

Meituan unveils LongCat-2.0, the model behind the stealth ‘Owl Alpha’ on OpenRouter

Source copy kept by wheresthe.ai for the story “Meituan unveils LongCat-2.0, the model behind the stealth ‘Owl Alpha’ on OpenRouter” (<https://wheresthe.ai/d/cn-meituan-longcat-2-open-2026-06/>).

Source: <https://venturebeat.com/technology/meituan-open-sources-longcat-2-0-the-1-6t-near-frontier-agentic-coding-model-thats-been-leading-openrouter-trained-entirely-on-chinese-chips>

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

SHA-256 of the saved text: c82140047dd723132fb0e0834d6d37d063f4f5ceca97fbf22ef345f72e7316f7

Featured

Carl Franzen

10:39 pm, PT, June 29, 2026

Credit: VentureBeat made with OpenAI GPT-Images-2.0

A few hours ago, Chinese delivery app company Meituan officially unveiled LongCat-2.0 on GitHub , Hugging Face , and its native platform, unmasking the model as the computational engine behind "Owl Alpha," the anonymous stealth model that has spent the last two months commanding global developer charts on OpenRouter. Developed to fundamentally disrupt closed-source enterprise dominance in autonomous software engineering, the 1.6-trillion-parameter Mixture-of-Experts (MoE) system brings a native 1-million-token context window to the public domain under a highly permissive, enterprise grade, commercially viable MIT license. However, the company has yet to post the full weights — both the Github and Hugging Face pages say "Model weights coming soon — stay tuned!"

Commercial access to the architecture introduces a highly aggressive pricing tier, deploying a mechanism where all context-cache hits are processed completely free of charge , running alongside a time-limited " Token Pack " flash-sale paradigm. There's also a typical "pay-as-you-go" API for non-cache hits standard priced at \$0.75/\$2.95 per million tokens in/out. However, a limited-time promotional discount aggressively slashes these operational expenditures down to \$0.30 per million tokens for uncached input and \$1.20 per million tokens for output, both on the cheaper-end of top performing models globally.

Model	Input (\$/1M)	Output (\$/1M)	Total (\$/1M)	Source
MiMo-V2.5 Flash	\$0.10	\$0.30	\$0.40	Xiaomi
deepseek-v4-flash	\$0.14	\$0.28	\$0.42	DeepSeek
deepseek-v4-pro	\$0.435	\$0.87	\$1.305	DeepSeek
MiniMax-M3	\$0.30	\$1.20	\$1.50	MiniMax
LongCat-2.0 — limited-time promo	\$0.30	\$1.20	\$1.50	LongCat
Gemini 3.1 Flash-Lite	\$0.25	\$1.50	\$1.75	Google
Qwen3.7-Plus	\$0.40	\$1.60	\$2.00	Alibaba Cloud
MiMo-V2.5	\$0.40	\$2.00	\$2.40	Xiaomi
LongCat-2.0 — standard	\$0.75	\$2.95	\$3.70	LongCat
Grok 4.3 (low context)	\$1.25	\$2.50	\$3.75	xAI
MiMo-V2.5 Pro (≤256K)	\$1.00	\$3.00	\$4.00	Xiaomi
Kimi-K2.6	\$0.95	\$4.00	\$4.95	Moonshot AI
GLM-5.2	\$1.40	\$4.40	\$5.80	Z.ai
GPT-5.6 Luna	\$1.00	\$6.00	\$7.00	OpenAI

