IntuitionLabs

INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

An Overview of Chinese Open-Source LLMs (Sept 2025)

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2025-09-29 · large language models · open source · china · artificial intelligence · mixture of experts · foundation models · natural language processing · ai

Original article: https://intuitionlabs.ai/articles/chinese-open-source-llms-2025

In this article

1. An Overview of Chinese Open-Source LLMs (Sept 2025)
2. Major Chinese Tech Companies
3. Leading Chinese AI Startups and Labs
4. Academic and Specialized Models
5. Post-September 2025 Supplement

An Overview of Chinese Open-Source LLMs (Sept 2025)

By mid-2025 China had become a global leader in open-source large language models (LLMs). According to Chinese state media, by July 2025 China accounted for **1,509** of the world's ~3,755 publicly released LLMs, ranking first globally; that count covers all released models, not only open-weight ones (www.reuters.com). This explosion reflects heavy state and industry investment in domestic AI and a strategic pivot by Chinese tech giants and startups toward publicly shared models. Availability and licensing vary by model: public weights do not necessarily carry an Apache- or MIT-style license, and some frontier systems remain API-only. The result is a "revival" of open-source AI, with dozens of Chinese LLMs now available for download or use via Hugging Face, GitHub, or cloud APIs (www.reuters.com) (www.reuters.com). These range from **general-purpose foundation models** dozens of billions of parameters in size to **specialized chatbots and domain experts**, many built on Mixture-of-Experts (MoE) architectures or with ultra-long context windows.

Update (October 2026): The leading Chinese open-weight models are now much larger and newer than the ones this overview was built around. As of October 10, 2026: DeepSeek's newest open model is **DeepSeek-V4.1-Flash** (September 10, 2026), a 552B-parameter MoE under the MIT license that DeepSeek says outperforms its own V4-Pro (api-docs.deepseek.com); Alibaba's **Qwen3.8** family (August 2026) brought a Max-class model, **Qwen3.8-2.4T-A95B** (2.4T total, 95B active parameters), to open release under a custom Qwen license, alongside the Apache 2.0 **Qwen3.8-27B** (huggingface.co); Moonshot AI published **Kimi K3** (2.8T total, 104B active, 1M-token context) in late July 2026 under its own Kimi K3 License (huggingface.co); Zhipu (Z.ai) released **GLM-5.2** (June 2026, MIT) and **GLM-5.3** (August 2026, custom license), both 744B total and 40B active parameters (huggingface.co); and MiniMax released **MiniMax-M3** (June 2026, about 428B total and 23B active parameters, MiniMax community license) (huggingface.co). Licensing has also shifted: several 2026 flagships ship under custom licenses rather than MIT or Apache 2.0, so check each model card before commercial use. The supplement at the end of this article lists the releases since September 2025.

Key Chinese open LLMs include offerings from major tech firms (e.g. Alibaba's *Qwen* series, Baidu's *Ernie*, and Tencent's *Hunyuan*), from leading startups (DeepSeek, Moonshot AI's *Kimi*, Zhipu AI's *ChatGLM*/GLM, Baichuan AI, MiniMax), and from academic labs (e.g. Fudan's *MOSS*). They are often openly shared with code and weights. For example, Alibaba open-sourced its Qwen 2.5 family (0.5B–72B parameters) in 2024, and in 2025 released further models like **Qwen3-Coder** (www.reuters.com) (www.reuters.com). Likewise, Zhipu AI has open releases of its *ChatGLM* and *GLM* models (most recently **GLM-4.5** and **GLM-4.5-Air** with 355B and 106B parameters) (www.reuters.com) (www.scmp.com). Competition is fierce: Chinese announcements often highlight that these open models now match or exceed U.S. models on benchmarks (for example, Alibaba claims Qwen-2.5-Max outperforms DeepSeek-V3 (www.reuters.com), and Alibaba said Qwen3-Coder matched leading U.S. models such as Anthropic's Claude and OpenAI's GPT-4 in certain areas, according to Reuters (www.reuters.com)). In short, by mid-2025 China's open-LLM ecosystem is vast and growing, featuring both general-purpose and domain-specific models at all scales.

