GPT-4 pricing for the same input length [59]. This rock-bottom pricing – achieved through model optimization and distributed inference methods [57] – essentially initiated a price war in China's cloud AI market, undercutting rival services from Baidu and Alibaba by an order of magnitude [57,59].

Regarding openness, ByteDance has not open-sourced Doubao's largest models (which remain proprietary), but it has open-access APIs and even GUI apps. The Doubao Chat app became one of China's most downloaded AI apps in 2024, boasting over 25 million monthly active users [57]. ByteDance has integrated Doubao capabilities into its productivity suite (Feishu/Lark) via an assistant named "My AI", handling tasks like email drafting and meeting analysis [60].

On the consumer side, Doubao powers an AI companion chatbot within the Douyin (TikTok China) app. In summary, ByteDance's Doubao effort, while starting a bit later, has quickly become influential due to the company's vast platform reach and disruptive pricing. Its model family demonstrates solid general abilities and multimodal features, aiming to drive adoption by making generative AI ubiquitously accessible and budget-friendly in enterprise settings [57]. ByteDance's strategy of unifying its LLM research under Doubao and offering it through Volcano Engine positions it as both a top provider and an enabler (hosting others' models) – capturing value from China's LLM boom even beyond its own models.

4.6. PanGu Series (Huawei)

Huawei's PanGu series represents one of China's earliest and most ambitious large-scale LLM endeavors, marked by an orientation toward scientific and industrial applications. The project began in 2020 within Huawei Cloud's AI unit, and by mid-2021 Huawei introduced PanGu-Alpha, a 200-billion-parameter Chinese language model that was the country's answer to GPT-3.

In April 2023, Huawei researchers went further by unveiling PanGu-Σ (Sigma) [61], a colossal multi-language model with 1.085 trillion parameters [62] – achieved via a sparse Mixture-of-Experts (MoE) architecture. PanGu-Σ [61] was trained for over 100 days on 329 billion tokens across 40 different natural and programming languages using Huawei's MindSpore framework and Ascend AI processors [62]. The model uses Random Routed Experts (RRE) layers (Huawei's variant of MoE) atop a Transformer, allowing it to dynamically route inputs to different "experts." This design yields 6.3× faster training throughput compared to standard MoE baselines, and importantly it enables extracting smaller sub-models specialized for tasks like conversation, translation, or coding on demand [62].

PanGu-Σ achieved state-of-the-art results on 16 Chinese NLP tasks in zero-shot settings when introduced [61,62]. Following this research milestone, Huawei focused on making LLMs useful for industry. At the Huawei Developer Conference in July 2023, it launched PanGu 3.0, positioning it as a foundational model for government, finance, manufacturing, mining, and meteorology [62].

Rather than a single chatbot, PanGu 3.0 was delivered as a customizable platform: enterprises could adapt the hierarchical model to their own data and needs, emphasizing the execution of domain-specific tasks over open-ended creative generation [62]. For example, Huawei worked with the China Meteorological Administration to create PanGu-Weather, a global weather forecasting model that leveraged PanGu's architecture and outperformed traditional numerical prediction methods [62].

Huawei also unveiled a 100-billion-parameter financial LLM for fintech analysis in late 2023 [62]. By mid-2024, Huawei announced PanGu 5.0, which integrated these models into its HarmonyOS and Harmony Cloud platforms to imbue smartphones (via the Celia assistant) and enterprise services with generative AI capabilities [62].

PanGu models are generally not fully open-sourced, but Huawei has published papers and allowed controlled access for research collaboration. The company leverages PanGu to enhance its cloud offerings – for instance, in 2024 Huawei Cloud started offering PanGu-derived APIs for text summarization, code generation, and even drug discovery (there is a PanGu Drug Molecule model) [63,64]. Multimodality is another theme: PanGu's multimodal branch (co-developed with CAS for Zidong Taichu 1.0 in 2021) learned unified representations for image-text-audio, setting a precedent for subsequent Chinese multimodal LLMs [65].

Overall, Huawei's PanGu stands out for its technical feats (pushing parameter count to trillion-scale and innovating on training efficiency) and its focus on real-world deployment in vertical domains. Huawei treats LLMs as a strategic component to "embed intelligence in everything" – from integrating AI into cloud infrastructure to enabling new applications like AI-powered weather forecasting and pharmaceutical research.

4.7. Hunyuan (Tencent)

Tencent's Hunyuan LLM [66] is the tech giant's internally-developed foundation model, notable for its integration into Tencent's broad product ecosystem and strong performance on Chinese language tasks. Hunyuan was unveiled in September 2023 at Tencent's Global Digital Ecosystem Summit [10]. The model boasts over 100 billion parameters and was trained on more than 2 trillion tokens of data [10], giving it a knowledge base spanning vast Chinese and English internet content.

Tencent's researchers have emphasized Hunyuan's balanced bilingual capability and its strength in handling Chinese idioms, slang, and long-form texts [67] – leveraging Tencent's experience from its popular WeChat and QQ platforms. Technically, Hunyuan follows a dense Transformer architecture with custom pre-training objectives [67]. Tencent noted that Hunyuan's training corpus included diverse sources (news, social media, code, and academic texts), and the model underwent extensive fine-tuning for alignment and factual accuracy.

In benchmarking, Tencent claimed Hunyuan outperforms OpenAI's ChatGPT (GPT-3.5) in several areas relevant to Chinese users: for example, generating lengthy, coherent texts (thousands of words) and solving complex mathematical problems in Chinese [10]. It also reportedly exhibits a 30% lower hallucination rate than Meta's LLaMA-2 on evaluation prompts [10]. These claims were backed by evaluations using the SuperCLUE and C-Eval benchmarks, where Hunyuan scored among the top models in late 2023.

Unlike some competitors, Tencent did not immediately open-source Hunyuan, but it did release an open-source multimodal extension called Hunyuan-VL (for vision-language tasks) and a series of smaller MoE models named Hunyuan-Large [67,68]. These moves aim to foster a developer community around Tencent's AI.

Strategically, Tencent leverages Hunyuan to maintain its dominance in social and gaming contexts – for instance, experimenting with AI NPCs in video games and advanced content recommendations. In corporate use, Tencent's fintech arm uses Hunyuan for financial analysis and customer service bots.

Tencent President Martin Lau described the landscape in late 2023 as a "war of a hundred models" and positioned Hunyuan as one of the front-runners, especially given Tencent's unmatched user data and social media insights to further refine the model [10,50] .

In summary, Hunyuan is Tencent's bid for LLM leadership, characterized by a large, high-quality bilingual model tightly integrated with the company's platforms, reaching over a billion users. Its development underscores Tencent's measured approach – achieving top-tier performance and reliability (emphasizing less hallucination and better long-text handling) before scaling out access.

4.8. Spark Model (iFlytek)

The Spark large model (Xunfei SparkDesk) [69] by iFlytek is a noteworthy entrant, especially given iFlytek's background in speech recognition and education technology [70]. Spark was first released in May 2023 as one of China's earliest ChatGPT-like systems, and it has since undergone rapid upgrades (v2.0 in June, v3.0 in August, v4.0 in October 2023, and v4.0+ in 2024) [70, 71]. By late 2023, iFlytek claimed Spark had reached or surpassed OpenAI's GPT-4 on key Chinese language benchmarks [72]. Specifically, Spark v3.5 (announced January 2024) was reported to outperform GPT-4 Turbo in Chinese-centric tasks such as language understanding, knowledge Q&A, and mathematical reasoning.

Spark v4.0, launched in June 2024, was said to rank first on eight international evaluation leaderboards and demonstrated superior performance to GPT-4 in areas like long-form text generation, complex comprehension, and logical reasoning in Chinese [72]. While made by the company, these claims were partly corroborated by third-party evaluations (the Xinhua News Agency's benchmarking in Aug 2023 put Spark v2.0 slightly ahead of GPT-3.5 in Chinese and closed the gap to GPT-4). Technically, the Spark model architecture is built on a transformer with around 100 billion parameters (exact size not publicly disclosed, but hinted by hardware requirements) and is trained on a massive bilingual corpus, with a special focus on Chinese educational and legal documents.

Uniquely, iFlytek has leveraged its speech tech expertise to imbue Spark with strong speech integration: Spark can convert text to highly realistic speech in multiple languages and understand spoken prompts, enabling voice conversations. By v4.0, Spark supported 74 languages and dialects for speech and introduced robust speech-to-text and text-to-speech capabilities that won international awards [72]. This makes Spark a multimodal model (text + audio) with an edge in applications like AI voice assistants and automatic lecture transcription.

Another differentiator is iFlytek's focus on educational applications – the company integrated Spark into an intelligent education assistant that can grade exam papers, generate feedback, and tailor study plans. In demonstrations, Spark's grading system could evaluate and score student essays and math solutions nearly as well as human teachers, dramatically reducing grading time [72]. This plays to iFlytek's strength in the education sector (where it has a significant market presence in China).

The Spark model is not open-sourced, but iFlytek provides access via its Open Platform. It was among the first LLMs granted public deployment permission, and iFlytek wastes no opportunity to publicize Spark's prowess. As of 2024, Spark had been integrated into iFlytek's consumer translation devices, office software, and innovative classroom solutions. The company's monetization involves API subscriptions for enterprises (in competition with Baidu and others) and bundling AI features with its hardware (like translation earbuds that use Spark for real-time bilingual conversation).

In summary, iFlytek's Spark stands out for its speech and education orientation. It combines a powerful LLM with speech synthesis/recognition and domain fine-tuning, aiming to be the go-to model for Chinese-language education and communication. Its rapid progress (multiple version leaps in months) showcases how competition has driven even specialized players like iFlytek to iterate toward GPT-4-level performance in their niche. Spark's evolution also illustrates the convergence trend between voice AI and text AI, merging iFlytek's long-held speech tech dominance with LLM capabilities to create versatile AI assistants.

4.9. Baichuan Model Family (Baichuan Intelligence)

Baichuan Intelligence [73], a startup founded in April 2023 by former Sogou CEO Wang Xiaochuan, has emerged as a leading contributor of open-source LLMs in China. Baichuan released a series of models, aiming to provide a "Chinese version of OpenAI's foundation models" [74,75].

The Baichuan family began with Baichuan-7B, a 7-billion-parameter bilingual model unveiled in June 2023 and made openly available on platforms like Hugging Face [74]. Trained on both Chinese and English text, Baichuan-7B achieved surprisingly strong results for its size and was licensed for commercial use, quickly being adopted by independent developers as a lightweight ChatGPT alternative.

Remarkably, Baichuan followed up just 26 days later by open-sourcing Baichuan-13B, a 13B-parameter model that further improved performance and was also released under an Apache 2.0 license [75]. These releases filled a critical gap for permissively licensed Chinese LLMs and earned Baichuan a reputation akin to "China's LLaMA provider."

The startup didn't stop at small models: in August 2023 it introduced Baichuan-53B, and by January 2024 Baichuan announced it had a 100-billion+ parameter model (Baichuan 3) in testing [74]. The Baichuan 2 series, launched in late 2023, featured improved training (on 2 trillion tokens) and extended context lengths – one variant, Baichuan2-192k, supports a 192,000 token window to facilitate long document processing [74].

Baichuan models are characterized by strong Chinese NLP capabilities (e.g. high scores on the Chinese CMMLU and Gaokao benchmarks) while maintaining competitive English understanding. The team attributes this to a balanced training corpus and advanced pre-training techniques like data mixing and RoPE positional encoding for long context.

An essential aspect of Baichuan's models is openness and compliance. The company's strategy appears to be providing base models for others to build on ("model-as-a-service"); indeed, by late 2023 it opened API access to its 53B model for enterprise clients and launched Baichuan Cloud for developers.

In mid-2024, Baichuan rolled out Baichuan-4, its fourth-generation model, alongside a chatbot product called Baixiao Yin [76]. While details on Baichuan-4 are limited, they presumably cross the 100B-parameter scale and incorporate retrieval augmentation for better factuality. Baichuan has garnered accolades, making the 2024 Forbes China AI 50 list and the Hurun Global Unicorn list with a valuation of around RMB 7.1 billion [74].

In summary, Baichuan Intelligence has quickly established itself as a key player by delivering high-quality LLMs and releasing them openly. This has catalyzed innovation in China's AI community – Baichuan models have been used as the backbone for countless downstream projects, from chatbots and coding assistants to academic research. The Baichuan saga exemplifies how a nimble startup can compete with tech giants by embracing openness and fast iteration, making advanced bilingual models accessible and customizable.

4.10. ChatGLM Series (Zhipu AI & Tsinghua University)

The ChatGLM [77] series, developed by Zhipu AI in collaboration with Tsinghua University's Knowledge Engineering Group (KEG) lab, is another pillar of China's LLM ecosystem, especially in open-source and research contexts.

Zhipu's journey began with the GLM-130B model introduced in 2022 – a 130-billion-parameter bilingual (Chinese-English) model based on the General Language Model (GLM) framework [78]. GLM-130B was notable for being one of the first open 100B+ models; although initially released for research, it demonstrated that Chinese academics could produce a GPT-3 class model and share it with the world.

Building on that, in early 2023 Zhipu released ChatGLM-6B, a 6-billion-parameter conversational model distilled from GLM-130B. ChatGLM-6B was open-sourced with an Apache License, allowing unrestricted use, and it quickly became popular due to its ability to run on a single consumer GPU (requiring as little as 6 GB VRAM) [78].