Grok 4.3 (high context) \$2.50 \$5.00 \$7.50 xAI

MiMo-V2.5 Pro (>256K) \$2.00 \$6.00 \$8.00 Xiaomi

Qwen3.7-Max \$2.50 \$7.50 \$10.00 Alibaba Cloud

Gemini 3.5 Flash \$1.50 \$9.00 \$10.50 Google

Gemini 3.1 Pro Preview ($\leq$ 200K) \$2.00 \$12.00 \$14.00 Google

GPT-5.6 Terra \$2.50 \$15.00 \$17.50 OpenAI

GPT-5.4 \$2.50 \$15.00 \$17.50 OpenAI

Gemini 3.1 Pro Preview (>200K) \$4.00 \$18.00 \$22.00 Google

Claude Opus 4.8 \$5.00 \$25.00 \$30.00 Anthropic

GPT-5.5 \$5.00 \$30.00 \$35.00 OpenAI

GPT-5.5 Instant (chat-latest) \$5.00 \$30.00 \$35.00 OpenAI

Sakana Fugu Ultra ($\leq$ 272K) \$5.00 \$30.00 \$35.00 Sakana AI

GPT-5.6 Sol \$5.00 \$30.00 \$35.00 OpenAI

Claude Fable 5 / Claude Mythos 5 \$10.00 \$50.00 \$60.00 Anthropic

What makes the release a definitive inflection point for global tech infrastructure is its operational independence: the massive model was trained entirely on a cluster of over 50,000 domestic Chinese Application-Specific Integrated Circuits (ASICs), proving that near-frontier AI models can be scaled successfully without relying on the typical U.S. Nvidia GPUs that have, to date, powered much of the global generative AI frontier model training effort. This successful deployment of alternative silicon signals a profound structural shift. If Chinese conglomerates can consistently iterate trillion-parameter architectures using homegrown ASICs rather than general-purpose GPUs, it would seem to threaten Nvidia's dominance in this sector. Crucially, this technological pivot arrives precisely as Washington pressures top-tier American labs to restrict access to their latest models. Following a U.S. governmental request, OpenAI was forced to limit access to its new GPT-5.6 models, while Anthropic was previously also ordered by the U.S. to restrict access to its latest Claude Fable 5 / Mythos 5 models, which it took entirely offline in response. At the same time, a growing chorus of technologists, activists, and industry experts warn that these defensive regulatory maneuvers have inadvertently backfired. By locking down Western closed-source models and driving up API costs, the U.S. government has left a wide operational window for global developers seeking affordable, high-performance alternatives like those found in Chinese open source models such as Meituan LongCat-2.0. The raw operational metrics backed up the developer enthusiasm: during its unbranded residency on OpenRouter, Owl Alpha accounted for approximately 10.1 trillion monthly tokens—averaging 559 billion tokens per day—representing a 242% month-over-month explosion in volume that propelled it into the platform's global top three. By the time Meituan stepped forward to claim the architecture, the model had already secured the top ranking on the Hermes Agent workspace, second place on Claude Code deployments, and third place across international OpenClaw environments. VB Transform · July 14–15 · Menlo Park · Inference & AI infrastructure GM got a 300% jump in merged PRs by rearchitecting for agents. Here's what they built. The infrastructure track at Transform covers real-time video generation, machine-to-machine reasoning stacks, and what it actually takes to run agents at enterprise scale. See the full agenda →

Technology: Engineering the 1M-Token Sparse Context At the core of LongCat-2.0 lies an aggressive optimization of Mixture-of-Experts (MoE) sparsity, scaling total parameters to 1.6 trillion while limiting active computation to an average of 48 billion parameters per token. Depending on the structural complexity of a query, the model's dynamic activation ranges from 33 billion to 56 billion parameters. This design implements a "Zero-Compute Experts" framework, ensuring that routine execution elements pass through lighter subnetworks, entirely eliminating the idle computational overhead that typically penalizes ultra-dense models. To sustain a functional 1-million-token context window without incurring catastrophic hardware bottlenecks, Meituan introduced LongCat Sparse Attention (LSA). Designed as an evolutionary iteration of DeepSeek Sparse Attention, LSA resolves the quadratic scoring costs and memory fragmentation that typically plague fine-grained sparse mechanisms through three distinct, orthogonal vectors: Streaming-aware Indexing (SI): This system restructures the token selection pipeline by blending hardware-aligned contiguous data reads with dynamic random selection. By converting

fragmented memory access into highly predictable, sequential blocks, the system achieves coalesced High Bandwidth Memory (HBM) utilization and elevated effective bandwidth.

