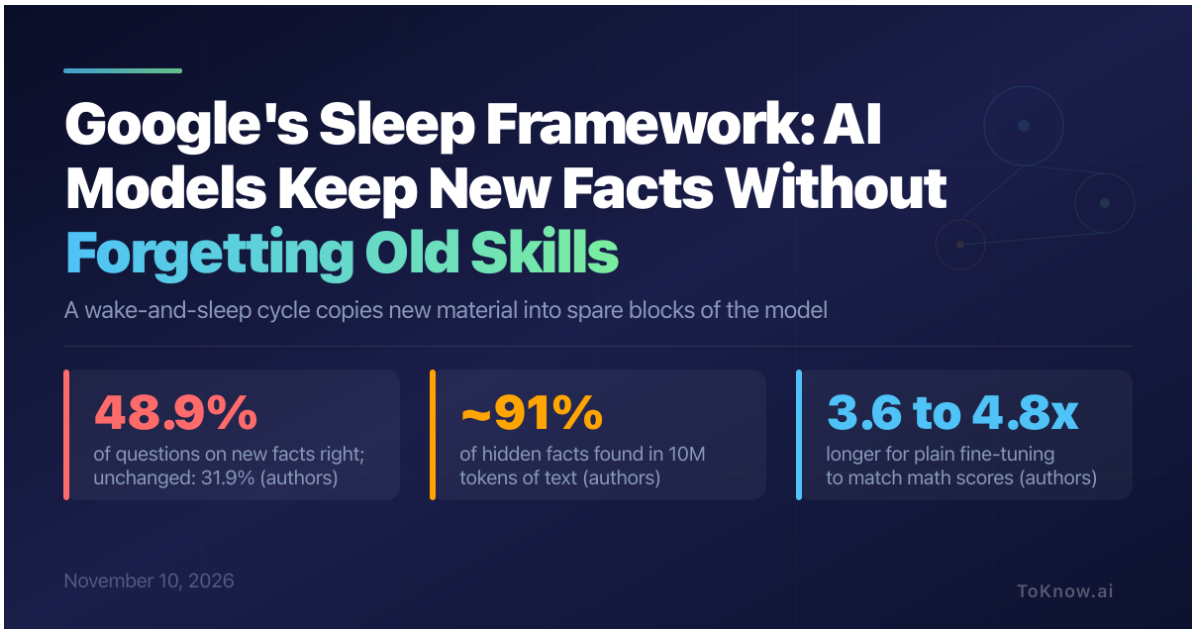
Google's Sleep Framework: AI Models Keep New Facts Without Forgetting Old Skills

Kabui, Charles

2026-11-12

Read at ToKnow.ai



Google researchers describe a training cycle, the sleep framework, that lets a language model learn new facts after release without losing old skills. A released model is frozen: it forgets what a prompt told it when the chat ends, and training it further on new data can erase old skills (catastrophic forgetting). Awake, the model keeps new material in short-term memory. In a sleep phase it stops taking input and trains spare expert blocks, specialist parts of its

network (a mixture of experts), to give the same answers without that material in view, then clears its short-term memory. The old parts stay frozen, like copying notes onto spare pages. In a dreaming step, it writes practice examples and trains on the best.

The aim is a model that keeps learning after release. The paper, from June and described in a Google blog on October 27, tested models of up to 8 billion parameters. After learning one passage of new facts, the model answered 48.9% of questions about it without seeing it again, against 31.9% for the unchanged model. On BABILong, which hides facts such as "Mary traveled to the office" in book-length text and asks "Where is Mary?", the paper's chart shows a Llama-based model finding about 91% of answers in 10 million tokens of text (about 7.5 million words). Matching the sleep method's maths scores took ordinary fine-tuning 3.6 to 4.8 times as long. The blog notes that "sleep" is borrowed language; people write the training rules.

Read More: Do AI Models Need Sleep? An Offline 'Rest' Pass That Boosts Reasoning, a different team's sleep-style pause that compresses recent context into the model before clearing its memory

Sources:

- The sleep framework: A new way for LLMs to retain knowledge in continual learning (Google Research blog)
- Language Models Need Sleep: Learning to Self-Modify and Consolidate Memories (arXiv paper, full text)
- BABILong: Testing the Limits of LLMs with Long Context Reasoning-in-a-Haystack (arXiv)
- Mixture of Experts Explained (Hugging Face)
- Understanding and counting tokens (OpenAI Help Center)

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