

When Relevance Is Not Evidence: A Controlled Diagnostic of Claim-Conditioned Support Ranking

Ke Lu

Independent Researcher

look970726@gmail.com

Abstract

Retrieval-augmented generation ranks passages by topical relevance or semantic similarity, but generation needs passages that actually *support* a target claim. We study when these targets come apart, framing the task as **claim-conditioned support ranking over a constructed candidate shortlist**: given a query, an explicit target claim, and a small fixed set of candidates, rank them by evidential support. We build a controlled, typed development set of **175 controlled contrastive examples** spanning seven spurious-evidence cues, each pairing one supporting passage with one targeted non-support passage expressing the designated cue, plus neutral distractors. We evaluate with a **corrected, tie-aware** protocol that uses role-free evaluator-facing opaque identifiers and neutral tie handling. Tested query-only relevance rankers (lexical, dense, and a cross-encoder) yield tie-aware focal-pair error rates of roughly 18–34%; in the BM25 cue decomposition, several strata have high error, including negation and relation directionality; a gold-claim NLI reference—which receives information the rankers do not—reduces but does not eliminate the error. We argue that evidence-ranking evaluation should explicitly *condition on the claim* whose support is assessed, and sketch a typed, claim-conditioned evaluation direction that separates relevance, support, and source attributes. This is a controlled diagnostic on development data: it does not estimate real-world prevalence or provide a deployable method.

1 Introduction

Retrieval-augmented generation (RAG) grounds language models in retrieved passages (Lewis et al., 2020), but retrieval is optimized for topical *relevance* or semantic *similarity*, whereas faithful generation requires passages that provide evidential *support* for the specific claim being made. These targets are related but not identical: a passage can

be highly similar to a query—sharing its entities, numbers, or phrasing—yet fail to support the intended answer, or even contradict it, because of negation, a reversed relation, a different entity, or a stale value. For instance, a passage asserting a cipher *is* permitted shares nearly all terms with the claim that it is *not*, yet must rank below genuine support (§2).

We isolate this distinction with a controlled diagnostic. The relevance/support distinction and entailment-aware retrieval are not new (§4); our aim is to isolate *when* relevance-oriented rankers prefer high-overlap non-evidence to genuine support. We frame **claim-conditioned support ranking over a constructed candidate shortlist**: given a query, an explicit target claim, and a small fixed candidate set, a system ranks candidates by evidential support for the claim. This is deliberately *not* corpus-scale first-stage retrieval; it is a controlled reranking probe that lets us build controlled contrastive pairs and read off pairwise errors directly. We use this diagnostic to ground a broader, forward-looking argument: evidence-ranking evaluation should condition on the claim whose support is assessed and should separate relevance from support and source attributes (§5).

Because different systems may see different information, we label four **information-access conditions** and keep them separate throughout: (A) query-only relevance ranking; (B) generated-claim-conditioned support ranking (a deployable setting we do *not* implement here); (C) gold-claim-conditioned diagnostic ranking; and (D) a gold-label oracle. Comparing Condition A with Condition C is an *unequal-information* comparison: it diagnoses the value of explicit claim access plus a support signal, not model architecture alone.

Our contributions are:

1. a precise formulation of claim-conditioned support ranking over a constructed shortlist,

with an explicit Condition A/C/D information-access distinction (§2);

2. a **corrected, tie-aware evaluation** for this setting—role-free evaluator-facing opaque identifiers, neutral tie handling, tie-aware pairwise error, and expected rank metrics—that removes a role-correlated tie-break present in the prior implementation (§2);
3. a **cue-decomposed diagnosis** on 175 controlled development examples: all tested query-only relevance rankers exhibit nonzero focal-pair error, while several BM25 strata show high error (negation, directionality, and redundancy/majority), with a gold-claim NLI model included as a diagnostic reference (§3).

Scope. All results are on *constructed development data* whose labels were produced and audited by the same project (§2). We make no claim about real-world prevalence, all-retriever generality, or a deployable mitigation, and we present no downstream generation results here (§6).

