

Moonshot releases Kimi K3, a 2.8-trillion-parameter open-weights model

Source copy kept by wheresthe.ai for the story "Moonshot releases Kimi K3, a 2.8-trillion-parameter open-weights model" (<https://wheresthe.ai/d/cn-moonshot-kimi-k3-release-2026-07/>).

Source: <https://pureai.com/articles/2026/07/17/china-moonshot-ai-releases-kimi-k3.aspx>

Text saved 18 Oct 2026 (the page as published).

SHA-256 of the saved text: fd4e0c31bc537e2e8d04ddcdeb0ab06cdc51551af97064b4e42c882b90fa6d3

China's Moonshot AI Releases Kimi K3, Its Largest Open-Weight Model -- Pure AI

[Skip to main content](#)

[Sections](#) [News](#) [Features](#) [Insights](#) [Human in the Loop](#) [Mind the Prompt](#) [The Pure AI Blog](#) [The Long View](#) [AI Castaway](#) [AI Watch](#)

[How-To](#) [AI Tracker](#)

[Events](#) [Videos](#) [Webcasts](#) [White Papers](#) [Training](#) [3-Hour Agent Workshop](#) [Live!](#) [360 Technical Training](#)

[Advertise](#) [Newsletters](#) [Glossary](#) [Contact Us](#)

[Expert Insights for the AI Business Revolution](#)

[Add as a preferred source on Google](#)

[News](#)

China's Moonshot AI Releases Kimi K3, Its Largest Open-Weight Model

By John K. Waters 07/17/2026

Key Takeaways

Chinese AI company Moonshot AI has released Kimi K3, a 2.8 trillion-parameter open-weight model, positioning it as the largest open-weight model released to date.

An independent benchmark from Artificial Analysis ranks K3 fourth among 189 tested models, behind Claude Fable 5 and two configurations of GPT-5.6 Sol, and ahead of Claude Opus 4.8 and GPT-5.5.

Analysts and Chinese officials have tied the open-weight strategy to export controls on advanced chips, a bid to build global developer adoption, and China's push to shape international AI standards.

Chinese artificial intelligence startup Moonshot AI released Kimi K3 this week, an open-weight model with 2.8 trillion total parameters that the Beijing-based company describes as achieving "open frontier intelligence." The release makes K3 the largest open-weight model released to date, exceeding DeepSeek's V4 Pro, at 1.6 trillion parameters, and Zhipu AI's GLM 5 series at 744 billion parameters.

K3 uses a "mixture-of-experts" architecture (MoE), an AI model design that contains many specialized neural networks, called "experts," but activates only a small number of them for each request. It also introduces two techniques developed internally at Moonshot, Kimi Delta Attention (KDA) and Attention Residuals, which the company says improve computing efficiency and reasoning quality. The model supports a 1-million-token context window and processes both text and images natively. Moonshot said K3 is built for long-horizon coding, knowledge work, and reasoning tasks that require sustained, multistep execution with limited human oversight.

An independent evaluation by Artificial Analysis, a third-party AI benchmarking firm, gives K3 a score of 57 on its Intelligence Index, ranking it fourth among 189 models tested. That places K3 behind Claude Fable 5 and two configurations of OpenAI's GPT-5.6 Sol, and ahead of Anthropic's Claude Opus 4.8, OpenAI's GPT-5.5, Claude Sonnet 5, and GLM-5.2. Moonshot's own launch materials report additional coding benchmark results, including a first-place finish on two internal coding tests, though those figures have not yet been independently reproduced using a common testing harness.

In its release announcement, Moonshot said, "K3 stands as Moonshot AI's most powerful open-source coding model to date." The company also said K3 performed competitively against Claude Fable 5, the most advanced model widely available from Anthropic, while outperforming Claude Opus 4.8 and OpenAI's GPT-5.6 Sol and GPT-5.5 on its benchmark set.

Kimi K3's API is live now through Moonshot's platform, the Kimi consumer app, Kimi Work , and the Kimi Code coding tool, with additional access available through OpenRouter . The API is compatible with the OpenAI software development kit, which lowers the integration barrier for developers already building on OpenAI or Anthropic tools. API pricing is set at \$3 per million input tokens and \$15 per million output tokens, the highest rate among Chinese AI labs currently, though still below the per-task cost of Anthropic's Opus 4.8, according to Moonshot. Full model weights are scheduled for release on July 27.

