long cat llm

LongCat is a recently released, open-source family of large language models (LLMs) from Meituan that target ultra-efficient and high-performance agentic tasks. The most prominent models are LongCat-Flash and LongCat-Flash-Thinking, featuring a Mixture-of-Experts (MoE) architecture with 560 billion total parameters, but activating only 18.6B–31.3B parameters per token on-demand, which keeps inference highly efficient while scaling to advanced, agentic reasoning abilities.[1] [2] [3] [4] [5]

Key Features

- **MoE Architecture**: Dynamically activates a subset of parameters for computation efficiency; on average only ~27B parameters are used per token, optimizing hardware usage and cost.[2] [5] [1]
- **Zero-Computation Experts**: Enables strategic allocation of computational resources for context-dependent reasoning.[3] [2]
- **High Throughput**: Achieves over 100 tokens per second (TPS) with a total cost of about $0.70 per million output tokens, even at massive scale (more than 20 trillion tokens trained in 30 days).[5] [2]
- **Open Source**: Released under the MIT License, available via Hugging Face and GitHub, supporting transfer learning, distillation, and community research.[4] [3]
- **Agentic Reasoning**: Specialized post-training in reasoning, instruction following, code, and multi-domain tool-use, performing highly in agentic (decision or planning) tasks.[3] [4]

Use Cases

- Complex multi-turn reasoning and dialogue
- Agent-based or multi-agent conversational AI
- High-performance QA and enterprise solutions
- Large-scale mathematical, logical, and programming tasks[4] [3]

Technical Highlights

Model	Total Parameters	Active per Token	Key Strengths	License
LongCat-Flash	560B	18.6B–31.3B	Efficient agentic reasoning, MoE design	MIT [1] [2] [3]

Model	Total Parameters	Active per Token	Key Strengths	License
LongCat-Flash-Thinking	560B	18.6B–31.3B	Formal, agentic, and chain-of-thought	MIT [4]

LongCat represents a state-of-the-art direction in scalable LLM architectures, highly focused on computational efficiency and agentic capabilities for advanced tasks. [1] [2] [3] [4]

⁂

1. https://huggingface.co/meituan-longcat/LongCat-Flash-Chat
2. https://arxiv.org/html/2509.01322v1
3. https://jimmysong.io/en/ai/longcat-flash-chat/
4. https://github.com/meituan-longcat/LongCat-Flash-Thinking
5. https://arxiv.org/abs/2509.01322
6. https://www.youtube.com/watch?v=uJ1wdEjDGck
7. https://github.com/meituan-longcat
8. https://www.digitalocean.com/community/tutorials/longcat-flash-chat-2025
9. https://huggingface.co/meituan-longcat/LongCat-Video
10. https://lmsys.org/blog/2025-09-01-sglang-longcat-flash/
11. https://www.youtube.com/watch?v=PvVGHf7LQ6w
12. https://www.reddit.com/r/StableDiffusion/comments/1ofnqvy/meituan_longcatvideo_mit_license_foundation_video/
13. https://www.youtube.com/watch?v=wNU19WORjfM
14. https://arxiv.org/abs/2510.22200
15. https://crepal.ai/blog/aivideo/longcat-video-meituan/
16. https://www.linkedin.com/posts/pasha-shaik_github-meituan-longcatlongcat-video-activity-7387859717408223232-f46i
17. https://longcat.chat
18. https://openrouter.ai/meituan/longcat-flash-chat:free
19. https://chatpaper.com/chatpaper/paper/204129
20. https://www.reddit.com/r/LocalLLaMA/comments/1n46mk9/longcatflashchat_is_here_yet_another_chinese_open/