2 Controlled Diagnostic Design

Shortlist structure. Each example is a query with an explicit target claim (the query’s intended answer expressed as a proposition) and a fixed shortlist of five candidate passages: one *supporting* passage, one *targeted non-support* passage designed to preserve cue-relevant query/entity/attribute signals without supporting the claim, and three neutral distractors. The supporting and non-support passages form a **controlled contrastive pair**. Negation and directionality often instantiate near-minimal surface edits; other cues preserve targeted query/entity/attribute overlap while varying referent, time, source style, or redundancy. Formally, given a query q , target claim c , and candidates $P = \{p_1, \dots, p_5\}$ with a supporting p^+ and targeted non-support p^- , a query-only ranker scores $f(q, p)$ and a claim-conditioned ranker $g(q, c, p)$; the diagnostic asks whether p^- outranks p^+ .

Cue types. The 175 examples are balanced across seven cue types (25 each) describing *why* the non-support passage is superficially attractive: entity overlap, number overlap, temporal mismatch, negation/polarity, relation directionality, authority style, and redundancy/majority. The examples are *self-contained*—each shortlist supplies all passages needed to judge support—and use either public facts or explicitly fictional mini-worlds.

Worked minimal pair (negation, fictional). To illustrate the self-contained design without external knowledge, we use an invented mini-world. *Claim*: “Under the Meridian Data Standard, the CX-2 cipher is not permitted for storing customer records.” *Supporting*: “The Meridian Data Standard states that CX-2 must not be used to store customer records; systems are required to use CX-4 or higher.” *Non-support (negation)*: “Under the Meridian Data Standard, CX-2 is a permitted cipher for storing customer records.” The non-support passage shares nearly all terms with the claim but asserts the opposite polarity, so lexical and similarity signals rank it alongside—or above—the supporting passage.

Label provenance (important caveat). The examples were constructed with *intended* support and non-support roles and subsequently subjected to a project-internal audit under a two-axis relation/attribute protocol. This is a **development quality-control process, not independent human reliability validation**: because the same project created and audited the examples, construction bias remains possible. The audit found no obvious malformed records but flagged temporal-scope ambiguity and construction confounds (e.g., the authority-style non-support passages are also contradictions) that would require independent annotation to resolve. We therefore treat labels as development annotations and report descriptive findings, not out-of-sample generalization.

Corrected, tie-aware metrics. Our unit of analysis is the support/non-support pair. **Spurious Win Rate (SWR)** is the mean over examples of 1 if the non-support passage outscores the supporting one, 0.5 on an exact score tie, and 0 otherwise; it is score-based and therefore invariant to how ties are ordered. **Expected Support@1** is the probability that a supporting passage is ranked first under uniform random tie-breaking within equal-score groups (computed exactly by enumeration for these five-candidate shortlists). We also report the support/non-support **tie rate**.

These corrections matter: the prior implementation broke score ties using role-bearing candidate identifiers, which alphabetically favored the non-support passage and inflated the headline error. Using role-free opaque identifiers and tie-aware scoring, BM25’s pairwise error falls from a strict role-resolved $57/175=0.326$ to a tie-aware 0.280 (16 exact ties), and the same artifact depressed Sup-

System	Info	SWR↓	E[S@1]↑
<i>Query-only relevance (Condition A)</i>			
Lexical overlap	query	0.246	0.744
BM25	query	0.280	0.697
Dense E5-small	query	0.274	0.726
Dense MiniLM	query	0.337	0.663
CE relevance	query	0.177	0.823
<i>Gold-claim diagnostic reference (Condition C)</i>			
BM25 + NLI	query+claim	0.086	0.909
<i>Gold-label upper bound (Condition D)</i>			
Oracle	gold labels	0.000	1.000

Table 1: Corrected tie-aware Spurious Win Rate (SWR; lower is better) and expected Support@1 (E[S@1]; higher is better) on the 175-example controlled development set. “CE relevance” is a relevance cross-encoder; “query+claim” denotes the *gold* target claim. **Conditions C and D receive information unavailable to Condition-A systems and are diagnostic references, not input-equivalent baselines.** Only the two lexical systems have exact support–non-support focal-pair score ties (rates 0.091 BM25, 0.297 lexical overlap), so their SWR requires the tie-aware treatment. Because each cue contributes 25 examples, aggregate metrics are balanced diagnostic averages, not estimates under a natural cue distribution.

port@1. All SWR numbers below use the corrected metric; the continuous-score systems have no exact support–non-support focal-pair ties, so their SWR values are unchanged.

