
Sufficient Context Is Not a Complete Answer: 70% of Required Elements Met, 40% of Arabic Answers Complete

Dmitrii Lozhkin
AVELIN AI · October 26, 2026

Abstract

Retrieval-augmented systems are usually judged by whether the right passages come back, yet a context holding everything an answer needs does not guarantee an answer that uses it. We measured the gap on 31 Arabic questions over Arabic and English scientific reports, with required answer elements written in advance and reference passages confirmed sufficient by an answer-blind check. Given these passages and asked to cover every part in at most 70 words, six open models met 70% of required elements but completed 40% of answers; by type, 29 of 36 direct-fact answers were complete against 24 of 96 causal or synthesis answers, a 56-point gap (document-group bootstrap 95% interval 31 to 77). Without the length clause the gap was 47 points (29 to 66); synthesis completeness doubled to 43%, while numerical answers did not improve. Dense BGE-M3 retrieval found sufficient evidence for 23 of 32 questions against 16 for BM25, but Gemma 2 9B completed only 44% of sufficient contexts. A blinded AI review gave no credit for padding.

1 Introduction

For retrieval, M3-Embedding reports that Arabic queries over English Wikipedia reach Recall@100 of 18.9 with BM25 and 71.1 with its dense retriever (MKQA, Table 2 of [1]), and multilingual RAG pipelines built on BGE-M3, Arabic included, have been evaluated end to end [2]. For generation, Joren et al. separate whether a context is sufficient from whether the answer is correct: on their curated human-labelled set, Gemma 2 27B answered 71.4% of sufficient-context questions correctly and hallucinated on 25.4% [3]. CRAG shows that question type matters, with straightforward RAG at about 44% accuracy and lower accuracy on questions of higher complexity [4].

These studies score an answer as right or wrong, mostly in English. We add an element-level measurement in Arabic, where a reviewer marks each fact, value and link a complete answer must contain, and report it by question type for six open models. We then test two explanations of the gap: the length clause, by regenerating every answer without it, and a reviewer preference for length, by padding answers.

2 Setup

Corpus and questions. The corpus holds 1,778 passages from 24 public scientific reports, 1,000 in Arabic and 778 in English, in 12 document groups. A *question* is an Arabic prompt with *required elements*, both written in advance with AI assistance: the facts, values, calculations or causal links a complete answer must state. The 32 questions here were written to have a *sufficient context*, reference passages that support every required element; half draw on Arabic sources, half on English, and 6 ask for a direct fact, 10 for a number or comparison, 6 for a causal explanation and 10 for a multi-passages synthesis. An answer-blind check (Review) confirmed 31; one numerical question was only partly supported, so results with supplied passages use 31 questions with 119 required elements, 2 to 5 each. A *complete answer* meets every element, computes any requested value and unit correctly, and has no critical error, materially unsupported claim or false refusal.

Contexts and models. With supplied passages, each question received its one or two reference passages, and Gemma 2 9B IT, SILMA 9B, ALLaM 7B, Falcon-H1 7B and Qwen3.5 4B and 9B (reasoning disabled) answered once; we compare question types, not models. With retrieval, Gemma answered from the top four passages of BM25 ($k_1 = 1.2$, $b = 0.75$) or of dense-only BGE-M3, with the original Arabic queries and no translation, hybrid scoring or reranking. All runs used 8-bit weights, a 4,096-token context, temperature 0 and seed 17.

Instruction. One Arabic instruction opened every prompt: answer only from the passages, in one to four short Modern Standard Arabic sentences of at most 70 words; answer every part of the question; give a calculation's result with its unit; mark the source after each factual sentence; say plainly when information is missing. No original answer reached its 256-token limit. After reproducing all 352 original answers byte for byte, we regenerated every prompt with only the length clause changed: to at most 150 words in up to eight sentences, or removed.

Review. One fixed procedure labelled all 768 answers and, in separate calls that never saw an answer, the evidence in all 96 contexts: a general-purpose AI model run statelessly without tools, one anonymous item per call, with model, instruction and retriever hidden. Two independent passes (element $\kappa = 0.95$) and a third on any disagreement or low confidence gave labels frozen before unblinding; on the original answers they agreed with an earlier four-round review at $\kappa = 0.82$ for elements and 0.83 for completeness. Paired contrasts use exact McNemar tests; intervals re-sample the 12 document groups.

