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

By InuitionLabs.ai • 9/28/2025 • 10 min read

large language models | open source | china | artificial intelligence | mixture of experts | foundation models | natural language processing

Chinese Open-Source LLM Landscape (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, far more than any other country (www.reuters.com). This explosion reflects heavy state and industry investment in domestic AI, open licensing (often Apache- or MIT-style), and a strategic pivot by Chinese tech giants and startups toward publicly shared models. 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.

Key Chinese open LLMs include offerings from major tech firms (e.g. Alibaba's *Qwen* series, Baidu's *Ernie*, ByteDance's *Kimi*), from leading startups (DeepSeek, Moonshot/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 Reuters reports Qwen3-Coder rivals OpenAI's GPT-4 on code tasks (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** (an enhanced 2.5B model) which Alibaba said surpasses competing models like DeepSeek-V3 (www.reuters.com), and **Qwen3-Coder** (an advanced 32B coder model) which it touted as outperforming domestic rivals and matching GPT-4 on code generation (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 maintains commercial models like Wanxiang for video, but its Qwen LLM line is fully open-sourced (www.reuters.com).)

- **ByteDance (Moonshot AI) – *Kimi* series (open source):** ByteDance's AI spin-off Moonshot AI has released a line of *Kimi* models. The original Kimi (K1) and upgraded **Kimi K1.5** appeared in 2024, offering 20B+ parameter multimodal capabilities. In July 2025 Moonshot unveiled ** Kimi K2** (open-source) with further improvements in code and reasoning (www.reuters.com). Reuters notes Kimi K2 excels in coding tasks and, like DeepSeek, is released under an open-source strategy (Moonshot explicitly "follows the example set by Meta" in open-sourcing its advanced models) (www.reuters.com). The Kimi line is distinguished by a very large context window (128K tokens) 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). The Ernie chatbot (Ernie Bot) was also made free to users from April 2025 (www.reuters.com). (Baidu also plans Ernie 5 with multimodal capabilities in H2 2025 (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* (proprietary):** Tencent's "Hunyuan" supermodel is often cited alongside Ernie and Yazhou (Chinese models), but unlike the above, it has remained closed at least to end-2025. (Chinese press notes it claims to match GPT-4, but we have no report of open release.)

- **SenseTime – Unified Multimodal Model:** SenseTime (AI firm, Hong Kong-listed) has released a new "unified large model" that handles text, images, and reasoning (www.scmp.com). Although not explicitly noted as open-source in reports, SenseTime often follows an open approach in infrastructure releases (e.g. the "LazyLLM" framework (www.sensetime.com)). Its in-house model is noteworthy as a Chinese multimodal LLM, complementing the language-only models above.

Leading Chinese AI Startups and Labs

- **DeepSeek ( ) – *DeepSeek* series (open source):** DeepSeek is a Hangzhou startup that burst onto the scene with high-efficiency MoE LLMs. In late 2024 it open-sourced **DeepSeek V3** (estimated ~250B parameters with only 37B active per query) on Hugging Face (www.reuters.com). In February 2025 DeepSeek also open-sourced **DeepSeek R1** (a reasoning-optimized model) and released its code repositories for full transparency (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 and released its Native Sparse Attention algorithm in tandem with these models (www.reuters.com) (www.reuters.com). Industry reports rank DeepSeek's models at the top of domestic Chinese benchmarks for reasoning and coding.

- **Zhipu AI (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) are bilingual (English/Chinese) chatbots released under Apache-2.0 licenses. 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) – "the most advanced open-source MoE model [s]" in China – built on Zhipu's own architecture (www.reuters.com) (www.scmp.com). (Zhipu's releases have collectively been downloaded millions of times worldwide.) Zhipu's road map ties closely to Chinese AI policy; as the Reuters Factbox notes, Zhipu has been dubbed one of China's "AI tigers" and its open-LLM efforts help the country fulfill government goals (www.reuters.com).