3 Results

We evaluate five query-only rankers—lexical overlap, BM25, two dense retrievers (all-MiniLM-L6-v2, e5-small-v2), and a relevance cross-encoder (ms-marco-MiniLM-L-6-v2)—and, as diagnostic references, a gold-claim NLI reranker (nli-deberta-v3-small; He et al., 2023) and a gold-label oracle.

Query-only relevance rankings exhibit systematic support-ordering errors. When evaluated under the support objective, every tested Condition-A system exhibits nonzero focal-pair error under the tie-aware metric (Table 1): SWR ranges from 0.177 (a supervised relevance cross-encoder) to 0.337 (dense MiniLM). The lexical and dense baseline point estimates are broadly similar on this development set, and we do not interpret their ordering. The strongest query-only system—the cross-encoder relevance reranker—still errs on roughly one pair in six. SWR isolates the focal support–non-

Cue type	BM25 (A)	BM25+NLI (C)
number overlap	0.00	0.00
entity overlap	0.08	0.00
temporal mismatch	0.08	0.04
authority style	0.24	0.00
redundancy/majority	0.48	0.16
relation directionality	0.50	0.28
negation/polarity	0.58	0.12

Table 2: Per-cue tie-aware SWR (↓), 25 examples per cue. For the query-only ranker (BM25) errors are highest for **negation**, **relation directionality**, and **redundancy/majority**. The gold-claim NLI reference (Condition C) reduces every cue with nonzero BM25 error, drives entity and authority to zero, and leaves number at zero, with the largest residual in **directionality**. Cue rows are descriptive construction strata, not causal-effect estimates.

support ordering, whereas expected Support@1 can additionally reflect how the supporting passage ranks against the three neutral candidates.

BM25 errors are highest for negation, directionality, and redundancy. Table 2 decomposes BM25 by cue. BM25 errors are highest for negation (0.58) and directionality (0.50), with the passage-level redundancy/majority stratum close behind (0.48), and near zero for number and entity overlap. Because that redundancy/majority stratum is not a controlled set-level majority manipulation, we do not interpret it as a majority-effect estimate (§6). Negation and directionality provide the clearest semantic contrast mechanisms, flipping the claim’s polarity or reversing its relation while sharing almost all of its terms.

A gold claim plus NLI helps, under unequal information. The Condition-C reference reranks the fixed five candidates by fusing the min–max-normalized BM25 relevance score rel with cross-encoder NLI probabilities (premise=passage; the hypothesis is a question–answer serialization of the gold claim, Question: {q}\nAnswer: {a}, where a is the gold answer associated with the target claim): $s = rel + 2p_{ent} - 2p_{con} - p_{neu}$. It lowers aggregate tie-aware SWR to 0.086, reduces all six cues with nonzero BM25 error, and leaves three cues at zero (Table 2). This reference receives information the Condition-A rankers do not, so the gap measures claim access plus a support signal, *not* a better retriever. Even so, a residual error remains: the largest residual under this NLI reference is relation directionality (0.28). The oracle (Condition D) reaches 0 by construction and only bounds

the setting.

4 Discussion

What this study does and does not show. On this controlled, cue-typed development set, tested query-only rankers exhibit support-ordering errors. In the BM25 cue decomposition, these errors are concentrated in negation and directionality, where the contrastive pairs are closely matched, and in the passage-level redundancy stratum (Table 2). These systems optimize query relevance, not claim support, so this is an *objective mismatch*, not a failure of relevance ranking on its own terms; a gold-claim NLI signal reduces but does not eliminate the observed errors.