Despite its small size, ChatGLM-6B showed surprisingly strong performance on Chinese queries and garnered tens of thousands of downloads globally. The ChatGLM series continued evolving: by mid-2023, ChatGLM2-6B was released with improved training stability, longer context (32k tokens), and an open commercial license, making it one of the most practical open Chinese models.

Meanwhile, Zhipu kept a larger model for cloud API use – ChatGLM-130B (sometimes referred to as GLM-130B-int8), which was made accessible via an interactive web demo and API. According to the team, ChatGLM (130B) achieved about 85% of ChatGPT's performance and outperformed ChatGPT on many Chinese language tasks [79]. Indeed, a Nature article in May 2024 noted that ChatGLM was closing the gap with ChatGPT and even surpassing it in Chinese-language benchmarks, as claimed by its creators [79].

The model family uses a bidirectional GLM architecture for pre-training (combining autoencoding and autoregressive blank filling) and then transitions to an autoregressive chat format [80]. This two-stage approach and intensive supervised fine-tuning and feedback tuning give ChatGLM robust conversational abilities. Zhipu has integrated these models into a platform called Qingyan for enterprises.

Nonetheless, the company thrives, prepping an IPO and expanding its model lineup (recently announcing ChatGLM3 in 2024). The impact of ChatGLM in China's LLM landscape is significant: it lowered the barrier to entry for researchers and startups by providing powerful models that can be run locally without relying on foreign APIs. Many university labs and smaller firms have built their own applications on ChatGLM-6B or GLM-130B. In essence, the ChatGLM series serves as a democratizing force, ensuring that advanced LLM technology is not limited to tech giants. By being open-source and bilingual, it also bridges the Chinese and international AI communities.

With continued academic collaboration (several papers [81,82] on GLM have been published at major conferences) and open innovation, ChatGLM and GLM models exemplify how China's AI researchers are contributing at the cutting edge while embracing openness – "China's ChatGPT," as Nature dubbed it, built in the open for all to use [79].

4.11. Zi Dong Tai Chu (Chinese Academy of Sciences)

Zi Dong Tai Chu (ZDTC) [83] is a multimodal LLM developed by the Institute of Automation, Chinese Academy of Sciences (CAS), representing the state-led effort to push AI frontiers. The name, taken from an idiom meaning "Genesis" , befits its broad ambition – ZDTC was designed from the outset to handle text, vision, and other modalities in a unified model [84].

The first version, ZDTC 1.0, was released in 2021 as a proof-of-concept "full-modal" model with 100 billion parameters, heralded as the world's first 100B-scale multimodal AI system [65]. It was jointly trained by CAS and Huawei on a diverse dataset of images, text (Chinese and English), and audio, enabling it to perform cross-modal generation (e.g. describe an image in text or generate an image from text) via a single model [65].

By mid-2023, CAS unveiled ZDTC 2.0, which expanded the model's training data to include video, live sensor signals, and 3D point cloud data. This allowed version 2.0 to interpret and generate a wider array of information – for instance, analyzing surveillance video feeds, performing audio-visual speech recognition, and understanding LiDAR data for autonomous driving. An example application demonstrated at the launch was using Zidong 2.0 in a neurosurgical robot: the model could integrate visual and tactile data during endoscopic surgery and provide real-time reasoning to assist the surgeon.

In another scenario, the model ingested legal case documents (text + scanned images) and extracted key information within seconds, showing promise in legal AI assistance [65]. Architecturally, ZDTC combines multiple subnetworks: a text encoder/decoder, a vision encoder, an audio processor, etc., all mapped into a shared semantic space. The training was performed on CAS's advanced supercomputing facilities with Huawei Ascend chips (the project is a showcase for China's domestic AI hardware).

In terms of performance, while not directly comparable to specialized text-only LLMs on pure NLP tasks, ZDTC is extremely valuable for tasks requiring multi-modal understanding. It has been evaluated on Chinese vision-language benchmarks and achieved state-of-the-art results in 2022–2023. By early 2024, CAS announced that ZDTC 3.0 was developing, focusing on enhanced reasoning and tool-use abilities [65,84].

In summary, ZDTC represents the cutting-edge of government-backed AI research in China – a large-scale, multi-domain AI brain intended to be a "universal" model for text, vision, and beyond. Its successive versions show a clear trajectory: rapidly incorporating more data modalities and cognitive abilities (from perception to reasoning to tool use). This ambitious initiative underscores China's drive to match and leapfrog by creating AI models that are broader in scope than any single-function LLM [85].

4.12. InternLM (SenseTime)

InternLM is a series of large language models jointly developed by SenseTime (a leading AI company) and the Shanghai AI Laboratory, exemplifying a collaboration between industry and state research to produce advanced multilingual models. The flagship InternLM, introduced in mid-2023, has about 104 billion parameters and was trained on an extremely large corpus of 1.6 trillion tokens spanning Chinese, English, and code [86,87].

SenseTime invested massive computing into this effort – reportedly around 10,000 GPUs were used to train InternLM [88]. Upon its debut, InternLM achieved breakthrough results; SenseTime reported that the refined InternLM-123B model (after further tuning) exceeded OpenAI's GPT-4 on 12 authoritative benchmark tests [88]. These tests ranged from knowledge quizzes to coding problems, indicating InternLM's all-round capability. The model architecture is a standard transformer with optimizations for long context (up to 32k tokens) and efficient inference.

SenseTime also emphasized alignment and safety: InternLM was fine-tuned with human feedback and additional training by Chinese linguists to ensure it follows instructions accurately in both Chinese and English, while filtering sensitive content.

In June 2023, SenseTime launched "SenseChat," a chat application based on InternLM, and opened it to the public after receiving regulatory clearance [88]. This made SenseTime one of the first companies (along with Baidu and Baichuan) to offer a legal ChatGPT-like service in China. SenseChat showcases InternLM's abilities in tasks like creative writing, programming Q&A, and multi-turn reasoning in a user-friendly interface.

Concurrently, SenseTime made InternLM available via its cloud API, allowing businesses to integrate the model into their workflows. For instance, financial firms use InternLM through SenseTime's platform for report analysis and customer interaction bots (internlm/internlm-xcomposer2-vl-7b - Hugging Face) [86].

The InternLM series also includes smaller derivative models: SenseTime released InternLM-7B and InternLM-20B models, including specialized versions like InternLM-Math-7B (fine-tuned for mathematical reasoning), which reportedly surpassed ChatGPT despite its small size [89]. These lighter models (collectively called InternLM 2.5) were open-sourced in late 2023 for researchers, underlining SenseTime's partial commitment to open science.

Moreover, in early 2024 a follow-up model InternLM 2.0 (sometimes dubbed InternLM2) was released, featuring enhanced long-context handling (up to 300k characters) and the ability to invoke tools such as calculators and search engines during its responses [90]. This model can read and summarize extremely lengthy texts (e.g. entire financial reports) and was offered free for commercial use with Shanghai AI Lab's permission, to encourage adoption [90].

SenseTime's commercialization strategy involves leveraging its reputation in computer vision and AI platforms. It is incorporating InternLM into its product suite – from the SenseNova AI-as-a-service platform to smart city solutions where an LLM can, for example, understand traffic incident reports. It is a multilingual, general-purpose LLM that SenseTime deploys across various domains, and its iterative improvements (e.g. InternLM2's tool-use and ultra-long memory) show a drive to lead the next generation of AI assistants. With its blend of openness (smaller models) and enterprise solutions (full 100B model via cloud), SenseTime positions InternLM as both a research platform and a competitive commercial AI service.

4.13. ABAB & MiniMax Models (MiniMax AI)

MiniMax [91], an AI startup backed by Alibaba and others, has emerged as a notable force in China's generative AI landscape. Its model family includes ABAB, a conversational model series that adopts a Mixture-of-Experts (MoE) architecture. In April 2024, MiniMax released ABAB-6.5, a 30B-parameter model optimized for open-domain dialogue and task specialization through expert routing [92].

In early 2025, MiniMax introduced the MiniMax-01 series, comprising a general-purpose dense model (MiniMax-Text-01) and a multimodal variant (MiniMax-VL-01) capable of processing both text and images [92]. These models were released under open licenses to encourage adoption and competition. Internal evaluations emphasize accuracy, long-context support, and the ability to handle multilingual and multimodal inputs. The company has integrated these models into applications such as virtual companions, image-generation tools, and multilingual text-to-speech systems.

MiniMax differentiates itself through low-cost, open access. In January 2025, it released three lightweight open models positioned as efficient ChatGPT alternatives. The company has secured over $600M in funding and collaborates with hardware makers to deploy its models on consumer devices.

In summary, MiniMax demonstrates a dual strategy of MoE research and practical deployment. By releasing competitive models and fostering a developer ecosystem, it has quickly become one of China's most active LLM startups.

5. European Union

The European Union's contributions to large language models are shaped by an emphasis on transparency, multilingual capabilities, and digital sovereignty. In contrast to the commercial race seen in the United States and China, EU efforts are often driven by public-private collaborations, prioritize compliance with data protection laws, and aim to support diverse linguistic and regulatory contexts across member states. Development practices tend to emphasize openness, responsible deployment, and alignment with European values. EU-based models are typically designed for efficient deployment, traceability, and adaptability to local infrastructure. The ecosystem benefits from EU-funded high-performance computing infrastructure, cross-border research coordination, and legal frameworks that guide both model development and responsible use. We highlighted the findings as Table 3.

Table 3. Comparison of Major European Union-Developed LLMs and Their Characteristics**.**

Company	LLM Series	Representative Models	Open-Source or Commercial Use	Characteristics
France	BLOOM	BLOOM 176B	Open-Source (Responsible AI License)	Multilingual (46 languages); trained on public compute; early open 100B+ model; strong research transparency

Germany	Luminous	Luminous-base, Luminous-supreme (70B)	Commercial API (partial open)	European-language focus; explainability tools; GDPR-compliant cloud deployment; strong efficiency
France	Mistral	Mistral 7B, Mistral NeMo 12B	Open-Source	Strong performance at small scale; Grouped-Query & Sliding Window Attention; efficient long context handling
Germany / EU Consortium	OpenGPT-X	Teuken-7B	Open-Source	Trained on EU supercomputers; 24 EU languages; public infrastructure use; compliant with EU AI Act goals

5.1. BLOOM (France): The BigScience Open Multilingual Model

BLOOM [11,93] is a 176-billion-parameter LLM developed through the BigScience project, a year-long global research workshop with strong European involvement (led in France) aimed at democratizing access to large models [94]. Backed by France's national supercomputing center (GENCI/IDRIS) and Hugging Face, BLOOM was trained on the French Jean Zay supercomputer using public research grants [93,95]. The project's motivation was to create an open-access model as an alternative to proprietary LLMs, aligning with European open science ideals.

Regarding model architecture and scale, BLOOM adopts a decoder-only Transformer architecture (similar to GPT-style models) with 176 billion parameters [95]. Training spanned 3.5 months on 384 A100 GPUs, consuming over 1 million compute hours [95] – a scale made possible by public supercomputing resources. Despite its size, BLOOM was one of the first openly released models in the 100B+ parameter range. Its performance on standard NLP benchmarks is competitive with other large models, and it demonstrated strong few-shot learning capabilities.

For multilingual capability, a distinguishing feature of BLOOM is its broad multilingual support. It was trained on the ROOTS corpus, containing text in 46 natural languages (including many European languages) and 13 programming languages [95]. At least 30% of its training data is non-English, giving it robust ability across diverse languages. This multilingual orientation addresses Europe's linguistic diversity and helps avoid an English-centric bias in AI technology. Early evaluations confirmed BLOOM's stable performance in French, Spanish, Arabic, and other languages, making it a valuable foundation model for multilingual NLP research [95].

Regarding use and policy alignment, BLOOM was released in 2022 with an open model license (a Responsible AI License) that allows free access for research and application while urging compliance with ethical use guidelines [95]. The model card and extensive documentation included transparency about training data and potential biases, aligning with the European policy emphasis on responsible AI. Notably, BLOOM's development involved AI ethicists and a diverse team (over 1000 contributors), and the release was accompanied by research on the model's environmental impact and bias analysis. This collaborative and transparent approach anticipated upcoming EU AI Act requirements for foundation models to disclose training data and risk assessments.

In summary, BLOOM represents a milestone in open science – a large-scale, multilingual model "made in Europe" that advances AI sovereignty while adhering to European values of openness and inclusivity [94].

5.2. Aleph Alpha's Luminous (Germany): Efficient Multilingual AI for Europe

Aleph Alpha [96], a Heidelberg-based AI startup, launched Luminous [97] as a family of large language models positioned as a European answer to US-developed GPT models [98]. Founded in 2019, Aleph Alpha received German government support and venture funding to pursue "AI sovereignty" – ensuring Europe has its own foundation models and infrastructure. The development of Luminous aligns with Europe's drive for digital sovereignty: keeping data processing in Europe (addressing GDPR and CLOUD Act concerns) and customizing AI to European languages and norms

[99]. By 2022–2023, Aleph Alpha had built several Luminous model variants and demonstrated their capabilities to industry and public-sector partners.