3 Results

Table 1: Required elements met and complete answers by question type (number of questions). Six models with supplied passages; retrieved: Gemma’s BM25 and dense contexts judged sufficient, from all 32 questions, 70-word clause.

Question type	Elements met	Complete, 70 words	Complete, no clause	Complete, retrieved
Direct fact (6)	100/108 (93%)	29/36	32/36	7/8
Number or comparison (9)	148/216 (69%)	21/54	16/54	4/13
Causal explanation (6)	94/138 (68%)	11/36	14/36	2/8
Synthesis (10)	156/252 (62%)	13/60	26/60	4/10
All (31)	498/714 (70%)	74/186	88/186	17/39

Completeness differed more between question types than between models. The six models completed 9 to 16 of 31 answers each (29% to 52%); the question types ranged from 22% to 81% (Table 1). A direct-fact question was completed by 4.8 of six models on average and a causal or synthesis question by 1.5 (exact permutation $p = 0.0012$). Direct-fact questions carry fewer elements, 3.0 against 3.8 to 4.2, which eases completion, but element credit itself differs, 93% against 62% to 69%, so the count alone does not explain the gap. Type is confounded with content (Section 4), so an independent type effect is not established.

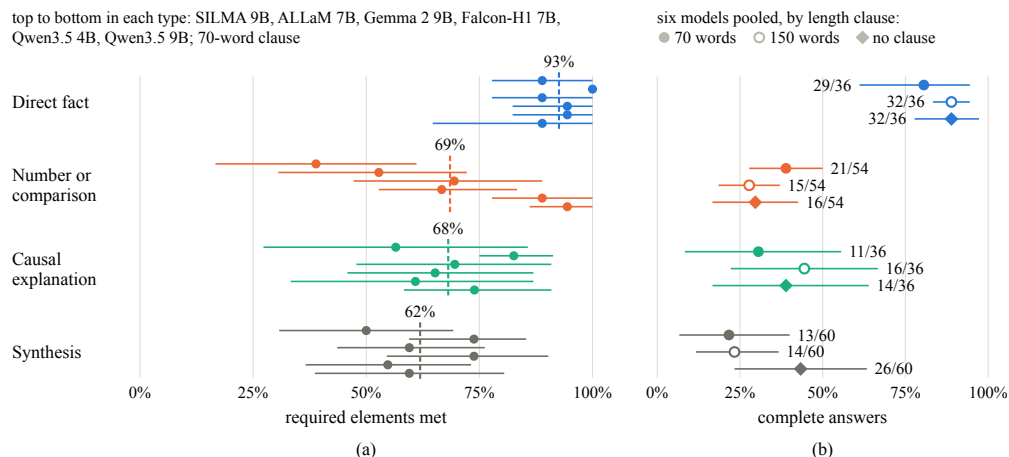


Figure 1: (a) Required elements met under the 70-word clause by question type, for each of six models on the 31 questions; dashed lines: six-model share. (b) Complete answers by question type and length clause, six models pooled. Whiskers: group bootstrap 95% intervals.

The gap persists without the 70-word clause. Removing the clause made causal and synthesis answers 1.33 times longer (median) and raised their completeness by 16.7 points on the same model–question pairs (10.8 to 22.7; 19 gained, 3 lost, McNemar $p = 0.0009$). Direct facts rose too, 29 to 32 of 36, so the gap narrowed from 56 to 47 points, a shrinkage of 8.3 (−17.1 to 26.5) whose upper bound is just below half the original gap (27.8), as pre-specified for a gap that persists. The same test on element credit, the pre-specified secondary, was inconclusive (6.6, −3.8 to 15.0). A 150-word clause gave longer answers (median 48.5 words; 43 without a clause, 38 at 70 words) but a smaller causal and synthesis gain (+6.2 points, −1.1 to 12.2). It held in all three ($p = 0.0012, 0.0002, 0.0048$).