Related work. The relevance/support distinction is long-standing, from the PASCAL RTE challenge (Dagan et al., 2006) to natural-logic accounts of surface versus semantic inference (MacCartney and Manning, 2007). It is also central to modern work: fact verification retrieves and labels evidence for a claim (Thorne et al., 2018), natural-language inference models entailment and contradiction (Bowman et al., 2015; Williams et al., 2018), attribution work asks whether generated statements are supported by cited sources (Rashkin et al., 2023; Gao et al., 2023), and knowledge-conflict studies examine models facing contradictory evidence (Longpre et al., 2021; Xie et al., 2024; Xu et al., 2024). Most closely, prior work uses claim-evidence entailment to improve evidence retrieval (Yang et al., 2021), learns support retrievers conditioned on question-answer pairs (Zhang et al., 2025), and systems such as SIDE rank sources by how well they support a claim (Petroni et al., 2023); Alt et al. (2026) rank evidence for sufficiency and non-redundancy, and utility-oriented retrieval such as SCARLet (Xu et al., 2025) optimizes downstream usefulness. Embedding retrieval also has *representational* limits—for a fixed dimension, some top- k sets are unreturnable by any query (Weller et al., 2026)—whereas our failures are task-level, arising even when support is present. Methodologically, our contrastive pairs follow controlled diagnostics that expose hidden failures—lexical-substitution NLI sets (Glockner et al., 2018) and overlap heuristics (McCoy et al., 2019)—and are closest to VitaminC (Schuster et al., 2021), whose contrastive pairs differ in support; unlike its verification *labeling*, we score claim-conditioned *ranking* with a tie-aware metric. Support-aware,

answer-conditioned, and entailment-based ranking are therefore *not* new, and we claim neither them nor the relevance/support distinction itself. Our contribution is a *controlled, cue-typed ranking diagnostic* that localizes support-ordering failures across controlled cue types, under a corrected tie-aware evaluation separating query-only relevance (Condition A) from gold-claim support (Condition C).

5 Toward Claim-Conditioned Evidence-Ranking Evaluation

The diagnostic grounds an evaluation direction for knowledge-grounded generation, bearing on evidence selection, provenance, and verification. On this controlled set, relevance-oriented rankers systematically exhibit cases in which high-overlap non-evidence outranks genuine support; adding gold-claim access plus an NLI signal reduces but does not eliminate these errors. **(1) Relevance is an inadequate sole objective for evidence selection:** our controlled contrastive pairs show that relevance-oriented ranking does not reliably separate support from high-overlap non-support. **(2) Evaluate against explicit claims:** evidence ranking should be scored relative to the specific claim, and each system’s information access (query-only vs. claim-conditioned) stated, since these settings are not comparable. **(3) Typed controlled diagnostics can complement end-to-end RAG evaluation:** cue-decomposed controlled contrastive pairs localize failure modes (here, in BM25’s cue decomposition; Table 2) that aggregate scores obscure. **(4) Separate the axes:** benchmarks and systems should distinguish query relevance, claim support, source attributes, and provenance rather than collapsing them into a single relevance score. Contradictory passages can be valuable counterevidence for verification; SWR measures whether a passage supports the designated target claim, not its overall usefulness for verifying or challenging that claim, so a claim-aware system should expose support and contradiction as distinct relations. We do *not* claim that structured knowledge bases or entailment models resolve this. A claim-aware, axis-separated view of evidence ranking provides a clearer basis for evaluating evidence selection in knowledge-grounded generation.