Regarding model architecture and scale, Luminous models [97] are decoder-only Transformer LLMs with parameter scales ranging from about 13B up to 70B. Specifically, Luminous-base has 13 billion parameters, Luminous-extended ~30B, and Luminous-supreme ~70B as the largest model [99]. They use standard architectures with enhancements such as rotary positional embeddings for efficiency [100]. Despite being smaller than GPT-3 (175B), Luminous-supreme was benchmarked to perform on par with GPT-3 on various language tasks [101]. A February 2023 performance report showed Luminous (70B) achieving comparable scores to GPT-3 while using less than half the parameters, indicating twice the efficiency in size vs. performance [101]. This efficiency reflects insights similar to the Chinchilla scaling law, emphasizing optimal training data volume over sheer model size. Aleph Alpha has also experimented with even larger models – planning a 300B-parameter Luminous-World model as of 2023 [101] – though their primary focus remains to make models efficient and controllable rather than simply the largest.

For the multilingual and European focus, Luminous was trained on a curated multilingual corpus focusing on major European languages. According to Aleph Alpha, the training data included English, German, French, Italian, and Spanish sources (among others) totaling ~400–590 billion tokens [100]. The resulting models can work in multiple languages, although their strongest capabilities are in English and German. By tailoring Luminous to European languages and cultural context, Aleph Alpha ensures the model serves local business and government needs. For example, the city of Heidelberg piloted a Luminous-based chatbot (“Lumi”) to assist citizens, which can answer questions in German and provide sources for its answers [101]. This application highlights Aleph Alpha’s emphasis on explainability and traceability in LLMs. The company introduced an “Explain” feature that allows users to see which input text segments influenced the model’s output [99] – an approach to mitigate hallucinations and increase trust. Such features directly address European principles of transparency and accountability in AI (as later formalized in the EU AI Act).

Regarding intended use and responsible AI, Aleph Alpha offers Luminous via a controlled API for enterprise and government use rather than releasing weights publicly (at least for the larger models). This managed approach ensures user data stays on European servers and complies with privacy regulations, a key concern for European clients [99]. At the same time, Aleph Alpha has advocated open research: they publish model performance benchmarks and allow academic access. In 2024, anticipating the AI Act, Aleph Alpha went a step further by releasing two open-source LLMs (7B parameters each) called “Pharma" – designed to be fully compliant with upcoming EU rules [102]. These models (one base and one aligned for less toxic outputs) are tuned to European languages (especially German, French, Spanish) and were open-sourced under a permissive license. The move demonstrated that European AI companies can embrace openness and compliance simultaneously, providing a “counterbalance to U.S. AI dominance” in line with EU policy goals [102].

In summary, Aleph Alpha’s Luminous initiative illustrates a European pathway to LLM development: create competitive models (70B+) that are efficient, multilingual, and trustworthy, deliver them in a sovereignty-preserving manner (European cloud and explainability features), and gradually open-source smaller versions to foster community adoption under European compliance standards.

5.3. Mistral AI (France): Small, High-Performance Models with Open Release

Mistral AI [103] is a French startup founded in 2023 by former Meta AI and DeepMind researchers aiming to develop state-of-the-art LLMs while embracing open-source principles. The venture quickly gained prominence by raising a record seed funding in Europe, signaling the strategic importance of LLMs for European tech sovereignty. Mistral’s formation came amid European calls for more home-grown AI capabilities and was bolstered by France’s talent and public support for AI research. By focusing on relatively smaller models that are highly optimized, Mistral aims to provide powerful LLMs that can be deployed by a wide range of companies and researchers (rather than only tech giants).

For model architecture and innovations, the first model, Mistral 7B [104], was released in September 2023 and drew widespread attention for its strong performance relative to its size. Mistral 7B is a 7.3-billion-parameter decoder-only Transformer model that outperforms Meta's LLaMA-2 model with 13B parameters on all tested benchmarks [105], and even rivals 30B+ models on specific tasks. This was achieved through several architectural and training innovations. Notably, Mistral 7B uses Grouped-Query Attention (GQA) to improve throughput at inference, and Sliding Window Attention to efficiently handle longer context lengths than typical models of its size [105]. These modifications allow Mistral to process long inputs at lower computational cost, which benefits applications like long document summarization. The training was likely done on large-scale GPU clusters (the team has mentioned collaboration with hardware partners); specific details were not fully public, but efficiency was a key design criterion. By carefully curating training data and hyperparameters, the Mistral team was able to push a 7B model to unprecedented levels of accuracy on reasoning, knowledge, and coding tasks [105]. This result echoed the "smaller is smarter" trend, showing that a well-trained 7B model can match or beat larger models that are not optimized as effectively.

Regarding open-source release and usage: Mistral AI released the Mistral 7B model [104] weights openly under the Apache 2.0 license in 2023. This means anyone can use, fine-tune, or deploy the model without restrictions – a stark contrast to most commercial LLM releases. The open release aligns with the European open science and innovation ethos, enabling researchers and startups to build on Mistral 7B for their own purposes. Alongside the base model, Mistral provided a demonstration chatbot (fine-tuned model) that surpasses LLaMA-2 13B-chat in conversational ability. The model's strong coding performance (nearly matching OpenAI's Code-Cushman on some code tasks) also makes it attractive for code assistant applications. Because the model is relatively small, it can be deployed on-premise or even on high-end consumer hardware, giving organizations control over their data – a feature appealing under EU data protection norms. Mistral's documentation and code were released on GitHub, inviting community contributions and transparent scrutiny of the model, which supports the EU's stance on AI transparency.

For multilingual support and extensions, while Mistral 7B's initial focus was on English performance (and it showed excellent results on English benchmarks [105], the developers did not ignore multilingualism. In mid-2024, Mistral AI introduced Mistral 7B v0.2 and Mistral NeMo (12B) [106], which were trained on a much broader set of languages. Mistral NeMo 12B, developed with NVIDIA, features a 128k token context window and is explicitly designed for global multilingual applications, with top-tier proficiency in French, German, Spanish, Italian, Portuguese, Chinese, Japanese, Korean, Arabic, Hindi, in addition to English [106]. These advancements illustrate that Mistral's roadmap includes catering to Europe's multilingual environment.

Alignment with European AI Policy, Mistral's open and transparent approach also extends to responsible AI considerations. Open model weights allow the community to audit and identify biases or risks. Moreover, Mistral AI provides a model moderation service and encourages users to fine-tune the base model for safe deployment [107]. This modular approach (separating base and moderation models) could facilitate compliance with the EU AI Act's risk mitigation requirements. France's national AI strategy emphasizes both excellence and ethics; Mistral's work, supported by European investors, exemplifies this by pushing technical boundaries while adopting open releases and discussing limitations openly.

In summary, Mistral AI has contributed a new class of efficient LLMs that empower users with limited compute resources, and by open-sourcing these models, they strengthen the European ecosystem of trustable and accessible AI.

5.4. OpenGPT-X (Germany/EU Consortium): Large Models for European Sovereignty

OpenGPT-X [108] is a Germany-led initiative to create large language models "Made in Europe", tailored to European languages and industrial use cases. It began as a multi-partner research project in 2021, funded by the German Federal Ministry for Economic Affairs and Climate Action (BMWK) as part of a broader European cloud and AI sovereignty effort. The consortium comprises Fraunhofer Institutes (IAIS and IIS as leads), industry partners (e.g., Deutsche Telekom), and research centers, including the Jülich Supercomputing Centre.

By mid-2022, the project set out to develop open, multilingual LLMs for Europe's needs, integrated with European high-performance computing infrastructure. This context reflects EU policies like the Digital Europe Programme, which encourages European collaboration in foundational AI technologies, and the Gaia-X initiative for federated cloud infrastructure.

For model development and infrastructure, In November 2024, OpenGPT-X announced its first major model release: "Teutonic Transformer 7B" (code-named Teuken-7B), a model with 7 billion parameters [109]. Teuken-7B was trained from scratch in all 24 official languages of the European Union, making it one of the few LLMs explicitly built to encompass the full linguistic diversity of the EU [109]. The training was conducted using Europe's supercomputers – notably the JUWELS supercomputer at Jülich, a leading Tier-0 system in Germany [109]. Leveraging such HPC resources was crucial to handle the massive training workload and large multilingual dataset. The result was a model with approximately 50% of its training data in non-English languages, ensuring strong multilingual capabilities [109]. OpenGPT-X reports that Teuken-7B achieves stable and reliable performance across many languages without the steep drop-off often seen when English-trained models are applied to lesser-resourced languages. This is especially valuable for European companies and public services operating in multiple countries. The model can be further fine-tuned for specific domains and remains commercially usable open-source, meaning companies can deploy it on-premises and keep sensitive data in-house. These features align well with European priorities around data sovereignty and trust.

Regarding technical features and research, in developing Teuken-7B, the OpenGPT-X team also tackled research questions related to energy efficiency and tokenization for multilingual AI [109]. They designed a new multilingual tokenizer capable of handling the varied character sets and word structures of Europe's languages (from Finnish to Greek) more compactly. A more efficient tokenizer reduces the number of tokens per sentence, thereby lowering computational cost – an important consideration for sustainable AI given Europe's focus on green computing. The project's researchers are investigating methods to scale models further (with larger parameter counts) while controlling the energy usage, potentially utilizing upcoming exascale supercomputers like JUPITER in Germany (the first EU exascale system, operational in 2024–25) for training next-generation LLMs. Indeed, EuroHPC Joint Undertaking, which funds JUPITER and other supercomputers, explicitly lists AI and big data as key applications for these infrastructures [110]. Thus, OpenGPT-X sits at the intersection of Europe's HPC strength and AI ambitions, illustrating how public infrastructure can drive cutting-edge model development outside Big Tech.

For use cases and policy alignment, the OpenGPT-X models are intended to serve European industry and public sectors with use cases such as enterprise document analysis, customer service in multiple languages, and government applications in local languages. By focusing on a "distinctly European perspective" in model development [109], the consortium ensures that cultural nuances and values (like GDPR compliance, avoidance of extremist content in European contexts, etc.) are taken into account. The openness of Teuken-7B (available on Hugging Face and usable without royalties) fosters an open ecosystem: companies can fine-tune it for their needs, and academic groups can study or improve it. This openness is coupled with understanding the EU's emerging AI regulation. The project frames itself as delivering "transparent and compliant" models within Europe's regulatory framework [111]. For example, documentation includes summaries of training data provenance, which anticipates the AI Act's requirement for foundation model providers to disclose a "sufficiently detailed summary of the training data" used.

Overall, OpenGPT-X exemplifies Europe's collaborative approach: a publicly funded, multi-stakeholder effort producing multilingual, open-source LLMs optimized for European languages and legal norms. It strengthens European digital sovereignty by reducing reliance on foreign AI APIs and showcasing that top-tier AI can be built in a European way, harnessing both technology and policy support.

6. United Kingdom

The United Kingdom has contributed to large language model development through both academic research institutions and startups, with a focus on efficient scaling, safety evaluation, and open access. The UK's ecosystem combines strong theoretical foundations with a commitment to transparency and public engagement. While fewer in number compared to U.S. or Chinese firms, UK-based efforts have had significant influence—particularly through innovations in compute-efficient training and open-source deployment. Research from institutions like DeepMind has introduced widely adopted scaling and optimization principles, while startups such as Stability AI are reshaping model accessibility through open models. These contributions align with broader UK goals of fostering innovation, reducing concentration of AI capabilities, and supporting AI safety via transparent research. We highlighted the findings as Table 4.

Table 4. Comparison of Major United Kingdom-Developed LLMs and Their Characteristics.

Company	LLM Series	Representative Models	Open-Source or Commercial Use	Characteristics
DeepMind	Gopher, Chinchilla	Gopher (280B), Chinchilla (70B)	Research Only	Gopher explored scaling and risks; Chinchilla introduced compute-optimal training; strong impact on efficiency paradigms
Stability AI	StableLM	StableLM Alpha 3B, 7B	Open-Source	Trained on 1.5T tokens; code + text generation; open weights; aligns with open-science and low-barrier deployment principles

6.1. DeepMind's Gopher and Chinchilla (UK): Pioneering Large-Scale and Efficient LLMs

DeepMind, based in London (UK), has been a major contributor to fundamental AI research, including large language models. Although owned by Alphabet, DeepMind operates with a research-driven ethos, and its work on LLMs has significantly influenced the field. Two of its notable LLM projects are Gopher and Chinchilla, which, while not products for direct public use, advanced state-of-the-art and yielded insights now leveraged by many subsequent European efforts. The UK's AI strategy has often highlighted safe and cutting-edge AI research, and DeepMind's exploration of LLM capabilities and limitations aligns with that focus.

Gopher [112] – Scaling to 280B: Announced in late 2021, Gopher was a Transformer-based language model with an enormous 280 billion parameters [113]. At that time, Gopher was one of the largest dense LLMs (surpassing GPT-3's 175B) and was trained on a massive text corpus to study how far scaling could push performance. DeepMind's researchers evaluated Gopher on 152 diverse tasks, finding that increasing model size yielded substantial gains in areas like reading comprehension, fact-checking, and toxicity detection [113]. Gopher achieved state-of-the-art results on most benchmarks, demonstrating the promise of ultra-large models [113]. However, the Gopher study was also a comprehensive analysis of risks: the authors examined Gopher's tendencies for bias, its failure modes in logical and mathematical reasoning, and the challenges in knowledge inaccuracies [113]. They highlighted that specific reasoning tasks did not improve beyond a point, indicating limits to pure scaling.

