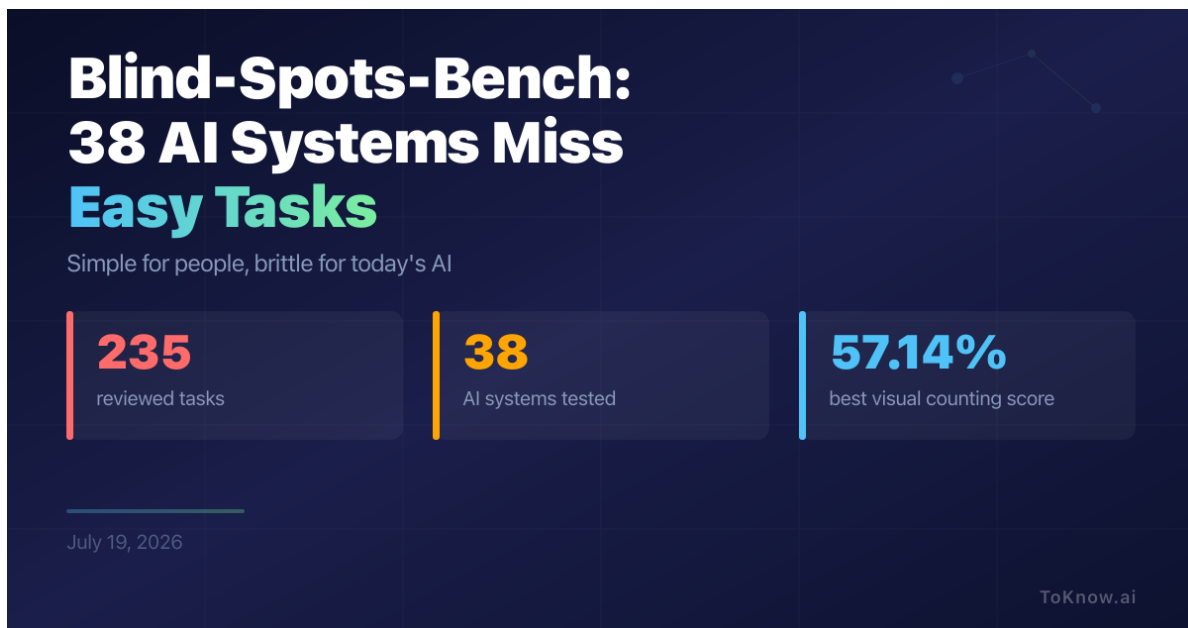


Blind-Spots-Bench: 38 AI Systems Miss Easy Tasks

Kabui, Charles

2026-07-19

[Read at ToKnow.ai](#)



EPFL researchers built [Blind-Spots-Bench](#) from problems that graduate AI students found easy but frontier models failed. Students submitted about 287 examples in late 2025; review and cleanup left 235 text, image, and mixed-media tasks with examples involving character-level text manipulation, spatial reasoning, clock reading, and object counting. The team tested 32 language or vision-language models plus six image generators. Averaged across four runs,

GPT-5.5 scored 84% on text tasks, while Gemini 3.1 Pro led inputs that combined text and images at 66.9%. Gemini 3 Pro Image topped the single-run image-generation evaluation at 54.8%. The weak spots were more revealing than the leaders: the best four-run average was 41.67% on visual attribute and pattern recognition, and 57.14% on visual counting.

These are the annoying failures that broad benchmark averages hide. A model can ace familiar math and coding tests, then miscount objects or break while copying a string into Python. Tool access did not reliably fix that: it raised Gemini 3.1 Flash Lite by 9.03%, but lowered GPT-5.4 by 5.32%. Those tool results came from one run and were compared with four-run base averages, so the deltas are only suggestive. The authors checked their automated grader against people on 154 outputs, finding 96.6% agreement for text and 90.9% for images. That audit is reassuring but small. The benchmark also comes from one course, and its public release makes future training contamination likely. It works best as a test suite for concrete failures, not another single-number intelligence ranking.

Read More: [How eight agent benchmarks reached 100% without solving their tasks.](#)

Sources:

- [Blind-Spots-Bench paper](#)
- [Blind-Spots-Bench dataset](#)
- [Evaluation code and per-question results](#)

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