

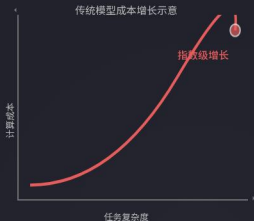
# MiniMax-M1: 开启高效推理新纪元

一款为复杂任务而生的超高效、长上下文语言模型

# 我们面临的挑战： 大模型的"思考成本"

## 为什么我们需要M1？

- 1 现实世界的任务日益复杂（如分析财报、解决软件bug），需要模型进行更深入、更长时间的思考。
- 2 传统大模型"想得越久，成本越高"，计算力的瓶颈限制了其应用深度。
- 3 急需一种既能深度思考，又能控制成本的全新解决方案。



# 我们的解决方案：MiniMax-M1

M1是什么？

全球首个开源、大规模、专为高效深度推理打造的语言模型

🕒 想得更久

**8** 万字

超长输出：支持8万字的思考与输出

👁️ 看得更全

**100** 万字

超长输入：能处理100万字的上下文  
信息

⚡ 算得更快

**1/4** 成本

超高效率：计算成本仅为同类模型的  
1/4

# 揭秘M1的“超级引擎”：闪电注意力

M1为何如此高效？

传统注意力



告别“暴力计算”：传统模型处理长文本就像地毯式搜索，效率低下。

# 75%

FLOPs 降低

闪电注意力



M1的智慧：“闪电注意力”像一个智能导航，用线性复杂度快速定位关键信息。

vs DeepSeek R1, 100K token 生成 FLOPs 降低 **75%**

# M1的"专家大脑": 混合专家模型

M1如何兼顾"博学"与"专注"?



**类比：** M1的大脑不是一个"万事通"，而是一个高效的"专家顾问团"。

**工作方式：** 遇到问题，路由网络只派出最相关的几位"专家"协作解决，其他专家则待命，节省能耗。

优势：海量知识，轻量激活

~~4560亿~~ → **459亿**

# 独特的"学习方法": CISPO强化学习算法

M1如何学会像人一样深度推理？

传统PPO算法：通过"试错"学习，但部分尝试被浪费

✗ 部分数据被丢弃



2倍

训练速度提升

CISPO创新：不放弃任何一次"尝试"，训练更稳定高效



# 性能展示: M1在真实世界中的表现

M1的实战能力如何？

**综合实力：**在数学、编程等核心能力上，与全球顶尖模型并驾齐驱

## <> 软件工程

解决真实GitHub问题的能力，展现卓越的工程实践水平。

M1 模型

领先

顶尖竞品

## 🛠️ 工具使用 (Agent)

作为智能体调用工具的能力，高效完成复杂任务。

M1 模型

超越

Gemini 2.5 Pro

## ≡ 长文本理解

在海量信息中精准提炼关键内容，阅读理解能力出众。

M1 模型

超越

OpenAI o3 & Claude 4

# 思考得更久，结果会更好

为什么80K比40K更强？

## 40K (标准版)

我们训练了模型的标准版本，在数学竞赛等复杂任务上表现出色。

**75.2%**

数学竞赛准确率

## 80K (深度思考版)

通过扩展训练，模型准确率获得持续且显著的提升。

**85.1%**

数学竞赛准确率 (+9.9%)

推理步数 (Inference Steps)


这证实了M1的核心设计理念：更长的推理过程能带来更优的解决方案。



# 总结与展望

## M1的价值与未来

**核心价值：** MiniMax-M1为解决需要深度思考和长文本处理的复杂现实世界问题，提供了一个高性价比且性能强大的开源解决方案。

**开放性：**  我们已将模型完全开源，并提供API，欢迎社区共同推动技术发展，构建开放共赢的生态。

**未来方向：** M1将成为下一代智能体 (Agent) 的坚实基础，赋能于以下前沿领域



自动化办公



科学发现



智能助理



教育培训

# 感谢聆听

欢迎提问与交流



model@minimax.io



<https://github.com/MiniMax-AI/MiniMax-M1>



报告日期：2025年6月19日

MiniMax Agent 制作