Additionally, DeepMind discussed mitigating toxic or biased outputs in the Gopher paper, reflecting a commitment to responsible AI research [113]. Although Gopher itself was not open-sourced (it remained an internal research model), the publication of its results provided the global community, including European researchers, valuable data on large-scale LLM behavior. This openness in publishing aligns with academic norms. It has influenced policy discussions by illustrating the need to carefully evaluate large models (an ethos mirrored in EU guidelines for AI testing).

Chinchilla [114] – The Efficiency Breakthrough: In 2022, DeepMind followed up with Chinchilla, a 70B-parameter model that became famous for redefining the scaling paradigm. The premise of Chinchilla was outlined in the paper "Training Compute-Optimal Large Language Models" by Hoffmann et al. [115]. Instead of making models larger, the researchers [115] asked: given a fixed compute budget, what is the optimal way to spend it on model size vs. training data? They concluded that many existing LLMs were underdigitized (under-trained) for their size [115]. Their experiments suggested an approximately equal scaling of model size and training tokens for optimal use of compute [115]. Chinchilla was the instantiation of this idea: using the same compute as was used for Gopher, but with a smaller model (70B) trained on 4× more data (about 1.4 trillion tokens) [115]. The result was striking – Chinchilla outperformed Gopher (280B) on a wide range of tasks, as well as outperforming other large models like OpenAI's GPT-3, Microsoft/NVIDIA's Megatron-Turing (530B), and AI21's Jurassic-1 (178B). For instance, Chinchilla achieved 67.5% accuracy on the MMLU academic exam benchmark, more than 7% higher than Gopher's score. This showed that a mid-sized model with sufficient training could beat much larger models – a crucial insight for any organization (especially those with limited resources) planning to train LLMs.

Impact on European efforts: The Chinchilla findings influenced many subsequent projects; for example, Aleph Alpha's Luminous 70B model leveraged the idea of extensive training to rival larger models [101], and projects like OpenGPT-X opted to start with a 7B well-trained model and scale gradually. Efficiency is also an ethical consideration: Chinchilla's strategy means less energy and carbon emissions for the same performance, aligning with Europe's focus on sustainable AI. The Chinchilla paper notes the advantage of smaller models for downstream use due to lower inference costs [115]. This resonates with European initiatives to reduce the environmental footprint of AI (e.g., the French research on the carbon footprint of LLMs).

Regarding safety and responsible AI, DeepMind's work on LLMs also emphasizes safe deployments. While Gopher and Chinchilla themselves were not released for public use, DeepMind used them to study and publish model bias, toxic language avoidance, and evaluation of harmful content generation [113]. They also explored techniques like reinforcement learning from human feedback (RLHF) in follow-up models (e.g., Sparrow, not detailed here) to align LLMs more with human preferences. These research outputs fed into global discussions in AI ethics. The UK, which has AI governance initiatives (the 2023 UK AI White Paper emphasizes innovation with safeguards), benefits from DeepMind's research leadership in this domain. Indeed, DeepMind's findings have likely informed EU and UK policymakers about the capabilities and risks of frontier models – underscoring why transparency and evaluation (as mandated in the EU AI Act for foundation models) are vital.

In summary, Gopher and Chinchilla are exemplars of the UK's contribution to LLM development: Gopher pushed the boundaries of scale and provided extensive analysis of a model's behavior at that scale [113], while Chinchilla introduced a new paradigm of compute-optimal training that has made LLM development more attainable and efficient [115]. European and UK-aligned teams have built on these insights to create more accessible models instead of mindlessly chasing model size. Though not open-source, DeepMind's academic publications (in venues like NeurIPS 2022 for Chinchilla) ensured that the broader AI community, including European researchers, could learn and benefit from these innovations. This knowledge-sharing reflects an open science spirit and contributes to responsible AI development globally.

6.2. Stability AI's Open-Source LLMs (UK): Democratizing Access

Stability AI [116], a startup headquartered in the UK, became widely known for open-sourcing image generative models (Stable Diffusion) [117]. In 2023, it expanded its mission to language models with the release of StableLM, aiming to provide an open-source foundation that developers and organizations can use freely. Stability's CEO has been vocal about "AI by the people, for the people", an ethos strongly resonant with the open-source community and complementary to European ideals of technological democratization. The effort also ties into the UK's role as a hub for AI startups and open research (London, in particular, hosts communities like EleutherAI and LAION with similar philosophies). Stability AI collaborated with some of these communities in developing its LLMs, leveraging public datasets and volunteer efforts.

Regarding model details and training, the initial release, StableLM Alpha (April 2023), included model sizes of 3 billion and 7 billion parameters [118]. These models were trained on 1.5 trillion tokens of text – a substantial training corpus for their size [118]. The dataset was an expanded version of The Pile (an open 800B-token text corpus from EleutherAI) with three times more data, including diverse web text, code, and multilingual content [118]. By using such a large token count (comparable to what one might use for a 100B+ model) on a smaller model, Stability AI followed the Chinchilla-style strategy to maximize performance. Indeed, StableLM's training approach was informed by the latest research: a balance of English and some non-English data and a significant portion of code data to bolster coding abilities. The 7B StableLM was reported to be proficient in generating both text and programming code, making it a general-purpose foundation model. Notably, Stability AI trained these models on its compute cluster with partners. However, details were not fully public, and the compute budget was likely substantial (on the order of many GPU-months) given the 1.5T token dataset.

Alignment with open science and policy, Stability AI's open-source LLM efforts directly contribute to the democratization of AI, a concept valued in European AI policy discourse. In an official statement, Stability AI asserted that "with the launch of StableLM, Stability AI is continuing to make foundational AI technology accessible to all… Models like StableLM demonstrate how small and efficient models can deliver high performance with appropriate training… and we want everyone to have a voice in their design" [118]. This ethos echoes the European Commission's encouragement of open platforms and the involvement of diverse stakeholders in AI development. Additionally, open models like StableLM increase transparency. Anyone can inspect the weights or analyze outputs, which helps identify biases or undesirable behavior. This transparency aligns with the EU AI Act's emphasis on transparency for foundation models. Indeed, an open model can more easily produce the documentation (datasheets, model cards) required by the Act, since the training process is not a proprietary secret. Stability AI also put a model safety team in place and released an AI safety disclosure alongside StableLM, acknowledging areas like potential bias or misuse and inviting feedback – a practice aligning with responsible innovation principles championed in Europe.

Regarding use cases, StableLM and its successors are building blocks for various use cases, from assisting in coding (the model can autocomplete code or explain code in natural language) to content generation and summarization. In Europe, small and medium enterprises (SMEs) that may be hesitant to send data to an API overseas can fine-tune StableLM on their data and create custom chatbots or assistants. The accessibility of an open model reduces barriers to AI adoption, potentially boosting innovation in local languages or niche domains that big players might not prioritize. For example, a startup in Spain could fine-tune StableLM on Spanish legal texts to build a legal assistant without needing to train a model from scratch or rely on an English-centric model from abroad. This empowerment of local innovation is a key point of emphasis in European digital strategy.

In summary, Stability AI's StableLM initiative plays a significant role in the ecosystem by providing open, transparent, and adaptable LLMs. As a UK-based company engaging a global open-source community, Stability bridges the gap between cutting-edge model research and the public. Their work reinforces the idea that responsible AI is furthered by openness – when more eyes can examine and improve a model, issues can be identified and addressed faster. This community-driven model development complements governmental efforts (like OpenGPT-X) and provides a check against the concentration of AI capabilities. It underscores the European and UK-aligned stance that the foundation of our "digital economy" should not lie exclusively in closed models from a few companies but in part in open models that anyone can use and scrutinize [118].

7. India

India has initiated several projects to develop large language models (LLMs) that cater to its diverse linguistic landscape. This analysis surveys key India-originated LLMs – BharatGPT, Airavata, IndicGPT, IndicBERT, Sanchay, and others – detailing their architectures, training data, language coverage, openness, goals, and institutional origins. Each model is discussed in turn, emphasizing technical specifics and published references. We highlighted the findings as Table 5.

Table 5. Comparison of Major India-Developed LLMs and Their Characteristics**.**

Company	LLM Series	Representative Models	Open-Source or Commercial Use	Characteristics
CoRover.ai	BharatGPT	BharatGPT 3B	Open-Source	Multilingual; 12+ Indian languages; optimized for edge/offline deployment; deployed in real-world apps like AskDisha
AI4Bharat (IIT Madras)	Airavata	Airavata v0.1	Open-Source	Hindi instruction-following model; fine-tuned from OpenHathi (LLaMA-2 based); licensed for reuse
AI4Bharat	IndicGPT	IndicGPT (GPT-2 based)	Open-Source	GPT-2 decoder trained on Indic texts; supports Hindi, Tamil, Bengali, Telugu; strong open-ended generation
AI4Bharat	IndicBERT	IndicBERT, IndicBERT v2	Open-Source	ALBERT-based encoder model; multilingual (12 Indic languages); high efficiency and performance on NLU tasks

7.1. BharatGPT (CoRover)

BharatGPT is a multilingual generative LLM developed by the Indian AI startup CoRover.ai as "India's answer" to existing chatbots [119]. Technically, BharatGPT uses a Transformer decoder architecture with around 3 billion parameters. The model was fine-tuned on a diverse conversational corpus in 12 languages, primarily Indian languages such as Hindi, Punjabi, Marathi, Malayalam, Odia, Kannada, Gujarati, Bengali, Urdu, Tamil, Telugu, as well as English [120]. The training data comprises authentic Indian dialogues and content, allowing BharatGPT to handle tasks like multilingual question-answering, summarization, and translation across these languages.

The model sets new standards in efficient inference on Indian languages – it is lightweight and quantized (GGUF format) for offline or edge deployment [120]. This design enables use in low-resource environments (e.g. on-device or rural internet settings). CoRover has open-sourced BharatGPT (3B Indic version) on HuggingFace [121], making it accessible for public use and fine-tuning. The strategic goal of BharatGPT is to provide a sovereign, home-grown AI platform for India, supporting chatbots and virtual assistants in local languages for banking, e-governance, customer service, etc. [120].

Indeed, BharatGPT has already been deployed in over 100 live applications powering chat services (such as the Indian Railways' AskDisha chatbot) and serving 1.3 billion+ users in 22 languages via text or voice interfaces [122]. CoRover's vision, aligned with "AI for Bharat," is to foster data sovereignty and tailor AI to Indian cultural-linguistic nuances [120].

7.2. Airavata (Hindi LLM by AI4Bharat)

Airavata [123] (sometimes informally called "Airavat") is an instruction-tuned Hindi large language model released in 2024 by AI4Bharat, IIT Madras's open-source AI initiative [124]. Airavata's architecture is based on Meta's LLaMA-2 (7B) Transformer model, which serves as the foundation. Specifically, Airavata was created by fine-tuning OpenHathi, a Hindi-centric LLaMA-2 extension, with supervised instruction datasets [125]. OpenHathi is a 7-billion-parameter Hindi foundational model developed by Sarvam AI – it extends LLaMA-2's vocabulary and was shown to achieve performance comparable to GPT-3.5 on Hindi tasks [125].

Building on this base, Airavata was tuned for Hindi instruction following. The fine-tuning used a diverse set of Hindi instruction data (termed IndicInstruct): for example, ~27k how-to articles from WikiHow were translated to Hindi, and a crowdsourced set of prompts (AnuDEsh) with expert model-generated responses [125]. Notably, the team avoided using proprietary model outputs (like GPT-4) for training, instead relying on human-curated and license-friendly data to ensure open usage. The resulting model, Airavata v0.1, can understand and generate Hindi responses for various queries, achieving strong results on Hindi benchmarks for both NLU and NLG tasks [124]. Airavata is open-sourced (CC BY license) – the model weights and the compiled instruction dataset have been released openly to encourage further research [124].

7.3. IndicGPT (AI4Bharat IIT Madras)

IndicGPT [126] refers to a GPT-style generative language model optimized for multiple Indian languages. This model was developed by researchers at IIT Madras (AI4Bharat) [127] as part of efforts around the national Bhashini mission.

Technically, IndicGPT is built on the GPT-2 architecture – a Transformer decoder – which has been fine-tuned on a corpus of Indian language texts [128]. The model reportedly uses the largest GPT-2 variant (on the order of 1.5 billion parameters) adapted to Indian languages. Languages covered include at least Hindi, Tamil, Bengali, and Telugu (representing major language families in India) [128]. The training data is a curated corpus of Indic-language content, including books, articles, and other text sources in these languages [128].

By fine-tuning GPT-2 on this multilingual collection (and likely extending its vocabulary for Indic scripts), the team enabled IndicGPT to generate fluent text in these languages. Its open-ended text generation performance is reportedly competitive with other Indic-specific models on tasks like story or paragraph generation [128]. For example, it can continue a prompt in Hindi or Tamil with contextually relevant text comparable in quality to models like multilingual GPT-Neo or others focusing on those languages.

7.4. IndicBERT (AI4Bharat Multilingual ALBERT)

IndicBERT [129] is a pioneering multilingual language model for Indic languages released in 2020 by AI4Bharat (IIT Madras) [130]. Unlike the generative models above, IndicBERT is an encoder-only Transformer (BERT-family) intended for understanding tasks (NLU). It uses the ALBERT architecture – a parameter-efficient variant of BERT – which drastically reduces model size via factorized embeddings and cross-layer parameter sharing [130]. IndicBERT has roughly 35 million parameters (an order of magnitude fewer than mBERT's 170M) while maintaining strong performance [129]. It was pretrained on a massive monolingual text corpus of about 9 billion tokens from 12 languages [129]. These languages include Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and English. Notably, "Indian English" was included to help with code-mixed text understanding. The training data (IndicCorp) was compiled from news, Wikipedia, and other web sources across the languages [130]. By jointly training on multiple Indic languages, IndicBERT learns cross-lingual representations beneficial for low-resource cases.

