

When do LLMs use evidence correctly?

Khabibillo Gulomov, Kyungpook National University

Daegu, South Korea khabibillo@knu.ac.kr khabibillo.com github.com/khabibro linkedin.com/in/gulomovkhabibillo

I study how reliably large language models use the evidence they are given: in multi-document QA, in retrieval- and search-augmented reasoning, and in tool-using agents, with evaluation designed to hold up statistically.

Featured paper

Evidence Position Sensitivity Across Document-Count Conditions in Multi-Document QA

Khabibillo Gulomov (sole author). Korea AI 2026, 7th Korea Artificial Intelligence Conference. Accepted for publication.

6 LLMs, 3B to 72B	24 conditions	2,417 examples per condition	72 contrasts, significance-tested
Question and design Does the position of supporting evidence among retrieved documents change answer quality, and does that effect depend on how many documents there are? Evidence was placed at the beginning, middle or end of 8, 10, 15 or 20 documents with total input length held fixed, and contrasts were tested with paired bootstrap and sign-flip tests.		Findings • The best evidence position depends on the model; none is best for all six. • A model's preferred position can change with document count. • Higher-hop questions can show different positional behavior. For RAG systems, passage ordering should be validated for the model and context size in use.	

Other research

Search-R1 reproduction, Data and Language Intelligence Lab, KNU 2026
Reproduced and evaluated search-augmented reasoning on NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue and Bamboogle.

Retrieval- and tool-augmented LLMs, Data and Language Intelligence Lab, KNU 2026
Evaluated SeaKR and ToolBench / ToolLLM: tool selection, retrieval, and the reliability of tool-using agents.

Methods: Python, PyTorch, Hugging Face, vLLM, LLM evaluation, paired bootstrap and sign-flip testing