Major Chinese Tech Companies

- **Alibaba Cloud – *Qwen* series (open source):** Alibaba has released dozens of Qwen models under open licenses. The original Qwen 2.5 family (0.5B–72B params) was open-sourced in 2024 (www.reuters.com). In 2025 Alibaba introduced additional Qwen variants: for example **Qwen2.5-Max** (a large-scale hosted MoE model) which Alibaba said surpasses competing models like DeepSeek-V3 (www.reuters.com), and **Qwen3-Coder-480B-A35B-Instruct** (480B total parameters and 35B active parameters), which Alibaba said outperformed domestic rivals in key coding capabilities and matched leading U.S. models, including Anthropic's Claude and OpenAI's GPT-4, in certain areas (vendor-reported) (www.reuters.com). Alibaba's models support multi-lingual and multimodal tasks (text, code, and image understanding). According to Reuters, Qwen Chat (Alibaba's chatbot service) allows developers to access all public Qwen models (including QwQ-32B, 32B parameters) by simply selecting them in the interface (www.reuters.com). (Alibaba also develops the Wan (Wanxiang) video-generation models; Wan2.1 and Wan2.2 were released as open weights under Apache 2.0 (github.com). Many Qwen models are open-weight, but this does not apply to every Qwen model; for example, Qwen2.5-Max was announced for Qwen Chat and Alibaba Cloud API access, without open weights (qwenlm.github.io).)
- **Moonshot AI – *Kimi* series (open weights):** Moonshot AI is an independent Beijing AI startup, not a ByteDance spin-off. Its Kimi team publicly documented **Kimi k1.5** in January 2025 as a multimodal model; the project describes 128K context scaling. In July 2025 Moonshot unveiled **Kimi K2**, an open-weight model with 1T total and 32B active parameters released under a Modified MIT license (huggingface.co), with further improvements in code and reasoning (for a US comparison, see OpenAI's gpt-oss open-weight models) (www.reuters.com). Moonshot said Kimi K2 outperforms mainstream open-source models such as DeepSeek-V3 in some areas and rivals leading U.S. models in certain functions such as coding; these are company claims reported by Reuters, which also notes that open-sourcing lets developers showcase their technical capabilities and expand developer communities (www.reuters.com). The Kimi line is distinguished by long context windows (128K tokens for k1.5 and the original K2, raised to 256K with Kimi-K2-Instruct-0905 (huggingface.co)) and multimodal input (e.g. image+text) (kimi-ai.chat) (www.reuters.com).
- **Baidu – *Ernie* series (open source as of 2025):** Baidu's Ernie models historically were proprietary, but due to intense competition, Baidu announced in early 2025 that its latest **Ernie** model would be made open-source (available from end-June 2025) (www.reuters.com). Baidu followed through at the end of June 2025 with the ERNIE 4.5 family: 10 variants, the largest with 424B total and 47B active parameters, all under Apache 2.0 (github.com). The Ernie chatbot (Ernie Bot) was also made free to users from April 2025 (www.reuters.com). (At the time Baidu also planned Ernie 5 with multimodal capabilities for H2 2025; Ernie 5.0 and 5.1 later launched without published weights, as described in the supplement below (www.reuters.com).) While Ernie had limited adoption relative to newer Chinese challengers, its open-sourcing is significant, bringing one of China's original AI models into the open ecosystem (www.reuters.com).

- **Tencent – *Hunyuan*:** Tencent released **Hunyuan-A13B** as an open-source Mixture-of-Experts LLM in June 2025. Tencent describes it as an 80B-parameter model with 13B active parameters and publishes the weights and associated materials in its official repository (Tencent Hunyuan).