During the evaluation, IndicBERT achieved on-par or better performance than multilingual BERT (mBERT) and XLM-R on the IndicGLUE benchmark and various tasks [129]. For example, in classification tasks like news headline prediction or named entity recognition in Indian languages, IndicBERT's accuracy is comparable to or exceeds mBERT despite having fewer parameters. This efficiency is due to its focused training – dedicating capacity to just 12 languages (as opposed to 100+ in mBERT) and using ALBERT's compression. The model covers Indo-Aryan (e.g. Hindi, Bengali) and Dravidian (e.g. Tamil, Malayalam) families, which are broadly useful in India.

8. Japan

Japan's strategic approach to developing large language models (LLMs) emphasizes linguistic precision, cultural relevance, and open collaboration. This section examines three key contributors to Japan's LLM ecosystem: Rinna, Sakana AI, and the National Institute of Information and Communications Technology (NICT). We highlighted the findings as Table 6.

Table 6. Comparison of Major Japan-Developed LLMs and Their Characteristics**.**

Company	LLM Series	Representative Models	Open-Source or Commercial Use	Characteristics
Rinna	Nekomata	Nekomata 7B, Nekomata 14B	Open-Source (based on Qwen)	Japanese-focused models fine-tuned from Qwen; support for complex grammar; optimized for local NLP tasks
Sakana AI	EvoLLM-JP, EvoVLM-JP	EvoLLM-JP, AI Scientist	Research & Evaluation Stage	Nature-inspired model merging; multimodal and bilingual capabilities; focus on automation and low-resource training
NICT + KDDI	LLM-Japan (LLM-jp)	LLM-jp-13B	In Development / Research	Multimodal models for text + images; national initiative; large-scale training on Japanese web data; hallucination mitigation focus

8.1. Nekomata Series (Rinna)

Rinna Co., Ltd. [131], a Tokyo-based AI company, has been at the forefront of Japanese-language LLM development. Their models are tailored to the nuances of the Japanese language, addressing challenges such as complex grammar and context sensitivity.

One notable contribution is the "Nekomata" series [132], which includes models like Nekomata 7B and 14B. These models are based on Alibaba Cloud's Qwen-7B and Qwen-14B and have been further trained on extensive Japanese corpora. This additional training enhances their performance in Japanese-language tasks and improves inference efficiency [133].

Rinna has also developed bilingual models, such as the Bilingual GPT-NeoX-4B MiniGPT-4 [134], which combines Japanese and English capabilities. This model integrates a vision encoder with a language model, enabling multimodal interaction.

8.2. AI Scientist Project (Sakana AI)

Sakana AI [135], established in Tokyo by former Google researchers David Ha and Llion Jones, adopts a nature-inspired approach to AI development. Their methodology involves "model merging", where existing models are combined and evolved to create new, optimized models. This technique reduces the computational resources typically required for training large models.

Among their innovations is the "AI Scientist," an autonomous system capable of generating research papers with minimal human intervention. While still under evaluation, this system represents a step toward automating scientific research processes [135].

Sakana AI also gained attention for winning innovation awards at the 2025 U.S.-Japan Defense Tech Innovation Challenge, where it presented LLM-based solutions in biodefense and disinformation mitigation [136].

Sakana AI has also introduced the "EvoLLM-JP" and "EvoVLM-JP" models [137], which are designed to handle Japanese-specific content effectively. These models demonstrate the company's commitment to developing AI solutions attuned to Japan's linguistic and cultural context.

8.3. Governmental Research (NICT)

The National Institute of Information and Communications Technology (NICT) [138] is pivotal in Japan's governmental AI research initiatives. In collaboration with KDDI [139], NICT is developing multimodal LLMs to process diverse data types, including text and images. A key focus of this research is to mitigate the issue of hallucinations in LLM outputs, thereby enhancing the reliability of AI-generated information.

In 2023, NICT collaborated with KDDI to launch a joint research project on multimodal LLMs, leveraging over 60 billion web pages for training and focusing on domain-specific accuracy and multi-input understanding [140]. This initiative aligns with the goals of the broader LLM-Japan initiative, which aims to build high-quality, open Japanese LLMs, such as the LLM-jp-13B, using the mdx and ABCI supercomputing infrastructures [141,142].

9. South Korea

South Korea has emerged as a key player in large language model (LLM) development, driven by its major technology companies and a strategic focus on Korean-language optimization and multimodal AI. Despite a relatively small population (~52 million), the country benefits from advanced internet infrastructure and a highly digital-native society, making it an ideal environment for domestic LLM innovation. Major firms, including NAVER, Kakao, and LG have developed high-performing LLMs specifically for Korean applications and integrated them into national AI platforms. We highlighted the findings as Table 7.

Table 7. Comparison of Major South Korea-Developed LLMs and Their Characteristics**.**

Company	LLM Series	Representative Models	Open-Source or Commercial Use	Characteristics
NAVER	HyperCLOVA	HyperCLOVA (204B), HyperCLOVA X	Commercial + API Access	GPT-3 scale; trained on Korean-heavy data; integrated into search, assistants, enterprise plugins
Kakao Brain	Honeybee	Honeybee Multimodal Module	Open-Source (Module)	Adds vision-language capabilities to LLMs; interprets images; extensible for other models
LG AI Research	ExaONE	ExaONE-32B, ExaONE-7.8B	Mixed (Public Access for Smaller Models)	Multimodal reasoning; text + vision; Korean-optimized; smaller variants open for research

9.1. HyperCLOVA (NAVER)

NAVER introduced HyperCLOVA [143] in 2021—a GPT-3-scale model with 204 billion parameters trained on 560 billion tokens of predominantly Korean text, plus multilingual data [144]. It was developed using NAVER’s proprietary 700-petaflop supercomputing infrastructure, reducing reliance on foreign cloud providers.

HyperCLOVA was rapidly integrated into NAVER's products, including its search engine and AI assistants. In 2023, NAVER launched HyperCLOVA X [145], a refined version with a similar scale but improved instruction tuning, stronger Korean reasoning ability, and bilingual support (Korean–English). NAVER also released CLOVA Studio [146] and CLOVA X [147], tools for developers to build applications using the HyperCLOVA backbone [143].

NAVER's strategy highlights vertical integration, combining general-purpose LLMs with domain-specific plug-ins for enterprise use—one of the earliest moves in Asia toward LLM-powered enterprise ecosystems.

9.2. Multimodal Open-Source Innovation (Kakao)

Kakao [148], the company behind Korea's dominant messaging app KakaoTalk, has focused on multimodal AI. In 2024, its research department Kakao Brain open-sourced a Honeybee [149] module, which can equip LLMs with image understanding capabilities. Honeybee allows a language model to perform vision-language tasks—such as interpreting a photo and answering related questions.

9.3. EXAONE (LG AI Research)

LG AI Research [150] developed ExaONE [151], a series of large models targeting reasoning and multimodal capabilities. The flagship ExaONE-32B model reportedly matches the performance of a 670B Mixture-of-Experts model on specific benchmarks [152]. In addition to the 32B version, LG released smaller variants (2.4B and 7.8B) for public use.

These models support both text-only and vision-language tasks and are tuned for Korean-language reasoning. LG's decision to release smaller models openly aligns with a national trend toward accessible, research-friendly AI development [151].

10. Canada

Canada's role in large language model (LLM) development is best characterized by its continued strength in foundational research, multilingual focus, and its innovative startups. While Canada does not host the largest LLMs by scale, it has made notable contributions through talent, research institutes, and developing enterprise and globally inclusive models. We highlighted the findings as Table 8.

Table 8. Comparison of Major Canada-Developed LLMs and Their Characteristics**.**

Company	LLM Series	Representative Models	Open-Source or Commercial Use	Characteristics
Cohere	Command, Aya	Command R+, Aya 101, Aya Expanse 8B/32B	Mixed (Commercial + Open)	Instruction-tuned for enterprise; Aya supports 100+ languages; focus on underrepresented languages and ethical training
Mila / Vector Institute	N/A (Research Contributions)	Contributed to BLOOM, multilingual benchmarks	Research Only	Focus on low-resource adaptation, sustainability, and LLM ethics; key roles in multilingual and open-access initiatives

10.1. Command and Aya Series (Cohere)

Cohere is Canada's leading LLM startup, founded by University of Toronto alumni. It has developed several major models under the Command (English) and Aya (multilingual) series. The Command R+ models [153] have focused on enterprise deployment, combining instruction tuning and reliability for use in business APIs. Their multilingual Aya models [154,155]—particularly Aya 101 (13B) [156] and the later Aya Expanse series (8B [157] and 32B [158])—targeted over 100 global languages, especially underrepresented ones.

Cohere's strategy also emphasizes open science [155]: the Aya Expanse models were open-sourced to help developers fine-tune models in their native languages. This multilingual direction reflects Canada's policy values of inclusivity and its bilingual context (English and French). Cohere's Cohere for AI division has further contributed to global benchmarks and open training pipelines.

10.2. Academic and Research Contributions: Mila, Vector Institute, and More

Canada's academic institutions—particularly Mila (Quebec AI Institute) [159] and the Vector Institute—have played a critical role in research relevant to LLMs. Mila researchers have worked on model efficiency, low-resource adaptation, and reducing environmental impact from training. They were also co-authors of the BLOOM project [95]. Mila has not released a large-scale LLM independently but has shaped global conversations around AI ethics, training efficiency, and multilingualism.

11. Other Notable Countries

Beyond the major players above, several other countries have made significant strides in LLM development, often focusing on regional languages or strategic autonomy. We highlighted the findings as Table 9.

Table 9. Comparison of Other Notable Countries' LLMs and Their Characteristics**.**

Country / Region	LLM Series / Project	Representative Models	Open-Source or Commercial Use	Characteristics
Israel	Jurassic	Jurassic-1 Jumbo, Jurassic-2	Commercial API	Early GPT-3-scale model; strong in English and Hebrew; among first non-U.S. LLMs of this scale
Russia	GigaChat, YaLM	GigaChat, YaLM-100B	Mixed (GigaChat API, YaLM Open)	YaLM-100B was the largest openly licensed model at its time; Russian-centric language focus; national compute use
UAE	Falcon, Jais	Falcon 40B, 180B; Jais 13B/30B	Open-Source	Falcon 40B/180B trained on 1T+ tokens; Jais optimized for Arabic-English tasks; strategic positioning in global AI
Saudi Arabia	Noor	Noor 10B, AraGPT-2	Research Only	Arabic-centric; KAUST and SDAIA involved; early contributions to regional LLM capacity
Australia	(Research & Policy)	No major native LLMs	Policy/Research Contributions	Focus on responsible AI, policy, and inclusion of Indigenous languages; active university involvement
Brazil / LatAm	Bode	Bode 7B, 13B (LLaMA-based)	Open-Source	Portuguese-focused; tailored LLaMA models; increasing local NLP tool development
Africa	InkubaLM	InkubaLM, SafaBERT	Open-Source (Small-scale)	Supports multiple African languages; foundation for future scaling; driven by Lelapa AI and Masakhane community

11.1. AI21 Labs (Israel)

Israel is a hotbed of AI startups and talent. AI21 Labs [160], co-founded by Israeli researchers, launched Jurassic-1 Jumbo (178B) [161,162] in 2021, one of the earliest GPT-3 rivals in size. Jurassic-1 and its successor Jurassic-2 [163] in 2023 have been offered through AI21's API, showcasing strong English and Hebrew capabilities.

11.2. GigaChat (Russia)

Facing technological sanctions, Russia has pushed for home-grown AI solutions. In 2023, Russia's largest bank Sberbank released GigaChat [164], a ChatGPT-like system touted for its Russian language prowess. GigaChat (based on Sber's 13B encoder-decoder model rugpt3.5 and larger MoE experiments) is still in beta and lags behind Western models in maturity. More impressively, in 2022

Russian tech firm Yandex open-sourced YaLM-100B [165], a 100-billion-parameter GPT-like model. Yandex highlighted that YaLM-100B was the largest open model at the time that allowed commercial use, surpassing Meta's open 66B release [166]. This move was aimed at spurring local innovation – Russian researchers and companies can build on YaLM without restrictions. The motivation is partly to reduce reliance on English-centric models and partly to assert technological independence. Russia also has dedicated labs like AIRI (Artificial Intelligence Research Institute) working on LLMs, and uses its national supercomputers (Christofari AI cluster) for training. The primary focus is Russian and related languages (like Ukrainian, which ironically may benefit from these advances too). However, Western export controls on GPUs have made it harder for Russia to scale beyond 100B. Still, by making models like YaLM open, Russia is trying to remain relevant in AI research.

11.3. Falcon (UAE & Saudi Arabia)

The United Arab Emirates has made a strong entrance via the Technology Innovation Institute (TII) in Abu Dhabi. In 2023, TII released Falcon 40B [167], a 40-billion-parameter model trained on 1 trillion tokens of refined web data. Falcon 40B was open-sourced under Apache 2.0, including for commercial use, immediately making it one of the most powerful openly licensed models. Falcon 40B quickly topped some open-model leaderboards (ranked #1 on HuggingFace at one point for raw performance).