Cross-Layer Indexing (CLI): Leveraging the empirical reality that attention saliency remains highly stable across adjacent hidden layers, CLI amortizes calculation costs. A single indexing pass successfully guides multiple consecutive layers during inference, a capability reinforced by cross-layer distillation throughout the training phase.

Hierarchical Indexing (HI): This approach applies a coarse-to-fine, two-stage scoring layout. The indexer performs a rapid, approximate block-level recall to filter candidates, before running fine-grained token selection exclusively on the remaining population.

Furthermore, Meituan integrated an N-gram Embedding module inherited from its lighter model lines. By expanding parameter allocation in sparse dimensions completely orthogonal to the MoE expert layout, the architecture appends 135 billion parameters to a 5-gram token combination framework. This expands the core embedding space by roughly 100-fold, allowing the model to capture dense local token relationships and accelerate large-batch inference operations by reducing memory Input/Output (I/O) bottlenecks.

Product: Post-Training, MOPD Framework and Benchmark Performance While generalist large language models prioritize fluid, conversational interfaces, LongCat-2.0 focuses explicitly on multi-step engineering tasks, tool integration, and automated repository manipulation — agentic tasks, in other words. In standardized assessments, LongCat-2.0 registers an empirical 59.5 on SWE-bench Pro, surpassing GPT-5.5's benchmark of 58.6. The model further establishes its agentic specialization by marking a 70.8 on Terminal-Bench 2.1, a 77.3 on SWE-bench Multilingual, and a 73.2 on the general corporate workflow simulator FORTE. LongCat-2.0 benchmark comparison bar charts. Credit: Meituan LongCat

This precise operational behavior is achieved through a structural post-training layer called Multi-Teacher Optimization via Mixture of Specialized Experts (MOPD). Rather than blending raw human feedback into a singular reward function, the MOPD architecture segregates post-training optimization into three independent, highly focused expert clusters. The Agent Experts are fine-tuned strictly for structural execution, specializing in precise tool invocation, multi-turn API parameter parsing, and self-correcting loop mechanisms to avoid execution stagnation.

The Reasoning Experts are optimized in isolation to advance multi-hop logic, complex chain-of-thought engineering, mathematics, and high-level STEM problem-solving.

The Interaction Experts focus entirely on human alignment, instruction-following nuances, factual grounding to suppress hallucinations, and maintaining rigid safety guardrails without diminishing the model's overall utility.

By segregating these vectors during post-training, LongCat-2.0 prevents functional degradation. A dynamic gate-routing mechanism then seamlessly fuses these specialized behaviors at runtime, allowing the final model to coordinate deep reasoning, stable tool execution, and safe user interaction simultaneously. While LongCat-2.0 generally trails premium frontier systems like Claude Opus 4.8 across broad general-agent benchmarks such as FORTE and BrowseComp, it explicitly punches above its weight in software engineering. What makes this open-weight architecture special is its hyper-focus on autonomous development; it manages to narrowly exceed OpenAI's proprietary GPT-5.5 on the rigorous software engineering benchmark SWE-bench Pro (scoring 59.5 against 58.6), proving it is highly capable and fiercely competitive for complex coding tasks despite a leaner computational footprint. Commercial Framework: Pay-As-You-Go vs. Flash-Sale Token Packs Meituan's deployment strategy introduces a specialized commercial model that splits network access between conventional real-time API billing and structured "Token Packs". For traditional enterprise integration, standard top-up accounts are available, deducting operational capital in real time based directly on token input and generation metrics. However, to accommodate the unpredictable compute bursts characteristic of autonomous development agents, Meituan launched a structured Token Pack framework. Purchased as fixed, one-time volumetric allocations valid for a strict 30-day window, these packages stack directly on top of an organization's existing baseline API account. To manage network load across its ASIC clusters, Meituan releases these high-volume packages via limited flash sales four times daily, precisely at 10:00, 16:00, 21:00, and 23:00 Beijing Time on a first-come, first-served basis. The economic standout of this framework is the zero-charge processing of context cache hits. In massive agentic environments where a coding assistant must repeatedly read, reference, and modify the same multi-million-token code repository over an extended session, standard architectures penalize developers by charging full pricing for repeated input context. Under Meituan's infrastructure, only cache-miss inputs and final

token generations consume the package quota. This architecture completely alters the operational cost economics of large-scale agent software development, enabling deep iterative context exploration without compounding costs.