Moonshot AI was founded in 2023 by Tsinghua University graduate Yang Zhilin and is backed by Alibaba and Tencent . Its Kimi chatbot ranked third among Chinese AI products by monthly active users before DeepSeek's R1 model disrupted the domestic market in January 2025, after which Kimi's ranking fell to seventh. Moonshot's shift toward open-weight releases began with Kimi K2 in July 2025 and accelerated with K2.5 in January 2026, a strategy the company has used to rebuild its position in China's AI market.

Moonshot's approach reflects a broader pattern among Chinese AI developers, including DeepSeek, Alibaba, Tencent, and Baidu , all of which have released open-weight models over the past year. Analysts have offered several explanations for the strategy. Bloomberg has reported that Chinese firms, facing U.S. restrictions on access to the most advanced AI chips and unable to match the capital spending of American rivals, are using open-weight releases to speed AI adoption across their domestic economy. Reuters has reported that open-sourcing allows Chinese companies to demonstrate technical capability and expand their developer communities and global influence, a strategy analysts say could help China counter U.S. efforts to limit its technological progress. Separately, MIT Technology Review has reported that Chinese firms increasingly view open-weight releases as the fastest way to close the capability gap with U.S. labs by attracting developers, driving adoption and helping set technical standards.

The release came one day ahead of the 2026 World Artificial Intelligence Conference in Shanghai, which opened July 17 with a keynote address from Chinese President Xi Jinping.

About the Author

John K. Waters is the editor in chief of a number of Converge360.com sites, with a focus on high-end development, AI and future tech. He's been writing about cutting-edge technologies and culture of Silicon Valley for more than two decades, and he's written more than a dozen books. He also co-scripted the documentary film Silicon Valley: A 100 Year Renaissance , which aired on PBS. He can be reached at .

Featured

When an AI Companion Goes Bankrupt, What Happens to Its Users' Secrets?

Nearly Half of Organizations Report Unexpected AI Agent Activity, Study Finds

Google Limits Initial Gemini 4 Argon Access to Cyber Defenders

Shadow AI Is Becoming a Security Blind Spot for Enterprises, Sophos Warns

Most Popular

Most Popular Articles

Most Emailed Articles

Nearly Half of Organizations Report Unexpected AI Agent Activity, Study Finds

When an AI Companion Goes Bankrupt, What Happens to Its Users' Secrets?

BambooHR Adds Claude Connector for HR Data and Workflows

Google Limits Initial Gemini 4 Argon Access to Cyber Defenders

Shadow AI Is Becoming a Security Blind Spot for Enterprises, Sophos Warns

Pure AI

Sign up for our newsletter.