Leading Chinese AI Startups and Labs

- **DeepSeek (深度求索) – *DeepSeek* series (open source):** DeepSeek is a Hangzhou startup that burst onto the scene with high-efficiency MoE LLMs. In late 2024 it released **DeepSeek V3** (671B total parameters, with 37B activated for each token) on Hugging Face (www.reuters.com). On January 20, 2025, DeepSeek released **DeepSeek R1**, a reasoning-optimized model, under the MIT License; its code, models, and technical report were made public (www.reuters.com). In September 2025 they published **DeepSeek V3.2-Exp** (an experimental intermediate version) on Hugging Face (www.reuters.com). The company emphasizes low running costs: it published its Native Sparse Attention research in February 2025 (arxiv.org), and V3.2-Exp introduced a separate mechanism, DeepSeek Sparse Attention (DSA), aimed at long-context efficiency (huggingface.co) (www.reuters.com) (www.reuters.com).
- **Zhipu AI (智谱 AI) – *ChatGLM/GLM* series (open source):** Zhipu AI (the startup spun out of Tsinghua University) has long produced open Chinese chat models. Its **ChatGLM** series (e.g. ChatGLM-6B, ChatGLM2-6B, ChatGLM3-6B) includes bilingual English/Chinese chat models; license and model-use terms vary by release. In 2025 Zhipu expanded into larger MoE architectures. Reuters reports that in July 2025 Zhipu released **GLM-4.5** (355B params) and **GLM-4.5-Air** (106B), which it described as China's most advanced open-source MoE models, built on Zhipu's own architecture; the model card lists 32B and 12B active parameters respectively and an MIT license (huggingface.co) (www.reuters.com) (www.scmp.com). Reuters describes Zhipu as one of China's "AI tigers" and notes that it is backed by local governments (www.reuters.com).
- **Moonshot AI – *Kimi* series:** Moonshot AI is an independent startup; its Kimi models are covered above under Major Chinese Tech Companies.
- **MiniMax (Shanghai Xiyu Technology):** a Shanghai AI startup, MiniMax in Jan 2025 launched its *MiniMax-01* LLM family. These include **MiniMax-Text-01** (general open LLM) and **MiniMax-VL-01** (multimodal text+vision) (www.scmp.com). MiniMax benchmarks claim parity with leading AI models in math, reasoning and instruction following. Significantly, its announcement emphasized the models are **low-cost and open-source**, aiming to rival US tech. The weights were published on Hugging Face under MiniMax's own model license (the code is MIT); MiniMax-Text-01 has 456B total parameters, 45.9B active per token, and handles up to 4 million tokens of context at inference (huggingface.co) (www.scmp.com). MiniMax's debut underscores the trend: Chinese startups repeatedly tout cost-effectiveness of open models (DeepSeek, MiniMax) versus expensive alternatives (www.scmp.com) (www.reuters.com).
- **Baichuan AI – *Baichuan* series (open source):** Baichuan AI (founded by Wang Xiaochuan, ex-Sogou) has released open LLMs in 2023–25. Its first **Baichuan-7B** and **Baichuan-13B** (both support Chinese+English) were released in mid-2023, with their code under Apache 2.0 and their weights under

a separate Baichuan community license that requires registration for commercial use. These models were explicitly described as “open-source” by the company (github.com). In September 2023 Baichuan released **Baichuan 2** as open models in 7B and 13B sizes (base and chat versions), trained on 2.6 trillion tokens (arxiv.org). (We omit press citations here, but the company’s GitHub and Hugging Face pages publicly host Baichuan weights.) Baichuan’s 13B model was noteworthy in China as a big open model rivaling Western efforts. The Factbox cited Baichuan among startups strong in open LLMs (www.reuters.com).

- **“01.AI” (OneZero) – *Yi* series (open source):** This Beijing startup has released the **Yi** models: Yi-6B and Yi-34B in November 2023 (initially under a Yi community license) and the Yi-1.5 series (6B, 9B and 34B) in May 2024, now published under Apache 2.0 and aimed at both Chinese and English tasks (huggingface.co). Similarly, Chinese firms like Vivo (with *BlueLM*) and Shenzhen YuanXiang (*XVERSE-7B/13B/65B*) have open LLMs on GitHub/Hugging Face (github.com).
- **Other startups:** StepFun publishes open-weight models under Apache 2.0, from **Step-3** (July 2025, 321B total and 38B active parameters) (huggingface.co) to **Step-3.7-Flash** (May 2026, 198B total and about 11B active) (huggingface.co). Availability should be assessed release by release: one open release does not establish that every model or service from the same company is open.