- **Moonshot AI ( ) –** *Kimi* series is covered under ByteDance above (Moonshot is backed by Alibaba/Tencent).

- **MiniMax ( ) –** 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. They were released with permissive licenses on Chinese repos (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 under Apache-2.0 licenses in mid-2023. These models were explicitly described as "open-source" by the company (github.com). In late 2023 and early 2024 Baichuan announced larger models (e.g. Baichuan-2 with 33B) also as open-source (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):** A Beijing startup sometimes called OneZero has released the **Yi** models (Yi-6B, Yi-1.5 [15B], etc), openly licensed and aimed at both Chinese and English tasks. (Reuters did not cover Yi explicitly, but industry trackers list "Yi" as an upcoming 15B bilingual model in early 2024.) 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:** Several other Chinese AI firms have released open LLMs or are preparing to. For example, 360's *Zhinao* (“ ”) model, SenseTime's unified model (above), and new ones like Hotchips' *RocLM*, No Flask's *TigerBot* or *Jinshan Yiyun* (Allscripts) in finance/media all hint at open releases. The Reuters Factbox names “Doubao” (ByteDance's inexpensive chatbot) and Tencent's “Hunyuan,” though Doubao appears closed and Hunyuan currently is not open (www.reuters.com).

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 offered a public API and shared model code under an Apache-2 license.) MOSS helped kickstart academic interest in Chinese open models.

- **Tsinghua/China Academy – ChatGLM:** Related to Zhipu, Tsinghua's OpenAI-like research groups have released ChatGLM (GLM-6B) and its successor ChatGLM2 (6B) and ChatGLM3, bilingual chat models trained on 100B+ tokens of Chinese/English. The early GLM models (GLM-10B, GLM-130B) laid the groundwork; GLM-130B (the largest GLM) was published in 2023 (www.index.dev). ChatGLM-6B (released 2021) and ChatGLM2-6B (2022) were open-sourced under MIT in 2023, sparking wide use in Chinese communities (unfortunately no Reuters cite, but widely reported).

- **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 2024–25 BAAI published **Aquila-7B** (code generation) and **AquilaChat-7B** on Hugging Face (github.com). It also shared “VisualGLM” models for vision–language tasks. (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:** Dozens of Chinese teams have fine-tuned and open-released models for special domains. Examples include **ChatLaw** (legal assistant model from Beijing Univ. of Posts & Telecom and others), **DoctorGLM** (ShanghaiTech University medical model) and its follow-ups, **EduChat** (education), **TigerBot** (financial questions), and many more. While individual citations are scarce, Chinese tech blogs and repositories catalog dozens of such models (often based on LLaMA/ChatGLM fine-tuning) targeting finance, healthcare, logistics, etc. All are typically released with open weights or via open APIs.

- **Others and Derivatives:** Finally, the Chinese open-LLM ecosystem includes numerous derivative and community models. For instance, open congregations on GitHub (like “Awesome Chinese LLMs”) list dozens of smaller projects: language-model variants like *MengZi* (by Research Institute of Intelligent Vision, 13B), *LingJing* (education 7B), *XPE* (customer service model by Xiaoduo.AI), *Skywork* (inspur/TianGong open 13B base/chat/MM), 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 the Chinese ecosystem of open LLMs is remarkably rich. Some *headline* models and their releases include: Alibaba's **Qwen-2.5** (0.5–72B, Sep 2024) and **Qwen3-Coder** (Jul 2025) (www.reuters.com) (www.reuters.com); Zhipu's **GLM-4.5** (355B) (

www.reuters.com); DeepSeek's **V3** (250B MoE) and **R1** (671B MoE, 37B active) (www.reuters.com) (www.reuters.com); Moonshot's **Kimi K1.5/K2** (multimodal, 128K context) (kimi-ai.chat) (www.reuters.com); and Fudan's **MOSS** (open-chat 2023) (pandaily.com). Alongside these are many smaller and specialized models, all of which (to varying degrees) share their parameters or code with the public. The result is an open-weight AI ecosystem in China that now outpaces the West in sheer number of models (www.reuters.com) (www.reuters.com), fueling everything from startup innovation to global research.