Email Address* * * Country*

United States of America Afghanistan Åland Islands Albania Algeria American Samoa Andorra Angola Anguilla Antarctica Antigua and Barbuda Argentina Armenia Aruba Australia Azerbaijan Austria Bahamas Bahrain Bangladesh Barbados Belarus Belgium Belize Benin Bermuda Bhutan Bolivia, Plurinational State of Bonaire, Sint Eustatius and Saba Bosnia and Herzegovina Botswana Bouvet Island Brazil British Indian Ocean Territory Brunei Darussalam Bulgaria Burkina Faso Burundi Cambodia Cameroon Canada Cape Verde (Cabo Verde) Cayman Islands Curaçao Central African Republic Chad Chile China Christmas Island Cocos (Keeling) Islands Colombia Comoros Congo Congo, the Democratic Republic of the Cook Islands Costa Rica Côte d'Ivoire Croatia Cuba Cyprus Czech Republic Denmark Djibouti Dominica Dominican Republic Ecuador Egypt El Salvador Equatorial Guinea Eritrea Estonia Ethiopia Falkland Islands (Malvinas) Faroe Islands Fiji Finland France French Guiana French Polynesia French Southern Territories Gabon Gambia Georgia Germany Ghana Gibraltar Greece Greenland Grenada Guadeloupe Guam Guatemala Guernsey Guinea Guinea-Bissau Guyana Haiti Heard Island and McDonald Islands Holy See (Vatican City State) Honduras Hong Kong Hungary Iceland India Indonesia Iran, Islamic Republic of Iraq Ireland Isle of Man Israel Italy Jamaica Japan Jersey Jordan Kazakhstan Kenya Kiribati Korea, Democratic People's Republic of Korea, Republic of Kuwait Kyrgyzstan Lao People's Democratic Republic Latvia Lebanon Lesotho Liberia Libya Liechtenstein Lithuania Luxembourg Macao Macedonia, the former Yugoslav Republic of Madagascar Malawi Malaysia Maldives Mali Malta Marshall Islands Martinique Mauritania Mauritius Mayotte Mexico Micronesia, Federated States of Moldova, Republic of Monaco Mongolia Montenegro Montserrat Morocco Mozambique Myanmar Namibia Nauru Nepal Netherlands New Caledonia New Zealand Nicaragua Niger Nigeria Niue Norfolk Island Northern Mariana Islands Norway Pakistan Oman Palau Palestinian Territory, Occupied Panama Paraguay Papua New Guinea Peru Philippines Pitcairn Poland Portugal Puerto Rico Qatar Réunion Romania Russian Federation Rwanda Saint Barthélemy Saint Helena, Ascension and Tristan da Cunha Saint Kitts and Nevis Saint Lucia Saint Martin (French part) Saint Pierre and Miquelon Saint Vincent and the Grenadines Samoa San Marino Sao Tome and Principe Saudi Arabia Senegal Serbia Seychelles Sierra Leone Singapore Sint Maarten (Dutch part) Slovakia Slovenia Solomon Islands Somalia South Africa South Georgia and the South Sandwich Islands South Sudan Spain Sri Lanka Sudan Suriname Svalbard and Jan Mayen Eswatini (Swaziland) Sweden Switzerland Syrian Arab Republic Taiwan, Province of China Tajikistan Tanzania, United Republic of Thailand Timor-Leste Togo Tokelau Tonga Trinidad and Tobago Tunisia Turkey Turkmenistan Turks and Caicos Islands Tuvalu Uganda Ukraine United Arab Emirates United Kingdom United States Minor Outlying Islands Uruguay Uzbekistan Vanuatu Viet Nam Venezuela, Bolivarian Republic of Virgin Islands, British Virgin Islands, U.S. Wallis and Futuna Western Sahara Yemen Zambia Zimbabwe

*

I agree to this site's Privacy Policy

Please type the letters/numbers you see above.

Most Popular Articles

Most Emailed Articles

Nearly Half of Organizations Report Unexpected AI Agent Activity, Study Finds

When an AI Companion Goes Bankrupt, What Happens to Its Users' Secrets?

BambooHR Adds Claude Connector for HR Data and Workflows

Google Limits Initial Gemini 4 Argon Access to Cyber Defenders

Shadow AI Is Becoming a Security Blind Spot for Enterprises, Sophos Warns

Upcoming Training Events

0

AM

Live! 360 6-Week Training & Certification Course: Mastering the Microsoft AI Framework: Building Enterprise-

Ready AI Agents with Microsoft Foundry October 6–November 10, 2026

Live! 360 Orlando November 15-20, 2026

Artificial Intelligence Live! Orlando November 15-20, 2026

AI Enterprise Architecture Live! Orlando November 15-20, 2026

Cybersecurity & Ransomware Live! Orlando November 15-20, 2026

Data Platform Live! Orlando November 15-20, 2026

TechMentor Orlando November 15-20, 2026

Live! 360 2-Day Hands-On Seminar: AI-Powered .NET Development with Claude & Claude Code December 8-9, 2026

Pure AI 3-Hour Virtual Training Workshop: AI Agents for Everyday Business - Automate Your Work Without Coding December 9, 2026

Virtual Hands-on Training Seminar: AI-Powered PowerShell and Infrastructure Automation with Claude Code December 10-11, 2026

TechMentor & Cybersecurity Live! @ Microsoft HQ August 9-13, 2027