Academic and Specialized Models

- **Fudan University – *MOSS***: Fudan’s NLP lab developed **MOSS**, a Chinese ChatGPT-like model. It was announced in April 2023 as “the first open-source conversational language model in China with plugin enhancements” (pandaily.com). (MOSS’s code is released under Apache 2.0, its data under CC BY-NC 4.0 and its model weights under GNU AGPL 3.0 (github.com).) MOSS helped kickstart academic interest in Chinese open models.
- **Tsinghua/China Academy – ChatGLM:** Related to Zhipu, Tsinghua’s THUDM research group has released ChatGLM-6B and its successors ChatGLM2-6B and ChatGLM3-6B, bilingual chat models; ChatGLM-6B was trained on about 1 trillion tokens of Chinese and English (github.com). The early GLM models (GLM-10B, GLM-130B) laid the groundwork; the GLM-130B paper appeared in October 2022 and was accepted to ICLR 2023 (arxiv.org) (www.index.dev). ChatGLM-6B was publicly available in March 2023, and the official project records the release of ChatGLM2-6B on June 25, 2023. The repositories set out the applicable model and commercial-use terms.
- **Beijing Academy (BAAI) – *Aquila* series:** BAAI (NonProfit Beijing AI Institute) has developed the *WuDao/Aquila* models. Their earlier releases (WuDao 2.0 multimodal, WuDao-EVA, etc) were partly open. In 2023 BAAI released **Aquila-7B** and **AquilaChat-7B** on Hugging Face; Aquila-7B is a base model, while AquilaChat-7B is a supervised fine-tuned and reinforcement-learning-based chat model (github.com). BAAI followed with the **Aquila2** series (7B and 34B) in October 2023 (github.com). (The VisualGLM-6B vision-language model, sometimes listed alongside these, comes from Tsinghua’s THUDM group and builds on ChatGLM-6B (github.com).) (BAAI’s models, though emerging from a Chinese institute, follow a global open science style: most of their code and smaller models are public.)

- **Domain-specific Chinese LLMs:** Chinese teams have developed specialist models for areas such as law, medicine, education, and finance. Availability, licensing, and whether weights are published must be checked for each release; a domain label does not establish that a model is open-weight or available through an open API. For example, the public ChatLaw repository describes a Chinese legal model and links to its model distribution (ChatLaw repository).
- **Others and Derivatives:** Finally, the Chinese open-LLM ecosystem includes numerous derivative and community models. For instance, curated lists on GitHub (like “Awesome Chinese LLMs”) track dozens of smaller projects: language-model variants like *Mengzi* (by Langboat; Mengzi3-13B is Apache 2.0) (huggingface.co), *Skywork* (Kunlun Group’s Skywork-13B, trained on 3.2 trillion tokens) (github.com), etc. Many of these arise from university labs or mid-tier AI firms and often reuse open frameworks (LLaMA, Bloom) to produce Chinese-optimized models.

In summary, by September 2025 China had a broad ecosystem of publicly released LLM code and weights, with licensing and availability varying substantially by model. Some *headline* models and their releases include: Alibaba’s **Qwen-2.5** (0.5–72B, Sep 2024) and **Qwen3-Coder-480B-A35B-Instruct** (480B total parameters, 35B active; Jul 2025) (www.reuters.com) (www.reuters.com); Zhipu’s **GLM-4.5** (355B) (www.reuters.com); DeepSeek’s **V3** (671B total parameters, 37B activated) and **R1** (671B total parameters, 37B activated) (www.reuters.com) (www.reuters.com); Moonshot’s **Kimi K2** (open-weight; 1T total parameters and 32B activated) and the separately documented **Kimi k1.5** research model, for which Moonshot reports 128K context scaling and multimodal reasoning (Moonshot AI’s Kimi k1.5 repository) (www.reuters.com); and Fudan’s **MOSS** (open-chat 2023) (pandaily.com). Alongside these are many smaller and specialized models; their public availability ranges from code or downloadable weights to limited demonstrations or hosted services, so each release’s repository, model card, and license should be checked individually. By Xinhua’s count of all publicly released LLMs (not only open-weight ones), China leads the world in sheer number of models (www.reuters.com) (www.reuters.com), fueling everything from startup innovation to global research.

Post-September 2025 Supplement

The September 2025 overview ends above. The material below is a separately dated supplement covering subsequent developments and is not part of the September 2025 inventory.