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.

IntuitionLabs - Industry Leadership & Services

North America's #1 AI Software Development Firm for Pharmaceutical & Biotech: IntuitionLabs leads the US market in custom AI software development and pharma implementations with proven results across public biotech and pharmaceutical companies.

Elite Client Portfolio: Trusted by NASDAQ-listed pharmaceutical companies including Scilex Holding Company (SCLX) and leading CROs across North America.

Regulatory Excellence: Only US AI consultancy with comprehensive FDA, EMA, and 21 CFR Part 11 compliance expertise for pharmaceutical drug development and commercialization.

Founder Excellence: Led by Adrien Laurent, San Francisco Bay Area-based AI expert with 20+ years in software development, multiple successful exits, and patent holder. Recognized as one of the top AI experts in the USA.

Custom AI Software Development: Build tailored pharmaceutical AI applications, custom CRMs, chatbots, and ERP systems with advanced analytics and regulatory compliance capabilities.

Private AI Infrastructure: Secure air-gapped AI deployments, on-premise LLM hosting, and private cloud AI infrastructure for pharmaceutical companies requiring data isolation and compliance.

Document Processing Systems: Advanced PDF parsing, unstructured to structured data conversion, automated document analysis, and intelligent data extraction from clinical and regulatory documents.

Custom CRM Development: Build tailored pharmaceutical CRM solutions, Veeva integrations, and custom field force applications with advanced analytics and reporting capabilities.

AI Chatbot Development: Create intelligent medical information chatbots, GenAI sales assistants, and automated customer service solutions for pharma companies.

Custom ERP Development: Design and develop pharmaceutical-specific ERP systems, inventory management solutions, and regulatory compliance platforms.

Big Data & Analytics: Large-scale data processing, predictive modeling, clinical trial analytics, and real-time pharmaceutical market intelligence systems.

Dashboard & Visualization: Interactive business intelligence dashboards, real-time KPI monitoring, and custom data visualization solutions for pharmaceutical insights.

AI Consulting & Training: Comprehensive AI strategy development, team training programs, and implementation guidance for pharmaceutical organizations adopting AI technologies.

Contact founder Adrien Laurent and team at https://intuitionlabs.ai/contact for a consultation.

DISCLAIMER

The information contained in this document is provided for educational and informational purposes only. We make no representations or warranties of any kind, express or implied, about the completeness, accuracy, reliability, suitability, or availability of the information contained herein.

Any reliance you place on such information is strictly at your own risk. In no event will IntuitionLabs.ai or its representatives be liable for any loss or damage including without limitation, indirect or consequential loss or damage, or any loss or damage whatsoever arising from the use of information presented in this document.

This document may contain content generated with the assistance of artificial intelligence technologies. AI-generated content may contain errors, omissions, or inaccuracies. Readers are advised to independently verify any critical information before acting upon it.

All product names, logos, brands, trademarks, and registered trademarks mentioned in this document are the property of their respective owners. All company, product, and service names used in this document are for identification purposes only. Use of these names, logos, trademarks, and brands does not imply endorsement by the respective trademark holders.

IntuitionLabs.ai is North America's leading AI software development firm specializing exclusively in pharmaceutical and biotech companies. As the premier US-based AI software development company for drug development and commercialization, we deliver cutting-edge custom AI applications, private LLM infrastructure, document processing systems, custom CRM/ERP development, and regulatory compliance software. Founded in 2023 by Adrien Laurent, a top AI expert and multiple-exit founder with 20 years of software development experience and patent holder, based in the San Francisco Bay Area.

This document does not constitute professional or legal advice. For specific guidance related to your business needs, please consult with appropriate qualified professionals.