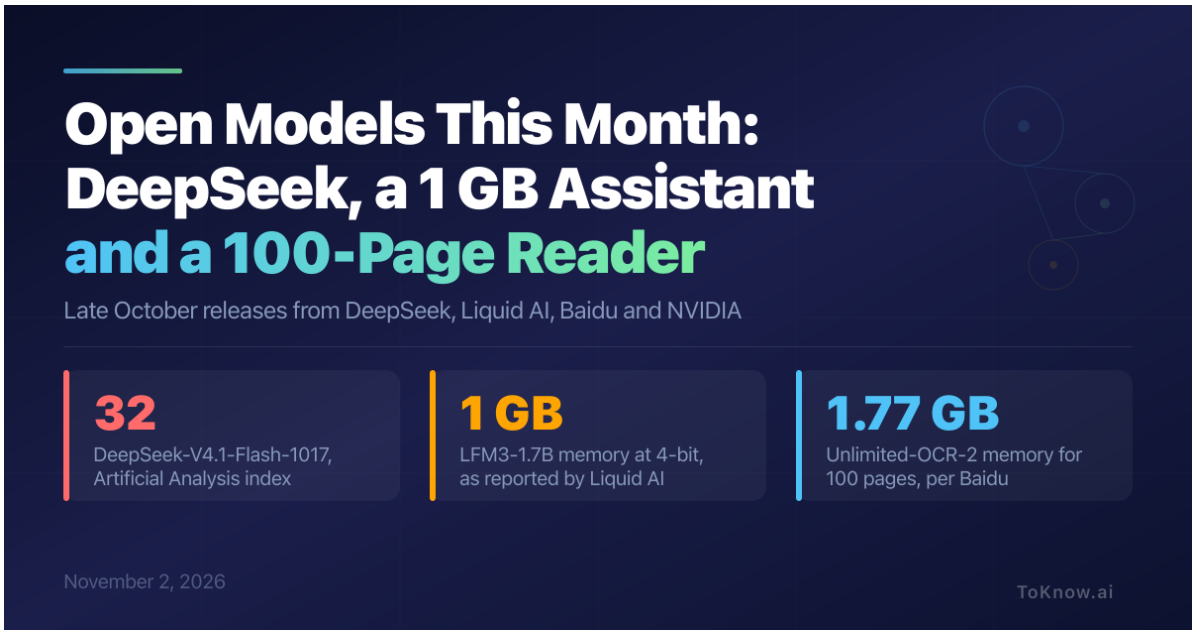
Open Models This Month: DeepSeek's Update, a 1 GB Assistant Model, and a Reader for 100-Page Documents

Kabui, Charles

2026-11-02

Read at ToKnow.ai



DeepSeek-V4.1-Flash-1017, released October 17, is a downloadable open-weight model for writing software and handling multi-step tasks, with 552 billion parameters (the numbers a model learns, roughly its size). DeepSeek reports 42.3 on Terminal-Bench 4.0 (command-line

tasks), up from 31.2 before the update. The independent tracker Artificial Analysis combines nine tests into one index and ranks it first among 23 similar-size open-weight models, at 32 against 59 for Gemini 4 Carbon. GLM-5.3 ties it at 32 but costs $5.01 per test task, against $0.60 for DeepSeek.

Smaller open-weight releases fit one device or job. Liquid AI's LFM3-1.7B is a text model for assistants that run on your own device. It needs about 1 GB of memory when compressed to 4-bit numbers, and Liquid says it beats its 2.6-billion-parameter predecessor. Baidu's Unlimited-OCR-2 turns scanned pages into text, reading a 100-page document in one pass. Baidu reports it needs 1.77 GB of short-term memory, against 47 GB with the usual method. NVIDIA's Nemotron 3.5 Transcribe turns hour-long recordings into text in up to 40 languages, but not live speech.

Baidu's test for long-document readers, MarathonDocBench, holds 100 documents of 100 pages or more. Unlimited-OCR-2 averages 77.06 out of 100 across eight accuracy measures and the next model listed scores 33.7, but Baidu wrote the test and lists no independent results. BizAgentBench gives AI agents 50 payment-company tasks and 601 business tools. The September DeepSeek-V4.1-Flash completes 43.3% of tasks without breaking a rule, against 64.0% for the top model tested, Claude Opus 5.

Read More: the previous open-source radar, on MiMo-V2.6, Kandinsky 6.0 Video and EmbeddingGemma 2 and the September radar, on the earlier DeepSeek-V4.1-Flash release

Sources:

- DeepSeek-V4.1-Flash-1017 release note (DeepSeek)
- DeepSeek-V4.1-Flash-1017 model card (Hugging Face)
- DeepSeek V4.1 Flash 1017: intelligence, performance and price analysis (Artificial Analysis)
- Introducing LFM3: Every Modality for Specialized Intelligence (Liquid AI)
- Unlimited OCR Works 2 (arXiv)
- MarathonDocBench dataset card (Hugging Face)
- NVIDIA-Nemotron-3.5-Transcribe-30B-A4B model card (Hugging Face)
- BizAgentBench: Unveiling Hidden Challenges in Deploying Long-Horizon Agents for Real-World Business (arXiv)

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