Newest open-weight LLM flagship per lab, as listed on each lab's official Hugging Face page or release notes on October 10, 2026:

Lab	Newest open-weight flagship	Weights published	Total / active parameters	License
DeepSeek	DeepSeek-V4.1-Flash	Sep 2026	552B MoE (8B active for input, 16B for output)	MIT
Alibaba (Qwen)	Qwen3.8-2.4T-A95B	Aug 2026	2.4T / 95B	Custom (qwen3.8-max)
Moonshot AI	Kimi K3	Jul 2026	2.8T / 104B	Custom (Kimi K3 License)

Lab	Newest open-weight flagship	Weights published	Total / active parameters	License
Zhipu (Z.ai)	GLM-5.3	Aug 2026	744B / 40B	Custom (glm-5.3)
MiniMax	MiniMax-M3	Jun 2026	about 428B / 23B	Custom (MiniMax community)
Tencent	Hy4-preview	Aug 2026	770B / 49B	Apache 2.0
Meituan	LongCat-2.0	Jul 2026	1.6T / about 48B	MIT
Xiaomi	MiMo-V2.6-Pro-RL	Sep 2026	1.02T / 42B	MIT
Ant Group (inclusionAI)	Ling-2.6-1T	Apr 2026	about 1T total	MIT
StepFun	Step-3.7-Flash	May 2026	198B / about 11B	Apache 2.0
Baidu	ERNIE 4.5 (largest: ERNIE-4.5-VL-424B-A47B)	Jun 2025	424B / 47B	Apache 2.0
ByteDance	Seed-OSS-36B	Aug 2025	36B dense	Apache 2.0

Sources for each row are cited in the bullets below and, for Baidu, in the main text above.

- **DeepSeek moved from V3.2-Exp to a V4 generation.** On April 24, 2026 DeepSeek published a preview release of **DeepSeek V4**, comprising **V4-Pro** (1.6T total parameters, 49B active) and **V4-Flash** (284B total, 13B active), pairing token-wise compression with DeepSeek Sparse Attention (DSA), with weights under the MIT license (huggingface.co); the company states that "1M context is now the default across all official DeepSeek services" (api-docs.deepseek.com). DeepSeek characterizes V4-Pro as delivering performance rivaling top closed-source models, a claim that, as with earlier releases, should be read as vendor-reported (api-docs.deepseek.com). Before V4, DeepSeek had released **DeepSeek-V3.2** and **V3.2-Speciale** (MIT) in December 2025 (huggingface.co). V4-Pro reached general availability on August 13, 2026 (api-docs.deepseek.com). On September 10, 2026 DeepSeek released **DeepSeek-V4.1-Flash**, a 552B-parameter MoE with a causal encoder-decoder design (8B active parameters for input, 16B for output) and MIT-licensed weights; DeepSeek said tests put it ahead of V4-Pro and began routing V4-Pro API requests to it on September 14, 2026, until a V4.1-Pro launches (api-docs.deepseek.com) (huggingface.co).
- **Alibaba shipped the Qwen3.5 and Qwen3.6 open-weight families.** Between February 16 and March 2, 2026 Alibaba released open weights for Qwen3.5 models ranging from a 397B-A17B MoE flagship down to a 0.8B model, followed by **Qwen3.6-35B-A3B** (April 16) and **Qwen3.6-27B** (April 22), all under Apache 2.0: the Qwen team states that "all our open-weight models are licensed under Apache 2.0" (github.com). Notably, Alibaba's top-end models have diverged from this pattern: **Qwen3.7-Plus** (June 2, 2026), which adds vision, deep reasoning, tool invocation, and autonomous iteration, is proprietary and API-only on the Bailian platform, with no published weights (marktechpost.com). Alibaba partly reversed that pattern with **Qwen3.8** in August 2026, which it describes as bringing "a Qwen-Max-class

model to open release" for the first time: **Qwen3.8-2.4T-A95B** (2.4T total, 95B activated parameters) is published under a custom qwen3.8-max license rather than Apache 2.0, while **Qwen3.8-27B** remains Apache 2.0 (huggingface.co) (huggingface.co).

