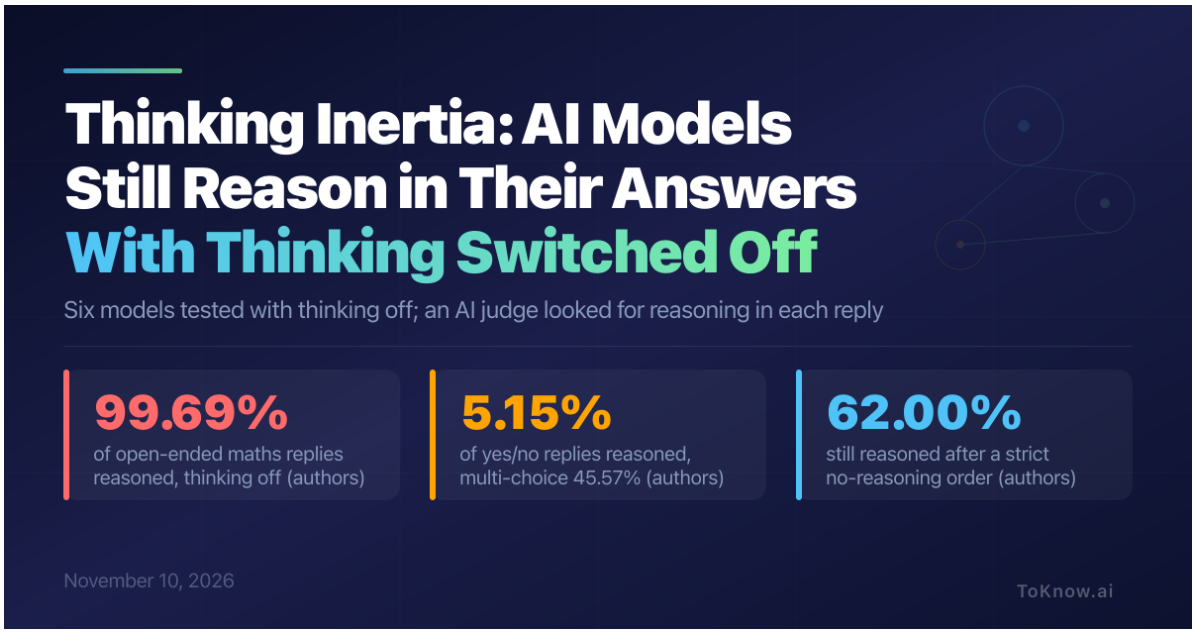
Thinking Inertia: AI Models Still Reason in Their Answers With Thinking Switched Off

Kabui, Charles

2026-11-13

Read at ToKnow.ai



A study of the "thinking off" switch on AI models found that the models still reason inside their answers. It matters to anyone who pays for AI answers by length or builds on that switch. Reasoning models write out step-by-step working before answering, which costs time and money, so many offer a switch to make them reply at once. In a paper accepted at the NeurIPS 2026 AI conference, researchers at Tsinghua, Oxford and Stanford tested six

models: three sizes of Alibaba's Qwen3, DeepSeek-V4-Flash, OpenAI's GPT-5.4 and Google's Gemini-3-Flash. An AI judge, GPT-5.5, checked the visible text before each final answer for a reasoning step, such as a calculation. With thinking off or at its lowest setting, it found reasoning in 5.15% of replies to yes/no questions, 45.57% to multiple-choice questions and 99.69% to open-ended maths problems.

Alibaba's Qwen3 model card says that with thinking off the model "will not generate any think content". In the tests, the working appeared in the answer text instead. A strict instruction, "give the final answer directly without any reasoning", still left reasoning in 62.00% of open-ended replies. Qwen3-32B mostly obeyed it and solved 44.4% of grade-school maths problems, against 93.4% with thinking off.

The authors suggest letting a model answer at once only when the question shows candidate answers, and keeping the working when dropping it costs accuracy. Rewritten as yes/no checks of a proposed answer, the same numeric questions drew reasoning in 36.90% of Qwen3-32B replies, against 90.14% with no candidate shown.

Read More: Antidoom, Liquid AI's training fix for small reasoning models that repeat the same text until they run out of room

Sources:

- Thinking Inertia: LLMs Keep Thinking When Told Not To (arXiv abstract, accepted by NeurIPS 2026)
- Thinking Inertia full paper, version 2 (arXiv)
- Thinking Inertia paper summary with author affiliations (HyperAI)
- Qwen3-32B model card, thinking switch (Hugging Face)
- Thinking Inertia: LLMs keep reasoning under no-think controls (AI Weekly)

Disclaimer: *For information only. Accuracy or completeness not guaranteed. Illegal use prohibited. Not professional advice or solicitation.* ***Read more:*** */terms-of-service*