

## LARGE LANGUAGE MODEL HIGHLIGHT

# MAI-Thinking-1

*Microsoft · USA · Proprietary · Announced June 2026*

## TOTAL PARAMETERS

**1T***35B active, MoE*

## TRAINING TOKENS

**33.5T***Web-scale*

## DATA RATIO

**34:1***Near-Chinchilla*

## TRAINING COST

**~\$11M***Hopper hardware basis***SUMMARY**

MAI-Thinking-1 by Microsoft (USA) is a Mixture-of-Experts model with 1T total parameters, 35B active parameters per token, trained on 33.5T tokens seen, for a training-token-to-total-param data ratio of 34:1, placing it in Near-Chinchilla territory.

On a normalized hardware basis (circa 2026 Hopper pricing), the estimated cost of a comparable new training run is ~US\$11M.

**BENCHMARKS (VS LEADING MODEL IN RED)**

|          |   |      |
|----------|---|------|
| MMLU     | Not reported.   |      |
| MMLU-Pro |  | 85.0 |
| GPQA     |  | 84.2 |
| HLE      | Not reported.   |      |

**ASSETS**

|               |   |
|---------------|---|
| Documentation | <a href="https://microsoft.ai/wp-content/uploads/2026/06/main_20260602_2.pdf">https://microsoft.ai/wp-content/uploads/2026/06/main_20260602_2.pdf</a> |
| Playground    | <a href="https://microsoft.ai/news/introducing-mai-thinking-1/">https://microsoft.ai/news/introducing-mai-thinking-1/</a>                             |
| More          | <a href="https://lifearchitect.ai/models-table-pro/">https://lifearchitect.ai/models-table-pro/</a>   |

**ADDITIONAL REMARKS**

Microsoft AI's reasoning model. 35B-active, ~1T-total parameters sparse MoE. Trained from the ground up without distillation from third-party models, on clean and commercially licensed data. Matches Claude Opus 4.6 on SWE-Bench Pro and preferred over Claude Sonnet 4.6 in blind human side-by-side evaluations. AIME 2025=97.0, AIME 2026=94.5.

**RATINGS (2026Q2)****88/100****COMPUTE PERCENTILE**

Percentile rank of estimated compute versus all models in Models Table Pro.

**A****DISCLOSURE**

Parameters disclosed · Tokens disclosed · Doc available · Playground available.

-

**ESTIMATE CONFIDENCE (\*)**

Model data disclosed by lab, estimates not required.

**C****OPTIMIZATION**

Tokens-to-total-params ratio: 34:1 · Near-Chinchilla.