Following this, the UAE announced Falcon 180B, which is on track as the next milestone (UAE's TII Launches Open-Source "Falcon 40B" Large Language Model for Research & Commercial Utilization | Technology Innovation Institute). These moves are part of UAE's broader strategy to be a global AI player; they even launched a global call for proposals, offering computing grants to anyone building on Falcon models.

Separately, a collaboration between UAE's G42/Inception and the Mohamed bin Zayed University of AI (MBZUAI) with Cerebras Systems yielded Jais [168], a 13B bilingual Arabic-English model. Jais was trained on a massive Arabic dataset (116B Arabic tokens) plus English data, making it the most advanced Arabic-focused LLM to date. Jais was open-sourced in August 2023 and achieved state-of-the-art on Arabic benchmarks.

Soon after, an improved 30B version was developed, underlining the Middle East's commitment to leadership in AI for Arabic. Saudi Arabia has also initiated efforts: KAUST and Horizon company developed Noor, a 10B Arabic model (mentioned in 2022), and Saudi NLP researchers have built experimental large models (e.g., AraGPT-2 and AraT5 series for Arabic). Saudi Arabia's SDAIA is reportedly investing in larger models to support Arabic translation and government AI services.

11.4. Research and Policy Initiatives by CSIRO Data61 (Australia)

Australia has not developed its large-scale LLMs but is actively engaged in LLM research and policy formulation. The national science agency, CSIRO, through its Data61 division, has highlighted the potential benefits and challenges of adopting AI foundation models tailored to Australian contexts, including considerations for Indigenous languages and cultural nuances [169]. Additionally, Australian universities have contributed to LLM research, particularly in areas like reinforcement learning with LLMs. On the policy front, Australia has developed frameworks to ensure the safe and responsible use of AI, including LLMs, in government and industry.

11.5. Bode (Brazil and Latin America)

In Latin America, particularly Brazil, there is a growing interest in developing LLMs tailored for Portuguese and Spanish. A notable contribution is the development of Bode, a fine-tuned LLaMA 2-based model for Portuguese prompts, available in 7B and 13B parameter versions. This model has demonstrated satisfactory results in Portuguese language tasks and is freely available for research and commercial purposes [170]. Additionally, the nonprofit LAION, while European-led, has collaborators in Latin America who assist in preparing multilingual data that includes Latin

American Spanish dialects. Due to resource constraints, these regions often leverage open models from elsewhere and fine-tune them for local language applications, such as agriculture or call centers.

11.6. InkubaLM (Africa)

In Africa, initiatives are underway to develop LLMs for African languages. A significant development is InkubaLM, introduced by Lelapa AI, Africa's first multilingual AI large language model. InkubaLM supports low-resource African languages, including Swahili, Yoruba, isiXhosa, Hausa, and isiZulu. The model aims to address the digital underrepresentation of these languages by providing tools for translation, transcription, and various natural language processing tasks [171]. While no African country has developed a native LLM exceeding 1 billion parameters, the Masakhane NLP community is actively involved in data collection and model development. Examples include SafaBERT, a BERT-like model for Swahili, and experimental GPT-2-sized models for languages like Zulu and Shona. These efforts are foundational to creating LLMs that understand and process African languages effectively.

12. Discussion

12.1. U.S. LLMs: Technical Leadership Anchored by Scale and Infrastructure

The U.S. continues to lead in LLM development due to early research breakthroughs, access to massive computational resources, and close coupling between private industry and academia. Models such as ChatGPT (OpenAI), Claude (Anthropic), and Gemini (Google) reflect not only architectural sophistication but also advanced strategies for model alignment and deployment.

The U.S. ecosystem benefits from vertically integrated infrastructure—NVIDIA GPUs, cloud services (Azure, AWS, Google Cloud), and commercialization pipelines (APIs, enterprise tools)—which collectively lower the barrier to scale large models rapidly. This self-sufficiency in compute and software stacks provides a structural advantage that few countries can replicate.

12.2. Compute Dominance: GPU Access as a Strategic Enabler

Underlying the U.S. leadership is its near-total control of cutting-edge AI hardware. Based in the U.S., NVIDIA supplies most of the world's high-performance GPUs (A100, H100), which are essential for training and inference at scale. This control extends to software (e.g., CUDA, TensorRT, and NeMo frameworks), creating an integrated supply chain advantage. U.S. firms not only build the models but also dominate the infrastructure on which they run. Due to export restrictions or cost, countries with limited access to such resources face structural constraints that force different strategic choices.

12.3. China: Innovation Under Hardware Constraints

China’s LLM development is marked by rapid model proliferation and strategic adaptation to compute limitations. Due to export controls on advanced GPUs, Chinese firms like DeepSeek, Moonshot AI, and Baichuan have invested in algorithmic efficiency (e.g., Mixture-of-Experts, low-bit precision training) and cost-effective model design. This has spurred innovation in architectures and training paradigms. Several models (e.g., Kimi, PanGu-Σ) demonstrate ultra-long context handling and domain-specific adaptation. The open release of several competitive models under liberal licenses also reflects a shift toward decentralized, developer-driven ecosystems. Overall, China has compensated for hardware disadvantage through algorithmic advances and aggressive openness in some sectors.

12.4. Open Source and Regional AI Sovereignty: Europe and Beyond

The European Union emphasizes transparency, accountability, and multilingual inclusivity. Initiatives such as BLOOM (France), Luminous (Germany), and OpenGPT-X (EU consortium) are driven by public institutions or regulated startups, often trained on national supercomputers with

doi:10.20944/preprints202504.2136.v1

open licenses. These projects aim to provide alternatives to U.S. proprietary models while aligning with local legal and ethical frameworks, including the EU AI Act.

Similarly, efforts in the UAE (Falcon), India (BharatGPT, Airavata), and Brazil (Bode) reflect a global interest in building local models that respect linguistic diversity and data sovereignty. These developments suggest that LLM proliferation is no longer limited to the largest economies, as smaller nations adopt tailored strategies through open-source collaboration, domain specialization, or regional HPC support.

12.5. Open Source vs. Commercial Use

The distinction between open-source and commercially licensed LLMs has become a central theme in the current landscape [35]. While open-source models (e.g., LLaMA, Mistral, DeepSeek) promise greater transparency and accessibility, many commercial models (e.g., GPT-4, Claude, Gemini) focus on deployment stability, compliance, and safety. Notably, benchmark performance does not consistently favor one category over the other—many open models now approach or match proprietary systems on standard tasks. However, the long-term ecosystem impact, including downstream innovation, cost structures, and platform lock-in, remains an open question for future study.

12.6. Multilingualism and Societal Relevance as Strategic Differentiators

While early LLM efforts centered on English, recent models increasingly emphasize language diversity and social relevance. Canadian, Indian, and European models have demonstrated multilingual capabilities (e.g., Aya, IndicGPT, OpenGPT-X), often with specific attention to underrepresented languages. In Asia, models such as HyperCLOVA (Korean), Kimi (Chinese), and Jais (Arabic-English) are optimized for linguistic and cultural contexts. Japan's Nekomata and South Korea's EXAONE also focus on precision in local language tasks. These trends reflect a recognition that language-specific performance and regulatory compliance are equally crucial as the model scales for downstream adoption.

12.7. National Capacity and the Structural Foundations of LLM Advancement

While LLM competition appears commercially driven, it increasingly reflects national-level strategic priorities. Training competitive LLMs requires large-scale compute infrastructure, high-density data pipelines, and a concentrated pool of AI talent—resources that align more closely with national capabilities than individual firms alone. Public investment, export controls, and geopolitical positioning all shape who can build and deploy frontier models. In this sense, LLMs are becoming instruments in a broader context of national technological sovereignty.

12.8. Algorithmic Advances as an Entry Point for Small Teams

Despite the dominance of large firms and nations in LLM development, there remain viable entry points for small research teams and academic labs. Algorithmic breakthroughs—such as improved optimizers, efficient fine-tuning (e.g., LoRA, QLoRA), long-context architectures, or better alignment methods—can have an outsized impact relative to their resource cost. Open research in these areas continues to shape how models are trained and deployed globally. As such, algorithmic innovation represents a critical lever for democratizing LLM progress.

13. Conclusions

This review examined the global development of large language models (LLMs) through a country-centered lens, identifying key patterns in institutional structures, resource allocation, and deployment strategies. While LLMs are often framed as commercial innovations, they increasingly reflected national priorities—from economic competitiveness and digital sovereignty to linguistic inclusivity and AI regulation.

The United States maintained a lead in foundational model development through institutions such as OpenAI (GPT series), Google DeepMind (Gemini, PaLM), Anthropic (Claude), Meta (LLaMA), and NVIDIA (Nemotron), supported by unparalleled access to GPU infrastructure and cloud-scale deployment capacity.

China followed a contrasting yet competitive trajectory: companies such as Baidu (ERNIE), Alibaba (Tongyi Qianwen), Huawei (PanGu), and newer entrants like DeepSeek and Moonshot AI rapidly scaled high-performing bilingual models, often under tight hardware constraints, by innovating on training efficiency and releasing open-source models to encourage domestic adoption.

European efforts, exemplified by BLOOM (France), Luminous (Germany), and OpenGPT-X (EU consortium), were shaped by a commitment to openness, multilingualism, and alignment with emerging AI regulations.

In parallel, countries like India (BharatGPT, Airavata), the UAE (Falcon), and South Korea (HyperCLOVA) invested in sovereign LLM infrastructure adapted to local linguistic and industrial needs.

Canada, Japan, and the UK—through efforts like Cohere (Aya, Command), Rinna (Nekomata), and DeepMind (Chinchilla, Gopher)—contributed models and theoretical insights that advanced both the technical and ethical frontiers of the field.

Across these findings, our analysis found that the distinction between open-source and commercial models had become increasingly fluid. Open models, including DeepSeek, LLaMA 2, BLOOM, and Mistral, frequently matched proprietary systems on key benchmarks. However, while performance parity was often achieved, the broader implications—especially regarding ecosystem resilience, innovation accessibility, and lock-in effects—remain areas for future empirical study.

Significantly, LLM development now operates as a proxy for broader measures of state capacity. Nations differ not only in their ability to produce large models but also in their ability to control supply chains for GPUs, structure data governance, and retain scientific talent. Integrating LLMs into education, government services, and national cloud infrastructure reflects their central role in shaping sovereign digital futures.

At the same time, algorithmic innovation remained essential for academic groups and small teams. Advances in training efficiency (e.g., QLoRA, MoE routing), alignment (e.g., Constitutional AI), and context scaling (e.g., Kimi, Claude, InternLM) showed that contributions to LLM research were not strictly gated by access to massive compute. These developments reinforced the view that creativity in model design, optimization, and deployment strategy can be as decisive as raw scale.

Overall, global LLM development co-evolved with geopolitical, regulatory, and infrastructural forces. Understanding this landscape required not only an analysis of model capabilities but also of governance structures, language needs, and platform strategies. As nations continue to shape—and be shaped by—the rise of LLMs, these systems remain a key domain where questions of openness, equity, and technological power intersect. The long-term trajectory of LLMs will be determined by scientific progress and how different societies choose to embed these technologies within their digital and institutional ecosystems.

References

1. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, 'Large language models in medicine', *Nature medicine*, vol. 29, no. 8, pp. 1930–1940, 2023.
2. W. C. Choi, I. C. Choi, C. I. Chang, and L. C. Lam, 'Comparison of Claude (Sonnet and Opus) and ChatGPT (GPT-4, GPT-4o, GPT-o1) in Analyzing Educational Image-based Questions from Block-Based Programming Assessments', in *2025 14th International Conference on Information and Education Technology (ICIET)*, IEEE, 2025.
3. W. C. Choi and C. I. Chang, 'Enhancing Education with ChatGPT 4o and Microsoft Copilot: A Review of Opportunities, Challenges, and Student Perspectives on LLM-Based Text-to-Image Generation Models', 2025.