- **Zhipu (now branded Z.ai) iterated from GLM-4.6 to GLM-5.3. GLM-4.6** (357B parameters, MIT license, 200K-token context) was published on Hugging Face shortly after this article's original date (huggingface.co). It was followed by GLM-4.7 (December 2025) and **GLM-5** (February 2026, reported at 744B total and 40B active parameters); GLM-5.1 weights were subsequently opened in April 2026, keeping Z.ai in the open-weight camp (presenc.ai). **GLM-5.2** followed in June 2026 under MIT with a 1M-token context (huggingface.co), and **GLM-5.3** in August 2026 on the same base model but under a custom glm-5.3 license that requires very large model-as-a-service operators (over $10 billion in annual revenue) to pass a Z.ai security review before commercial use (huggingface.co).
- **Moonshot AI extended the Kimi K2 line into K2.6 and K2.7-Code, then moved to K3.** The coding-focused **Kimi K2.7-Code** shipped in June 2026 with open weights on Hugging Face under a Modified MIT license; Moonshot reports a 21.8% improvement on its Kimi Code Bench v2 over K2.6 and states that K2.7-Code cuts reasoning token usage by roughly 30% relative to K2.6 (devops.com). In late July 2026 Moonshot published **Kimi K3**, a natively multimodal model with 2.8T total and 104B activated parameters and a 1M-token context, under a custom Kimi K3 License rather than Modified MIT (huggingface.co).
- **MiniMax pivoted its M-series toward agents and code. MiniMax M2** (a compact MoE with 230B total and 10B active parameters, aimed at coding and agentic workflows) was open-sourced on Hugging Face with API pricing of $0.30 per million input tokens and $1.20 per million output tokens, which MiniMax describes as "8% of Claude 4.5 Sonnet's" price at roughly double the inference speed (minimax.io). A later release, **MiniMax M2.7**, had its weights published in April 2026 under a non-commercial license (commercial use requires MiniMax's written authorization) (huggingface.co), with reported scores of 56.22% on SWE-Pro and 57.0% on Terminal Bench 2, figures MiniMax positions as the strongest among open-source models (marktechpost.com). **MiniMax-M3** (June 2026) is natively multimodal with a 1M-token context, about 428B total and 23B activated parameters, and a custom MiniMax community license (huggingface.co).
- **Tencent's open-weight models moved to the Hy name and to Apache 2.0.** Hunyuan-A13B (June 2025) was released under the Tencent Hunyuan Community License, which does not apply in the European Union, the United Kingdom or South Korea (huggingface.co). Its 2026 successors are Apache 2.0: **Hy3** (July 2026, 295B total and 21B active parameters) (huggingface.co) and **Hy4-preview** (August 2026, 770B total and 49B active) (huggingface.co).
- **Larger companies outside the original list now ship open-weight flagships.** ByteDance released **Seed-OSS-36B** (August 2025, 36B dense, Apache 2.0) (huggingface.co). Meituan went from **LongCat-Flash** (August 2025, 560B total, MIT) (huggingface.co) to **LongCat-2.0** (July 2026, 1.6T total and about 48B active, MIT) (huggingface.co). Ant Group's inclusionAI released the trillion-parameter **Ling-1T** (October 2025, MIT) (huggingface.co) and later **Ling-2.6-1T** (April 2026, MIT) (huggingface.co). Xiaomi went from **MiMo-V2-Flash** (December 2025, 309B total and 15B active) (huggingface.co) to **MiMo-V2.6-Pro-RL** (September 2026, 1.02T total and 42B active), both MIT (huggingface.co).

- **Baidu's Ernie line has split into an open tier and a closed frontier tier.** Baidu open-sourced its **Ernie 4.5** series in 2025, but its newer **Ernie 5.0** and **Ernie 5.1** (released May 8, 2026) remain API-only: an independent review of Ernie 5.1 notes that "as with Ernie 5.0, Baidu hasn't released model weights, so the benchmark scores and efficiency claims can't be independently verified" (the-decoder.com). This mirrors the pattern already seen with Alibaba's Qwen3.7-Plus: as Chinese labs push toward frontier-level agentic and reasoning capability, the newest flagship tier increasingly stays closed even as each lab continues to open-source its mid-tier and prior-generation models.

