Tencent Hunyuan

2025-05-20

Hunyuan-TurboS: Advancing Large Language Models through Mamba-Transformer Synergy and Adaptive Chain-of-Thought

Tencent Hunyuan Team

LMSYS CHATBOT ARENA

排名前8的腾讯混元 TURBOS模型技术报告

简介

HUNYUAN-TURBOS 是业界首个部署的大规模 MAMBA2 + GQA ATTENTION + MOE 混合LLM模型

非开源模型

128层　560B-A56B

预训练

16T tokens

三阶段预训练：

- 正常预训练
- 退火阶段训练（Annealing），快速降低学习率，对模型做精细调优
- 长上下文扩展（Long-Context Extension），先 4K → 32K，再 32K → 256K

后训练

四阶段后训练

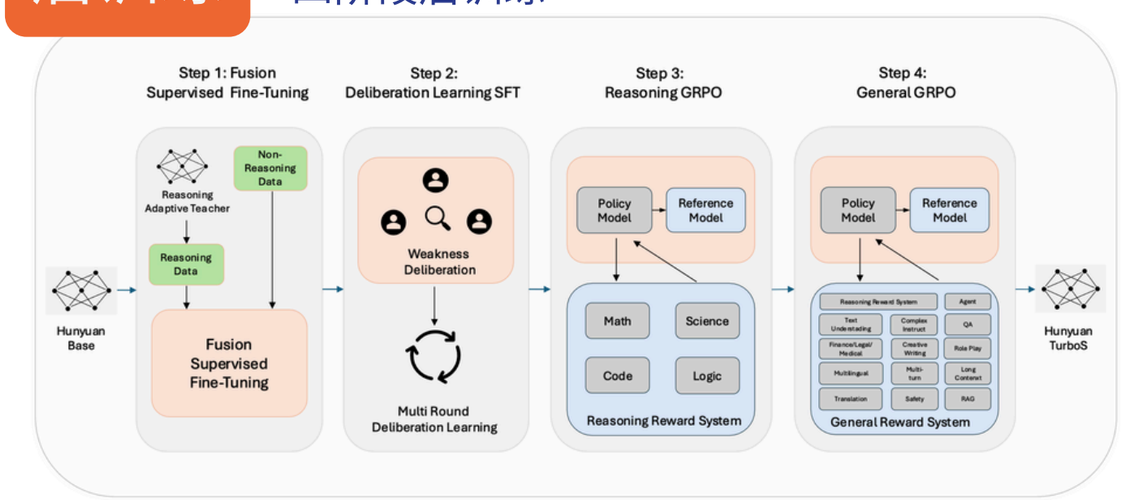

Figure 3: A diagram illustrating the four steps of Hunyuan-TurboS post-training.

效果对比

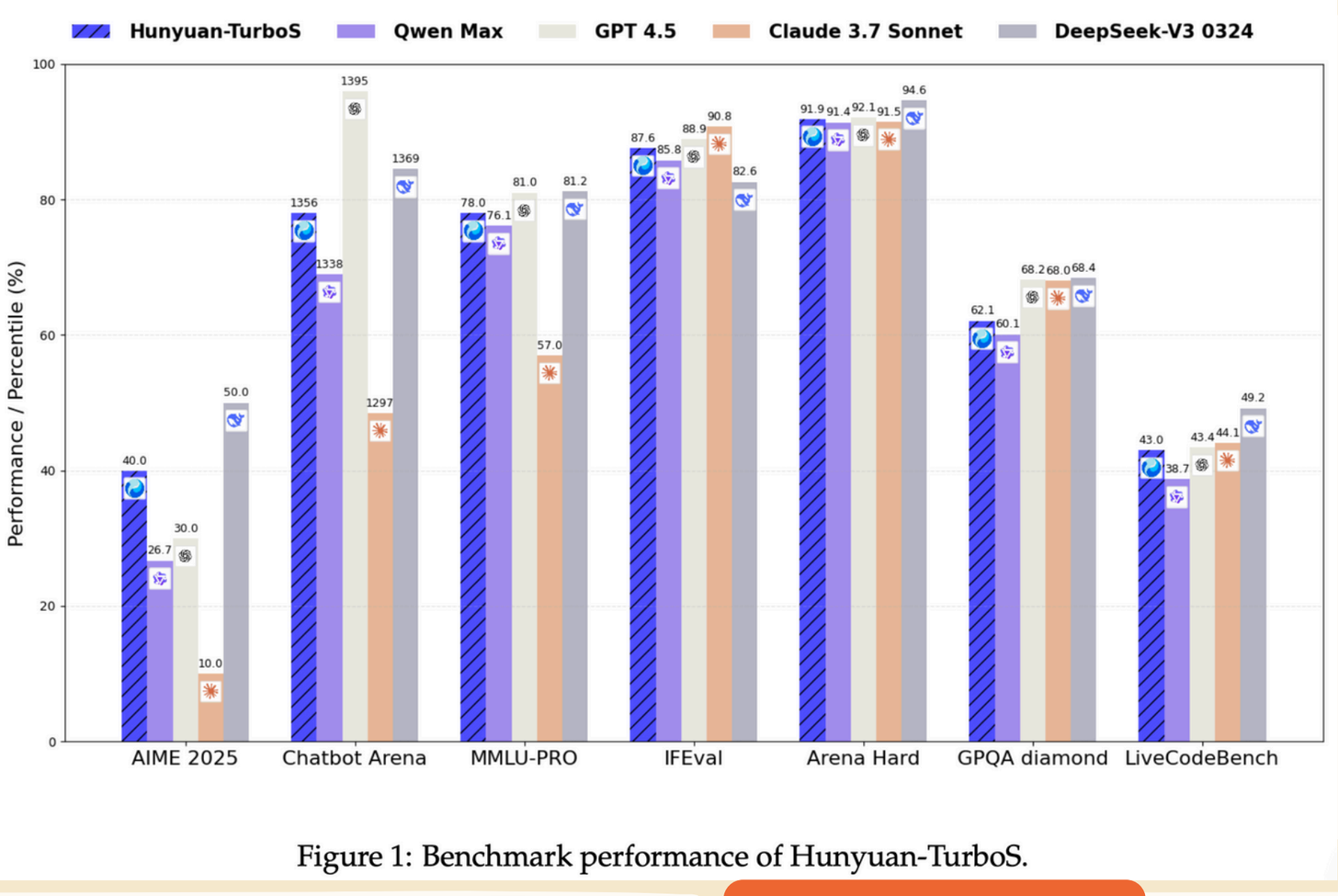

Figure 1: Benchmark performance of Hunyuan-TurboS.

架构细节

Table 1: Overview of the key hyper-parameters of Hunyuan-TurboS.

Configuration	Value
# Layers	128
# Attention Heads	64
# Key/Value Heads	8
# Mamba2 SSM Groups	16
# Shared Experts of MoE Layers	1
# Specialized Experts of MoE Layers	32
# Activated Specialized Experts of MoE Layers	2
# Trained Tokens	16T
Mamba2 d_state size	128
Mamba2 chunk size	128
Vocabulary Size	128K
Hidden Size	5120

CONCOLUSION

名副其实的混合模型结构😂

后训练中的自适应LONG/SHORT COT融合挺有意思的，让模型学习何时用LONG COT何时用SHORT COT

@机器爱学习