4. C. I. Chang, W. C. Choi, and I. C. Choi, ‘Challenges and Limitations of Using Artificial Intelligence Generated Content (AIGC) with ChatGPT in Programming Curriculum: A Systematic Literature Review’, in *Proceedings of the 2024 7th Artificial Intelligence and Cloud Computing Conference*, 2024.
5. C. I. Chang, W. C. Choi, and I. C. Choi, ‘A Systematic Literature Review of the Opportunities and Advantages for AIGC (OpenAI ChatGPT, Copilot, Codex) in Programming Course’, in *Proceedings of the 2024 7th International Conference on Big Data and Education*, 2024.
6. W. C. Choi, J. Peng, I. C. Choi, H. Lei, L. C. Lam, and C. I. Chang, ‘Improving Young Learners with Copilot: The Influence of Large Language Models (LLMs) on Cognitive Load and Self-Efficacy in K-12 Programming Education’, in *Proceedings of the 2025 International Conference on Artificial Intelligence and Education (ICAIE)*, Suzhou, China, 2025.
7. R. Aghaei et al., ‘Harnessing the Potential of Large Language Models in Modern Marketing Management: Applications, Future Directions, and Strategic Recommendations’, *arXiv preprint arXiv:2501.10685*, 2025.
8. P. Homoki and Z. Z\Hodi, ‘Large language models and their possible uses in law’, *Hungarian Journal of Legal Studies*, vol. 64, no. 3, pp. 435–455, 2024.
9. C. Culver, P. Hicks, M. Milenkovic, S. Shanmugavelu, and T. Becker, ‘Scientific computing with large language models’, *arXiv preprint arXiv:2406.07259*, 2024.
10. Reuters, ‘China’s Tencent debuts large language AI model, says open for enterprise use’. [Online]. Available: https://www.reuters.com/technology/chinas-tencent-says-large-language-ai-model-hunyuan-available-enterprise-use-2023-09-07/
11. BigScience, ‘BigScience Large Open-science Open-access Multilingual Language Model’. 2025. [Online]. Available: https://huggingface.co/bigscience/bloom
12. OpenAI, ‘OpenAI’. 2025. [Online]. Available: https://openai.com/
13. L. Ouyang et al., ‘Training language models to follow instructions with human feedback’, *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.
14. J. Achiam et al., ‘Gpt-4 technical report’, *arXiv preprint arXiv:2303.08774*, 2023.
15. Anthropic, ‘Anthropic’. [Online]. Available: https://www.anthropic.com/
16. Anthropic, ‘Claude’. 2025. [Online]. Available: https://claude.ai
17. Google, ‘Google DeepMind’. 2025. [Online]. Available: https://deepmind.google/
18. A. Chowdhery et al., ‘Palm: Scaling language modeling with pathways’, *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.
19. R. Anil et al., ‘Palm 2 technical report’, *arXiv preprint arXiv:2305.10403*, 2023.
20. G. Team et al., ‘Gemini: a family of highly capable multimodal models’, *arXiv preprint arXiv:2312.11805*, 2023.
21. Meta, ‘Meta AI’. 2025. [Online]. Available: https://www.meta.ai/
22. H. Touvron et al., ‘Llama: Open and efficient foundation language models’, *arXiv preprint arXiv:2302.13971*, 2023.
23. Microsoft, ‘Microsoft AI’. 2025. [Online]. Available: https://microsoft.ai/
24. Microsoft, ‘Microsoft Azure: Cloud Computing Services’. 2025. [Online]. Available: https://azure.microsoft.com/
25. J. Stratton, ‘An introduction to microsoft copilot’, in *Copilot for Microsoft 365: Harness the Power of Generative AI in the Microsoft Apps You Use Every Day*, Springer, 2024, pp. 19–35.
26. xAI, ‘xAI’. 2025. [Online]. Available: https://x.ai/
27. xAI, ‘Grok 3 Beta — The Age of Reasoning Agents’. [Online]. Available: https://x.ai/news/grok-3
28. Amazon Web Services, ‘Artificial Intelligence (AI) on AWS - AI Technology’. 2025. [Online]. Available: https://aws.amazon.com/ai/
29. Amazon Web Services, ‘Amazon Titan Foundation Models’. [Online]. Available: https://aws.amazon.com/tw/bedrock/amazon-models/titan/
30. Amazon Web Services, ‘Amazon Nova Foundation Models’. [Online]. Available: https://aws.amazon.com/ai/generative-ai/nova/
31. NVIDIA, ‘NVIDIA Llama Nemotron’. [Online]. Available: https://www.nvidia.com/en-us/ai-data-science/foundation-models/llama-nemotron/

32. Vespa, 'We Make AI Work at Perplexity – Delivering AI-Powered Search at Scale'. [Online]. Available: https://vespa.ai/perplexity/
33. Rest of World, 'DeepSeek and chip bans have supercharged AI innovation in China'. [Online]. Available: https://restofworld.org/2025/china-ai-boom-chip-ban-deepseek/
34. DeepSeek, 'DeepSeek'. 2025. [Online]. Available: https://www.deepseek.com/
35. W. C. Choi and C. I. Chang, 'Advantages and Limitations of Open-Source Versus Commercial Large Language Models (LLMs): A Comparative Study of DeepSeek and OpenAI's ChatGPT', 2025.
36. Wikipedia, 'DeepSeek'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/DeepSeek
37. Nature News, 'China's cheap, open AI model DeepSeek thrills scientists', *Nature*, 2025, doi: 10.1038/d41586-025-00229-6.
38. X. Bi et al., 'Deepseek llm: Scaling open-source language models with longtermism', *arXiv preprint arXiv:2401.02954*, 2024.
39. DeepSeek, 'deepseek-ai (DeepSeek)'. 2025. [Online]. Available: https://huggingface.co/deepseek-ai
40. Reuters, 'China's DeepSeek releases AI model upgrade, intensifies rivalry with OpenAI'. [Online]. Available: https://www.reuters.com/technology/artificial-intelligence/chinas-deepseek-releases-ai-model-upgrade-intensifies-rivalry-with-openai-2025-03-25/
41. Moonshot AI, 'Kimi - Moonshot AI'. 2025. [Online]. Available: https://kimi.moonshot.cn/
42. Global Times, 'China's Moonshot AI fuels domestic large model app frenzy, aiming to overtake ChatGPT'. [Online]. Available: https://www.globaltimes.cn/page/202403/1309421.shtml
43. South China Morning Post, 'Alibaba-backed Moonshot AI claims breakthrough in expanded Chinese-character prompt for Kimi chatbot'. [Online]. Available: https://www.scmp.com/tech/big-tech/article/3256109/alibaba-backed-moonshot-ai-claims-breakthrough-expanded-chinese-character-prompt-kimi-chatbot
44. South China Morning Post, 'Chinese AI start-up Moonshot cuts LLM feature price amid fierce domestic competition'. [Online]. Available: https://www.scmp.com/tech/tech-trends/article/3273619/chinese-ai-start-moonshot-cuts-llm-feature-price-amid-fierce-domestic-competition
45. K. Team et al., 'Kimi k1. 5: Scaling reinforcement learning with llms', *arXiv preprint arXiv:2501.12599*, 2025.
46. South China Morning Post, 'Chinese AI start-up MiniMax releases low-cost open-source models that rival top chatbots'. [Online]. Available: https://www.scmp.com/tech/big-tech/article/3294900/chinese-ai-start-minimax-releases-low-cost-open-source-models-rival-top-chatbots
47. Wikipedia, 'ERNIE Bot'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/Ernie_Bot
48. Y. Sun et al., 'Ernie: Enhanced representation through knowledge integration', *arXiv preprint arXiv:1904.09223*, 2019.
49. A. Velinov, 'Chinese LLMs vs Western LLMs - Developments, Comparisons, and Global Outlook'. [Online]. Available: https://www.linkedin.com/pulse/chinese-llms-vs-western-developments-comparisons-global-velinov-sqeyf/
50. Reuters, 'Alibaba upgrades AI model Tongyi Qianwen, releases industry-specific models'. [Online]. Available: https://www.reuters.com/technology/alibaba-upgrades-ai-model-tongyi-qianwen-releases-industry-specific-models-2023-10-31/
51. Alibaba Cloud, 'Tongyi Qianwen - Alibaba Cloud'. 2025. [Online]. Available: https://tongyi.aliyun.com/
52. UC Today, 'What is Tongyi Qianwen? Alibaba's ChatGPT Rival'. [Online]. Available: https://www.uctoday.com/collaboration/what-is-tongyi-qianwen-alibabas-chatgpt-rival/
53. CNBC, 'Alibaba opens AI model Tongyi Qianwen to public in competition with Baidu, SenseTime'. [Online]. Available: https://www.cnbc.com/2023/08/30/alibaba-opens-ai-model-tongyi-qianwen-to-public.html
54. R. Zheng et al., 'Secrets of rlhf in large language models part i: Ppo', *arXiv preprint arXiv:2307.04964*, 2023.
55. ByteDance, 'ByteDance'. 2025. [Online]. Available: https://www.bytedance.com/
56. ByteDance, 'Doubao AI Tool'. 2025. [Online]. Available: https://www.doubao.com/chat/
57. China Internet Watch, 'ByteDance's AI Surge: Introducing the Doubao Model Family'. [Online]. Available: https://www.ciw.news/p/bytedance-ai-surge-doubao

58. Wikipedia, 'Doubao chatbot'. 2025. [Online]. Available: https://zh.wikipedia.org/wiki/%E8%BE%93%E5%8C%85_(%E8%81%8A%E5%A4%A9%E6%9C%BA%E5%99%A8%E4%BA%BA)
59. South China Morning Post, 'TikTok owner ByteDance launches low-cost Doubao AI models for enterprises, initiating a price war in crowded mainland market'. [Online]. Available: https://sc.mp/2h1kd?utm_source=copy-link&utm_campaign=3262781&utm_medium=share_widget
60. TMTPost, 'Alibaba and ByteDance Intensify AI Race Despite Rule Tightening'. [Online]. Available: https://en.tmtpost.com/post/6485054
61. X. Ren et al., 'Pangu-Sigma: Towards trillion parameter language model with sparse heterogeneous computing', *arXiv preprint arXiv:2303.10845*, 2023.
62. Wikipedia, 'Huawei PanGu'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/Huawei_PanGu
63. Counterpoint Research, 'Huawei Expands its AI Ambition With Pangu Large Models'. [Online]. Available: https://www.counterpointresearch.com/insights/huawei-expands-its-ai-ambition-with-pangu-large-models/
64. Huawei, 'AI for Good at HAS 2024: GenAI for Modern Drug Discovery'. [Online]. Available: https://www.huawei.com/en/huaweitech/industry-trends/has24-takeaways-pangu-omdia
65. 1AI, 'The Chinese Academy of Sciences' self-developed AI model "Zidong Taichu 3.0" was released in the first half of this year to optimize intelligent driving training'. [Online]. Available: https://www.1ai.net/en/4953.html
66. Tencent, 'Tencent Hunyuan'. 2025. [Online]. Available: https://hunyuan.tencent.com/
67. X. Sun et al., 'Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent', *arXiv preprint arXiv:2411.02265*, 2024.
68. Analytics India Magazine, 'Tencent Launches Hunyuan Large, Outperforms Llama 3.1 70B & 405B'. [Online]. Available: https://analyticsindiamag.com/ai-news-updates/tencent-launches-hunyuan-large-outperforms-llama-3-1-70b-405b/
69. iFLYTEK, 'iFLYTEK Spark Model'. 2025. [Online]. Available: https://xinghuo.xfyun.cn/
70. Baidu Baike, 'iFLYTEK Spark Cognitive Model'. 2025. [Online]. Available: https://baike.baidu.com/item/%E8%AE%AF%E9%A3%9E%E6%98%9F%E7%81%AB%E8%AE%A4%E7%9F%A5%E5%A4%A7%E6%A8%A1%E5%9E%8B/62912838
71. Live Science, 'New Chinese AI model "better than industry leader" in key metrics'. [Online]. Available: https://www.livescience.com/technology/artificial-intelligence/chinese-ai-model-spark-35-better-than-open-ai-gpt4
72. China Daily, 'iFlytek unveils upgraded LLM that "outperforms GPT-4 Turbo"'. [Online]. Available: https://www.chinadaily.com.cn/a/202406/28/WS667e67f6a31095c51c50b617.html
73. Wikipedia, 'Baichuan'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/Baichuan
74. AI Fun, 'Baichuan Intelligence'. [Online]. Available: https://www.aifun.cc/en/sites/baichuanzhineng.html
75. A. Yang et al., 'Baichuan 2: Open large-scale language models', *arXiv preprint arXiv:2309.10305*, 2023.
76. AI Fun, 'Bai Xiaoying'. [Online]. Available: https://www.aifun.cc/en/sites/baixiaoying.html
77. Zhipu AI, 'ChatGLM'. 2025. [Online]. Available: https://chatglm.cn/
78. SiliconANGLE, 'Chinese AI startup Zhipu raises $300M in funding'. [Online]. Available: https://siliconangle.com/2023/10/20/chinese-ai-startup-zhipu-raises-300m-funding/
79. Nature News, 'China's ChatGPT: why China is building its own AI chatbots', *Nature*, 2024, doi: 10.1038/d41586-024-01495-6.
80. T. GLM et al., 'Chatglm: A family of large language models from glm-130b to glm-4 all tools', *arXiv preprint arXiv:2406.12793*, 2024.
81. C. Liu et al., 'CPMI-ChatGLM: parameter-efficient fine-tuning ChatGLM with Chinese patent medicine instructions', *Scientific reports*, vol. 14, no. 1, p. 6403, 2024.
82. H. Zhou et al., 'GLMLog: Log Anomaly Detection Method Based on ChatGLM', in *2024 10th International Conference on Computer and Communications (ICCC)*, IEEE, 2024, pp. 1923–1927.
83. Institute of Automation, Chinese Academy of Sciences, 'Zi Dong Tai Chu'. 2025. [Online]. Available: https://taichu-web.ia.ac.cn/