Synthesis lost content to compression; numbers did not. Under the clause synthesis answers ran longest, a median of 55 words with 15 of 60 over the limit, and 28 of their 47 incomplete answers touched every element but met some only partially. Without it they grew to 80.5 words, 37 elements moved from partial to met, and complete synthesis answers doubled from 13 to 26 of 60 (+21.7 points, 15.0 to 28.3; none lost). Numerical answers stayed near 30 words and did not improve, 21 then 16 of 54 complete, while answers failing the calculation criterion, with a requested value wrong or missing, rose from 14 to 19 (one a false refusal): length was not what they lacked.

Element credit followed content, not length. AI judges prefer longer answers in pairwise comparisons [5, 6]; this review checks fixed elements instead. Appending two to four fluent sentences without facts to 48 answers (1.6 times the words) changed element credit by -1.8 points (-4.8 to 0.0), deleting the sentences without an element from 32 answers changed it by -3.3 (-7.3 to 0.0), and deleting the sentence that carried one element removed its credit in 81 of 82 answers. The gains without the clause are not a length bonus.

Dense retrieval reached English evidence that BM25 could not. BM25 put English passages in 2 of its 128 top-four slots, dense retrieval in 38. Judged without the answer, dense contexts were sufficient on 23 of the 32 questions against 16 (9 to 2 discordant, $p = 0.065$), mostly English-source ones: 8 of 16 against 2 (Arabic: 15 against 14).

Fewer than half of sufficient contexts became complete answers. Gemma completed 6 of 16 sufficient BM25 contexts and 11 of 23 dense ones, 17 of 39 or 44% (bootstrap 23% to 61%), meeting 110 of their 146 required elements (75%). Without the length clause they gave 9 and 14 complete answers, 59% (41% to 76%).

4 Limitations

The questions come from 12 document groups and were chosen purposively; permutation tests treat them, and McNemar tests the model-question pairs, as independent. Only 5 groups contain both a direct-fact and a causal or synthesis question, so permuting types within groups gives $p = 0.063$, while the group bootstrap interval for the gap stays above zero. Type is confounded with question content. The length test changed only the 70-word clause, not other brevity cues, and two of six models lengthened causal and synthesis answers less than 1.2-fold. With the partly supported 32nd question, six models met 68% of elements and completed 39% of answers; the type comparisons contain no numerical question and do not change. All labels come from one AI review procedure, whose agreement with the earlier review was lowest on causal elements ($\kappa = 0.70$). Each question was generated once per condition; translated queries, hybrid retrieval and other generators were not run on retrieved contexts. Per-question records and source titles are not released.

5 Conclusions

Under a 70-word instruction a sufficient context gave a complete Arabic answer less than half the time, supplied or retrieved, so counting sufficient contexts alone overstated complete answers more than twofold. Completeness varied more with question type than with the model, and removing the length clause left 47 of the 56-point gap. The clause compressed synthesis answers; numerical answers failed on calculation either way.

6 Recommendations

1. Score answers per required element and report the all-elements rate beside it: 70% against 40% here.
2. Report completeness by question type before averaging: 81% of direct-fact and 25% of causal or synthesis answers were complete, and the 40% average describes neither.
3. Test a length limit on synthesis questions first: dropping a 70-word clause took them from 22% to 43% complete.
4. Recompute each requested calculation from its operands: 14 of 54 numerical answers given the right passages failed the calculation criterion, and 19 without the length clause.

Disclosure. Experiments and analysis were produced with AI coding agents under the author’s direction. All numbers were re-computed from the frozen per-answer labels.

References

- [1] J. Chen, S. Xiao, P. Zhang, et al. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In Findings of ACL, 2024. arXiv:2402.03216.
- [2] N. Chirkova, D. Rau, H. Déjean, et al. Retrieval-augmented generation in multilingual settings. arXiv:2407.01463, 2024.
- [3] H. Joren, J. Zhang, C.-S. Ferng, et al. Sufficient context: A new lens on retrieval augmented generation systems. In ICLR, 2025. arXiv:2411.06037.
- [4] X. Yang, K. Sun, H. Xin, et al. CRAG – Comprehensive RAG benchmark. In NeurIPS Datasets and Benchmarks, 2024. arXiv:2406.04744.
- [5] K. Saito, A. Wachi, K. Wataoka, Y. Akimoto. Verbosity bias in preference labeling by large language models. arXiv:2310.10076, 2023.
- [6] Y. Dubois, B. Galambosi, P. Liang, T. B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In COLM, 2024. arXiv:2404.04475.