Two broader patterns are worth noting. First, benchmark claims from all of these labs are increasingly framed against Western frontier models on agentic coding suites (SWE-style benchmarks, terminal benchmarks) rather than general knowledge tests, reflecting where commercial demand has moved. Second, licensing is no longer uniformly permissive at the top end: while most of Qwen's open-weight line remains Apache 2.0 and GLM-4.6 carries an MIT license (huggingface.co), some flagship models (Qwen3.7-Plus, and periodically Zhipu's newest tier before delayed weight releases) are held back as proprietary or released on a lag, and several 2026 flagships that are released carry custom licenses (Qwen3.8-2.4T-A95B, GLM-5.3, Kimi K3, MiniMax-M3, and MiniMax M2.7's non-commercial terms). The overall picture from September 2025 still holds: China's open-weight ecosystem remains large and active, but the strategic calculus of which weights get released, and when, has become more selective.

Sources: Chinese AI news reports and analyses (Reuters, SCMP, Pandaily, etc.) provide details on each model. For example, Reuters documents Alibaba’s Qwen releases (www.reuters.com) (www.reuters.com) and Baidu’s Ernie open-sourcing (www.reuters.com), Reuters/SCMP describe DeepSeek, Zhipu and Moonshot releases (www.reuters.com) (www.reuters.com) (www.reuters.com), and Pandaily reports Fudan’s MOSS (pandaily.com). These reports (cited above) detail the scale, licensing, and capabilities of the models listed here.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://intuitionlabs.ai/articles/chatgpt-understanding-architecture-llm
2. https://www.reuters.com/technology/chinas-ai-startup-zhipu-releases-open-source-model-glm-45-2025-07-28/
3. https://www.reuters.com/technology/artificial-intelligence/chinas-ai-firms-take-spotlight-with-deals-low-cost-models-2025-02-14/
4. https://intuitionlabs.ai/articles/mixture-of-experts-moe-models
5. https://api-docs.deepseek.com/news/news260910
6. https://huggingface.co/Qwen/Qwen3.8-2.4T-A95B
7. https://huggingface.co/moonshotai/Kimi-K3
8. https://huggingface.co/zai-org/GLM-5.3
9. https://huggingface.co/MiniMaxAI/MiniMax-M3
10. https://www.reuters.com/technology/alibaba-accelerates-ai-push-by-releasing-new-open-source-models-text-to-video-2024-09-19/

11. https://www.reuters.com/world/china/alibaba-launches-open-source-ai-coding-model-touted-its-most-advanced-date-2025-07-23/
12. https://www.scmp.com/tech/big-tech/article/3320136/alibaba-zhipu-roll-out-new-ai-models-amid-heated-open-source-race?module=top_story&pgtype=section
13. https://www.reuters.com/technology/artificial-intelligence/alibaba-releases-ai-model-it-claims-surpasses-deepseek-v3-2025-01-29/
14. https://intuitionlabs.ai/articles/key-innovations-behind-chatgpt
15. https://www.reuters.com/technology/alibaba-shares-surge-after-it-unveils-reasoning-model-2025-03-06/
16. https://github.com/Wan-Video/Wan2.2
17. https://qwenlm.github.io/blog/qwen2.5-max/
18. https://intuitionlabs.ai/articles/kimi-k2-open-weight-llm-analysis
19. https://huggingface.co/moonshotai/Kimi-K2-Instruct
20. https://intuitionlabs.ai/articles/openai-gpt-oss-open-weight-models
21. https://www.reuters.com/business/media-telecom/chinas-moonshot-ai-releases-open-source-model-reclaim-market-position-2025-07-11/
22. https://huggingface.co/moonshotai/Kimi-K2-Instruct-0905
23. https://kimi-ai.chat/models/kimi-k1-5/
24. https://www.reuters.com/technology/artificial-intelligence/baidu-make-ernie-ai-model-open-source-end-june-2025-02-14/
25. https://github.com/PaddlePaddle/ERNIE
26. https://github.com/Tencent-Hunyuan/Hunyuan-A13B
27. https://www.reuters.com/technology/artificial-intelligence/chinas-deepseek-releases-ai-model-upgrade-intensifies-rivalry-with-openai-2025-03-25/
28. https://www.reuters.com/technology/artificial-intelligence/deepseek-share-some-ai-model-code-doubling-down-open-source-2025-02-21/
29. https://www.reuters.com/technology/deepseek-releases-model-it-calls-intermediate-step-towards-next-generation-2025-09-29/
30. https://arxiv.org/abs/2502.11089
31. https://huggingface.co/deepseek-ai/DeepSeek-V3.2-Exp
32. https://huggingface.co/zai-org/GLM-4.5
33. https://www.scmp.com/tech/big-tech/article/3294900/chinese-ai-start-minimax-releases-low-cost-open-source-models-rival-top-chatbots?module=Tech+Trends&pgtype=section
34. https://huggingface.co/MiniMaxAI/MiniMax-Text-01
35. https://github.com/baichuan-inc/baichuan-13B
36. https://arxiv.org/html/2309.10305v1
37. https://huggingface.co/01-ai/Yi-1.5-34B
38. https://github.com/xverse-ai/XVERSE-7B
39. https://huggingface.co/stepfun-ai/step3
40. https://huggingface.co/stepfun-ai/Step-3.7-Flash
41. https://pandaily.com/fudan-universitys-chatgpt-like-model-moss-officially-open-sourced