84. Baidu Baike, 'Zi Dong Tai Chu'. 2025. [Online]. Available: https://baike.baidu.com/item/%E7%B4%AB%E4%B8%9C%E5%A4%AA%E5%88%9D/57969205
85. Chinese Academy of Sciences, 'Chinese Academy of Sciences Launches Its Next-generation AI Model'. [Online]. Available: https://english.cas.cn/newsroom/cas_media/202306/t20230620_332105.shtml
86. I. Team, 'Internlm: A multilingual language model with progressively enhanced capabilities'. 2023.
87. The Decoder, 'With "InternLM", China enters the race for large language models'. [Online]. Available: https://the-decoder.com/with-internlm-china-enters-the-race-for-large-language-models/
88. China Daily, 'SenseTime's large language model service open for registration'. [Online]. Available: https://global.chinadaily.com.cn/a/202308/31/WS64eff802a31035260b81f363.html
89. InternLM Team, 'InternLM/InternLM: Official release of InternLM series'. 2025. [Online]. Available: https://github.com/InternLM/InternLM
90. Yicai Global, 'SenseTime, Shanghai AI Lab Debut New-Generation LLM'. [Online]. Available: https://www.yicaiglobal.com/news/chinas-sensetime-three-other-institutes-debut-new-gen-llm
91. MiniMax, 'MiniMax'. 2025. [Online]. Available: https://www.minimaxi.com/
92. Wikipedia, 'MiniMax'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/MiniMax_(company)
93. Wikipedia contributors, 'BLOOM (language model)'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/BLOOM_(language_model)
94. BigScience, 'Introducing The World's Largest Open Multilingual Language Model - BLOOM'. 2025. [Online]. Available: https://bigscience.huggingface.co/
95. T. Le Scao et al., 'Bloom: A 176b-parameter open-access multilingual language model', 2023.
96. Aleph Alpha, 'Aleph Alpha'. 2025. [Online]. Available: https://aleph-alpha.com/
97. Aleph Alpha, 'Luminous'. 2025. [Online]. Available: https://www.luminous.com/
98. Wikipedia, 'Aleph Alpha'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/Aleph_Alpha
99. adesso, 'Quickstart with a European-based large language model: Aleph Alpha's "Luminous"'. [Online]. Available: https://www.adesso.de/en/news/blog/quickstart-with-a-european-based-large-language-model-aleph-alphas-luminous.jsp
100. Aleph Alpha, 'Aleph Alpha Luminous Benchmarks'. 2025. [Online]. Available: https://aleph-alpha.com/wp-content/uploads/Performance-Report_Luminous_Aleph-Alpha.pdf
101. Aleph Alpha, 'Luminous: European AI closes gap to world leaders'. [Online]. Available: https://aleph-alpha.com/luminous-european-ai-closes-gap-to-world-leaders/
102. Techzine Europe, 'Aleph Alpha is releasing two open-source models fully compliant with the European AI Act'. [Online]. Available: https://www.techzine.eu/blogs/privacy-compliance/123863/aleph-alphas-open-source-llms-fully-comply-with-the-ai-act/
103. Mistral AI, 'Mistral AI'. 2025. [Online]. Available: https://mistral.ai/
104. A. Q. Jiang et al., 'Mistral 7B', 2023. [Online]. Available: https://arxiv.org/abs/2310.06825
105. Mistral AI, 'Announcing Mistral 7B'. [Online]. Available: https://mistral.ai/news/announcing-mistral-7b
106. Mistral AI, 'Mistral NeMo'. [Online]. Available: https://mistral.ai/news/mistral-nemo
107. Mistral AI, 'Models Overview'. 2025. [Online]. Available: https://docs.mistral.ai/getting-started/models/models_overview/
108. OpenGPT-X Consortium, 'OpenGPT-X, Multilingual. Open. European'. 2025. [Online]. Available: https://opengpt-x.de/
109. A. Herten and S. Kesselheim, 'Multilingual and Open Source: OpenGPT-X Releases Large Language Model'. [Online]. Available: https://www.fz-juelich.de/en/news/archive/announcements/2024/multilingual-and-open-source-opengpt-x-releases-large-language-model
110. The Register, 'Germany to host Europe's first exascale supercomputer'. [Online]. Available: https://www.theregister.com/2022/06/16/first_european_exascale/
111. LUMI Supercomputer, 'Open LLMs for transparent AI in Europe'. [Online]. Available: https://www.lumi-supercomputer.eu/open-euro-llm/

112. DeepMind, 'Language modelling at scale: Gopher, ethical considerations, and retrieval'. [Online]. Available: https://deepmind.google/discover/blog/language-modelling-at-scale-gopher-ethical-considerations-and-retrieval/
113. J. W. Rae et al., 'Scaling language models: Methods, analysis & insights from training gopher', *arXiv preprint arXiv:2112.11446*, 2021.
114. DeepMind, 'An empirical analysis of compute-optimal large language model training'. [Online]. Available: https://deepmind.google/discover/blog/an-empirical-analysis-of-compute-optimal-large-language-model-training/
115. J. Hoffmann et al., 'Training compute-optimal large language models', *arXiv preprint arXiv:2203.15556*, 2022.
116. Stability AI, 'Stability AI'. 2025. [Online]. Available: https://stability.ai/
117. Wikipedia, 'Stability AI'. 2025. [Online]. Available: https://en.wikipedia.org/wiki/Stability_AI
118. A. Alford, 'Stability AI Open-Sources 7B Parameter Language Model StableLM'. [Online]. Available: https://www.infoq.com/news/2023/05/stablelm-release/
119. CoRover Pvt. Ltd, 'BharatGPT'. [Online]. Available: https://corover.ai/bharatgpt/
120. Analytics India Magazine, 'CoRover.ai's BharatGPT Surpasses 2,000 Downloads on Hugging Face Within Days'. [Online]. Available: https://analyticsindiamag.com/ai-news-updates/corover-ais-bharatgpt-surpasses-2000-downloads-on-hugging-face-within-days/
121. CoRover, 'CoRover/BharatGPT-3B-Indic'. [Online]. Available: https://huggingface.co/CoRover/BharatGPT-3B-Indic
122. Computer Weekly, 'Indian large language models gain momentum'. [Online]. Available: https://www.computerweekly.com/news/366575652/Indian-large-language-models-gain-momentum
123. AI4Bharat, 'ai4bharat/Airavata'. [Online]. Available: https://huggingface.co/ai4bharat/Airavata
124. J. Gala et al., 'Airavata: Introducing hindi instruction-tuned llm', *arXiv preprint arXiv:2401.15006*, 2024.
125. The Economic Times, 'AI4Bharat releases Hindi LLM "Airavata"'. [Online]. Available: https://economictimes.indiatimes.com/tech/technology/ai4bharat-releases-hindi-llm-airavata/articleshow/107146024.cms
126. IndicGPT, 'aashay96/indic-gpt'. 2025. [Online]. Available: https://huggingface.co/aashay96/indic-gpt
127. AI4Bharat, 'AI4Bharat'. [Online]. Available: https://ai4bharat.iitm.ac.in/
128. Brijeshkumar Y. Panchal, 'Examine the Opportunities and Challenges of Large Language Model (LLM) For Indic Languages', *jisem*, vol. 10, no. 26s, pp. 301–325, Mar. 2025, doi: 10.52783/jisem.v10i26s.4236.
129. AI4Bharat, 'ai4bharat/indic-bert'. [Online]. Available: https://huggingface.co/ai4bharat/indic-bert
130. D. Kakwani et al., 'IndicNLPSuite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for Indian languages', in *Findings of the association for computational linguistics: EMNLP 2020*, 2020, pp. 4948–4961.
131. Rinna Co., Ltd, 'Rinna Co., Ltd'. 2025. [Online]. Available: https://rinna.co.jp/global/
132. Rinna Co., Ltd, 'Rinna Releases "Nekomata" Series: Japanese Continued Pre-trained Models Based on Qwen'. [Online]. Available: https://rinna.co.jp/news/2023/12/20231221.html
133. Rinna Co., Ltd., 'Japanese-Language AI Models Based on Tongyi Qianwen (Qwen) Were Launched by rinna'. Feb. 2023. [Online]. Available: https://www.alibabacloud.com/blog/rinna-launched-ai-models-trained-in-the-japanese-language—nekomata-series—based-on-alibaba-clouds-qwen-models_600719
134. Rinna Co., Ltd., 'rinna/bilingual-gpt-neox-4b-minigpt4'. 2023. [Online]. Available: https://huggingface.co/rinna/bilingual-gpt-neox-4b-minigpt4
135. Sakana AI, 'The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery'. [Online]. Available: https://sakana.ai/ai-scientist/
136. Sakana AI, 'Sakana AI Wins Award at US-Japan Competition for Defense Innovation'. Mar. 24, 2025. [Online]. Available: https://sakana.ai/defense-challenge-2025/
137. T. Akiba, M. Shing, Y. Tang, Q. Sun, and D. Ha, 'Evolutionary optimization of model merging recipes', *Nature Machine Intelligence*, pp. 1–10, 2025.
138. National Institute of Information and Communications Technology, 'National Institute of Information and Communications Technology'. 2025. [Online]. Available: https://www.nict.go.jp/en/

139. KDDI Corporation, 'KDDI'. 2025. [Online]. Available: https://www.kddi.com/english/
140. Telecompaper, 'KDDI and NICT Start Joint Research on Multimodal LLMs'. Jul. 2024. [Online]. Available: https://www.telecompaper.com/news/kddi-and-nict-start-joint-research-on-multimodal-llms–1505017
141. L.-J. Consortium, 'LLM-jp'. 2023. [Online]. Available: https://llm-jp.nii.ac.jp/en/resources/
142. A. Aizawa et al., 'Llm-jp: A cross-organizational project for the research and development of fully open japanese llms', *arXiv preprint arXiv:2407.03963*, 2024.
143. NAVER Corporation, 'HyperCLOVA X'. 2025. [Online]. Available: https://clova.ai/en/hyperclova
144. B. Kim et al., 'What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers', *arXiv preprint arXiv:2109.04650*, 2021.
145. K. M. Yoo et al., 'Hyperclova x technical report', *arXiv preprint arXiv:2404.01954*, 2024.
146. NAVER Corporation, 'CLOVA Studio'. 2025. [Online]. Available: https://clova.ai/en/clova-studio
147. NAVER Corporation, 'CLOVA X'. 2025. [Online]. Available: https://clova-x.naver.com/welcome
148. Kakao Corporation, 'Kakao'. 2025. [Online]. Available: https://www.kakaocorp.com/
149. J. Cha, W. Kang, J. Mun, and B. Roh, 'Honeybee: Locality-enhanced projector for multimodal llm', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13817–13827.
150. LG AI Research, 'LG AI Research'. 2025. [Online]. Available: https://www.lgresearch.ai/
151. LG AI Research, 'EXAONE Official Site'. 2025. [Online]. Available: https://www.lgresearch.ai/exaone
152. L. Research et al., 'EXAONE 3.5: Series of Large Language Models for Real-world Use Cases', *arXiv preprint arXiv:2412.04862*, 2024.
153. Cohere, 'Cohere's Command R+ Model'. 2025. [Online]. Available: https://docs.cohere.com/v2/docs/command-r-plus
154. Cohere, 'Introducing Aya'. 2025. [Online]. Available: https://cohere.com/research/aya
155. A. Üstün et al., 'Aya model: An instruction finetuned open-access multilingual language model', *arXiv preprint arXiv:2402.07827*, 2024.
156. Cohere, 'Cohere Aya 101'. [Online]. Available: https://huggingface.co/CohereLabs/aya-101
157. Cohere, 'Aya Expanse - 8B'. 2025. [Online]. Available: https://huggingface.co/CohereLabs/aya-expanse-8b
158. Cohere, 'Aya Expanse - 32B'. 2025. [Online]. Available: https://huggingface.co/CohereLabs/aya-expanse-32b
159. Mila, 'Mila – Quebec AI Institute'. [Online]. Available: https://mila.quebec/
160. AI21 Labs, 'AI21 Labs'. [Online]. Available: https://www.ai21.com/
161. AI21 Labs, 'Jurassic-1 Jumbo'. [Online]. Available: https://llmmodels.org/tools/jurassic-1-jumbo/
162. O. Lieber, O. Sharir, B. Lenz, and Y. Shoham, 'Jurassic-1: Technical details and evaluation', *White Paper. AI21 Labs*, vol. 1, no. 9, pp. 1–17, 2021.
163. AI21 Editorial Team, 'Announcing Jurassic-2 and Task-Specific APIs'. [Online]. Available: https://www.ai21.com/blog/introducing-j2/
164. Sberbank, 'GigaChat'. [Online]. Available: https://giga.chat/
165. Yandex, 'YaLM-100B: Pretrained language model with 100B parameters'. [Online]. Available: https://github.com/yandex/YaLM-100B
166. Yandex, 'Yandex publishes YaLM 100B, the largest GPT-like neural network in open source'. [Online]. Available: https://yandex.com/company/press_center/press_releases/2022/2022-23-06
167. Technology Innovation Institute, 'tiiuae (Technology Innovation Institute)'. [Online]. Available: https://huggingface.co/tiiuae
168. Inception AI, 'JAIS'. [Online]. Available: https://inceptionai.ai/jais/index.html
169. C. Data61, 'Artificial Intelligence Foundation Models', CSIRO, Feb. 2024. [Online]. Available: https://www.csiro.au/-/media/D61/Files/2400012DATA61REPORTAIFoundationModelsWEB240208-1.pdf
170. G. L. Garcia et al., 'Introducing bode: a fine-tuned large language model for Portuguese prompt-based task', *arXiv preprint arXiv:2401.02909*, 2024.
171. L. AI, 'Lelapa AI launches Africa's first AI large language model'. [Online]. Available: https://africaworld.princeton.edu/news/2024/lelapa-ai-launches-africa%E2%80%99s-first-ai-large-language-model

doi:10.20944/preprints202504.2136.v1

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.