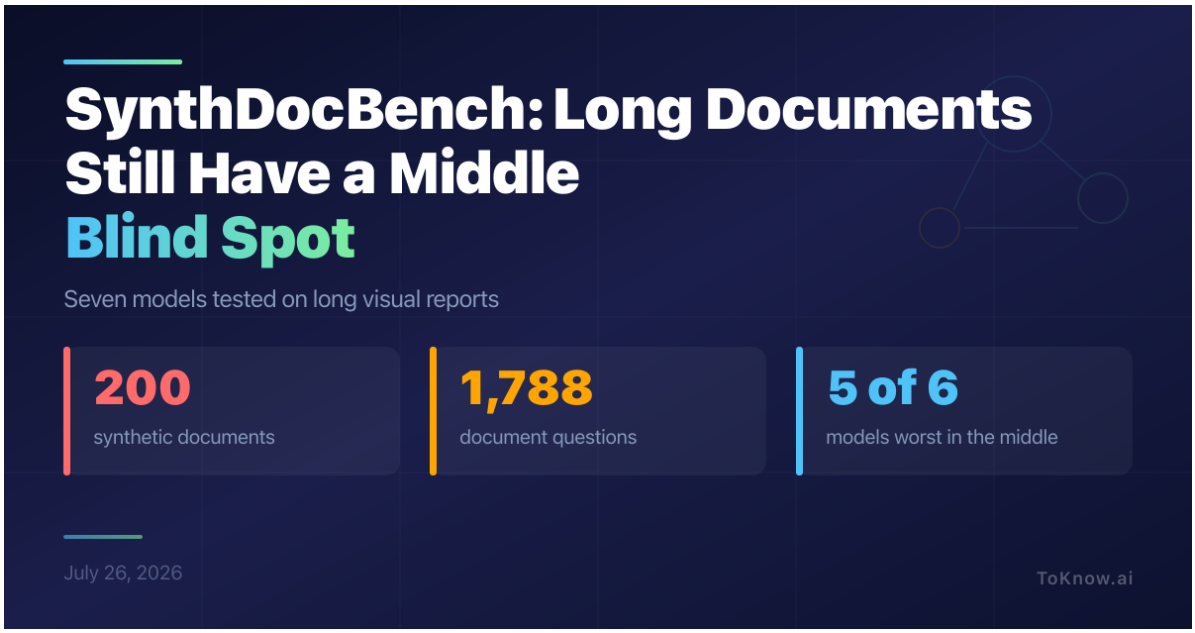
SynthDocBench: The Middle of a Long Document Is Still a Blind Spot

Kabui, Charles

2026-07-26

Read at ToKnow.ai



ServiceNow researchers built SynthDocBench to test whether vision-language models can reason across long, mixed-format documents rather than isolated pages. The benchmark contains 1,788 questions over 200 synthetic documents with 3,340 charts. Each document averages 51.1 pages and 20,568 words. Its generator varies length, layout, chart content, and question difficulty independently across six layouts, with a 40% random override designed to stop

models from learning easy layout shortcuts. Across seven frontier models, accuracy fell as documents grew. In the controlled position test, five of six analysed models scored lowest when the evidence sat in the middle third. Five also declined from early to late evidence, with the largest drop reaching 8.3%. Chart reading weakened further when charts appeared inside long reports.

That matters for anyone asking AI to review contracts, annual reports, research papers, or case files. A large context window can accept every page while still missing the evidence that answers the question. SynthDocBench makes that failure easier to isolate because its charts have hidden structured data and deterministic answers. A 100-question manual audit accepted more than 96% of generated items. Still, synthetic reports are cleaner than real scans, and no independent replication has appeared yet. The benchmark diagnoses a risk; it does not measure the error rate on your own documents.

Read More: How Prompt Repetition Addresses the Text-Only "Lost in the Middle" Problem

Sources:

- SynthDocBench Paper
- SynthDocBench Code and Evaluation Harness
- SynthDocBench Dataset
- Lost in the Middle Paper