42. https://github.com/OpenMOSS/MOSS
43. https://github.com/THUDM/ChatGLM-6B
44. https://arxiv.org/abs/2210.02414
45. https://www.index.dev/blog/chinese-open-source-llms
46. https://github.com/wgwang/awesome-LLMs-In-China/blob/main/README.md
47. https://github.com/FlagAI-Open/Aquila2
48. https://github.com/THUDM/VisualGLM-6B
49. https://github.com/PKU-YuanGroup/ChatLaw
50. https://huggingface.co/Langboat/Mengzi3-13B-Base
51. https://github.com/SkyworkAI/Skywork
52. https://github.com/MoonshotAI/Kimi-k1.5
53. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro
54. https://api-docs.deepseek.com/news/news260424/
55. https://huggingface.co/deepseek-ai/DeepSeek-V3.2
56. https://api-docs.deepseek.com/news/news260813
57. https://huggingface.co/deepseek-ai/DeepSeek-V4.1-Flash
58. https://github.com/QwenLM/Qwen3.6
59. https://www.marktechpost.com/2026/06/02/alibabas-qwen-team-launches-qwen3-7-plus-adding-vision-deep-reasoning-tool-invocation-and-autonomous-iteration-on-the-bailian-platform/
60. https://huggingface.co/Qwen/Qwen3.8-27B
61. https://huggingface.co/zai-org/GLM-4.6
62. https://presenc.ai/research/zhipu-glm-model-lineage-2026
63. https://huggingface.co/zai-org/GLM-5.2
64. https://devops.com/moonshot-ais-kimi-k2-7-code-targets-token-efficiency-in-agentic-coding/
65. https://www.minimax.io/news/minimax-m2
66. https://huggingface.co/MiniMaxAI/MiniMax-M2.7
67. https://www.marktechpost.com/2026/04/12/minimax-just-open-sourced-minimax-m2-7-a-self-evolving-agent-model-that-scores-56-22-on-swe-pro-and-57-0-on-terminal-bench-2/
68. https://huggingface.co/tencent/Hunyuan-A13B-Instruct
69. https://huggingface.co/tencent/Hy3
70. https://huggingface.co/tencent/Hy4-preview
71. https://huggingface.co/ByteDance-Seed/Seed-OSS-36B-Instruct
72. https://huggingface.co/meituan-longcat/LongCat-Flash-Chat
73. https://huggingface.co/meituan-longcat/LongCat-2.0
74. https://huggingface.co/inclusionAI/Ling-1T
75. https://huggingface.co/inclusionAI/Ling-2.6-1T
76. https://huggingface.co/XiaomiMiMo/MiMo-V2-Flash

77. https://huggingface.co/XiaomiMiMo/MiMo-V2.6-Pro-RL

78. https://the-decoder.com/baidus-ernie-5-1-cuts-94-percent-of-pre-training-costs-while-competing-with-top-models/

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.