Licensing: Open-Source Structural Freedom By registering the LongCat-2.0 repository under the open-source MIT License, Meituan positions the architecture with maximum legal flexibility for enterprise integration. In contrast to copyleft paradigms like the GNU General Public License (GPL)—which legally obligates developers to open-source any derivative frameworks or internal software that links to the code—the MIT license permits near-unrestricted freedom. For corporate engineering teams, this legal standard ensures that LongCat-2.0 can be deeply modified, compiled, and hard-coded directly into closed-source commercial applications, proprietary dev tools, and internal automation backends. Corporations can fork the repository, optimize the internal LSA mechanisms for private databases, and sell the resulting software stack to end users without any obligation to disclose their proprietary intellectual property or structural enhancements.

Meituan's Evolution: From Delivery Super App to AI Powerhouse Founded in March 2010 by serial entrepreneur Wang Xing, Meituan initially launched as a Groupon-style daily deals website before rapidly evolving into one of China's dominant "super apps". Following a massive 2015 merger with Dianping, the Beijing-based tech giant solidified a dominant market share over the country's urban delivery corridors, bridging local consumer reviews, instant retail, hotel bookings, and food delivery. Operating as a publicly traded powerhouse on the Hong Kong Stock Exchange, Meituan claims over 770 million annual transacting users and supports a network of more than 14.5 million merchants. However, faced with intense domestic market competition, severe margin compression, and a sliding profit margin, the company aggressively pivoted its strategy beyond logistics. Meituan publicly committed to investing "billions" into artificial intelligence and domestic chip capabilities to revitalize its technology-driven offerings. This strategic shift into the global AI race began materializing in late 2025 with the release of LongCat-Flash, a 560-billion-parameter Mixture-of-Experts foundation model, followed quickly by the advanced reasoning model LongCat-Flash-Thinking. By open-sourcing these frontier-class models under enterprise-friendly licenses, Meituan signaled its ambition to become a foundational player in global AI infrastructure rather than remaining strictly a regional e-commerce and delivery giant.

Enterprise Implications: Autonomous Operational Workflows For modern enterprises, the release of LongCat-2.0 unlocks clear operational strategies across software engineering, system operations, and long-form data interpretation. The combination of an open-weight, MIT-licensed model with an expansive 1-million-token context window means organizations can bypass the data privacy concerns and recurring overhead associated with hosting proprietary third-party APIs. In large-scale enterprise development environments, teams can leverage the model's specialized Agent Experts to orchestrate autonomous codebase migrations. Instead of dedicating hundreds of developer hours to manually rewriting legacy application frameworks, engineers can pass an entire enterprise repository along with modern SDK documentation directly into the 1-million-token context window. LongCat-2.0 can map the dependencies, execute the repository-level structural updates, compile the new codebase, and catch compilation and execution bugs autonomously within local sandbox environments before generating a final pull request. The model's architectural separation via the MOPD gate-routing mechanism yields significant advantages for strict enterprise compliance. By routing specific operational queries through isolated expert clusters, a financial institution or healthcare firm can deploy deep logic and mathematical reasoning passes without risking factual hallucination or violating strict safety bounds. The Interaction Experts function as an implicit guardrail layer, suppressing errors and enforcing instruction-following protocols without degrading the raw processing power of the internal Reasoning Experts. Combined with the zero-cost caching model, enterprises can maintain hyper-focused autonomous software networks that can repeatedly inspect corporate data pools, continuously maintaining and optimizing internal infrastructure at a fraction of standard operational costs.

More