6 Limitations

This is a controlled diagnostic, and its scope is correspondingly narrow. (i) Systems rank *constructed five-candidate shortlists*, not corpus-scale retrieval; SWR is a focal-pair quantity, whereas Support@1 is a five-candidate shortlist quantity; neither measures corpus-scale retrieval recall. (ii) The 175 examples are *development/legacy data*: the same project developed the cue taxonomy and metrics on this family ($\text{seed} \subset 70 \subset 175$), so results are descriptive, not out-of-sample estimates. The benchmark is intentionally small and controlled rather than representative of open-domain retrieval traffic: its purpose is to test whether specific confounding cues can induce ranking failures under tightly matched conditions, not to estimate their prevalence in deployed RAG systems. (iii) Labels were constructed and audited within the same project, rather than independently human-validated; construction bias remains possible, and temporal-scope and authority/contradiction confounds are unresolved. (iv) The legacy construction places *exactly one* supporting and *one* targeted non-support passage per shortlist, which cannot represent multiple or partial support; relatedly, the redundancy/majority cue is passage-level and does not constitute a controlled set-level majority manipulation. (v) Cue type is partially confounded with relation label by construction—for example, the negation, directionality, authority, and legacy redundancy strata use contradictory targeted passages—so per-cue differences should not be interpreted as isolated cue effects. (vi) We test a *limited set of local models* and one generic NLI model. (vii) The NLI reference uses a *gold claim* (Condition C); a deployable generated-claim setting (Condition B)—draft an answer, extract its claims, retrieve and rerank support and counterevidence, then revise or abstain—is proposed future work, not implemented or evaluated here (cf. SIDE (Petroni et al., 2023) for an editor-facing retrieval-and-verification system). Condition C also uses fixed, hand-chosen fusion weights set on an earlier 70-example seed subset and later examined, but not re-tuned, on the 175-example development family; it should therefore be read as an *in-development diagnostic reference*, not a held-out estimate of deployable performance. (viii) We report *no* downstream generation results: an earlier Flan-T5 probe used contexts that changed under the corrected tie handling and has not been rerun, so downstream propagation remains an open

validation step. Whether these ranking errors propagate into generation errors depends on downstream model behavior and remains an important direction for future evaluation. We do not argue these limitations are harmless; rather, the justified conclusion is limited to the controlled, development-set diagnostic above.

Acknowledgments

We used generative AI assistance (Anthropic Claude) for literature search, research framing, construction of controlled diagnostic examples, LLM-assisted label quality control, design and implementation of evaluation metrics, code generation and debugging, experiment planning, analysis, and manuscript drafting and revision. Reported results and references were cross-checked against the released artifacts and cited sources. The author reviewed the study design and final manuscript and takes full responsibility for the work.

Code, data, and result artifacts are available at <https://github.com/kelu-look/when-relevance-is-not-evidence> (release tag v1.0-akbc2026).

References

- Guy Alt, Eran Hirsch, Serwar Basch, Ido Dagan, and Oren Glickman. 2026. [User-centric evidence ranking for attribution and fact verification](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7237, Rabat, Morocco. Association for Computational Linguistics.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 632–642.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. [The PASCAL recognising textual entailment challenge](#). In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment (MLCW 2005)*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190. Springer.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. [Enabling large language models to generate text with citations](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6465–6488.
- Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. [Breaking NLI systems with sentences that](#)

- require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 650–655, Melbourne, Australia. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *International Conference on Learning Representations (ICLR)*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. [Entity-based knowledge conflicts in question answering](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7052–7063.
- Bill MacCartney and Christopher D. Manning. 2007. [Natural logic for textual inference](#). In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 193–200, Prague. Association for Computational Linguistics.
- R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. [Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- Fabio Petroni, Samuel Broscheit, Aleksandra Piktus, Patrick Lewis, Gautier Izacard, Lucas Hosseini, Jane Dwivedi-Yu, Maria Lomeli, Timo Schick, Michele Bevilacqua, Pierre-Emmanuel Mazaré, Armand Joulin, Edouard Grave, and Sebastian Riedel. 2023. [Improving Wikipedia verifiability with AI](#). *Nature Machine Intelligence*, 5:1142–1148.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. [Measuring attribution in natural language generation models](#). *Computational Linguistics*, 49(4):777–840.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. [Get your vitamin C! robust fact verification with contrastive evidence](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 624–643, Online. Association for Computational Linguistics.
- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: a large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 809–819.
- Orion Weller, Michael Boratko, Iftekhar Naim, and Jinyuk Lee. 2026. [On the theoretical limitations of embedding-based retrieval](#). In *The Fourteenth International Conference on Learning Representations (ICLR)*.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1112–1122.
- Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. [Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts](#). In *International Conference on Learning Representations (ICLR)*.
- Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024. [Knowledge conflicts for LLMs: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8541–8565.
- Yilong Xu, Jinhua Gao, Xiaoming Yu, Yuanhai Xue, Baolong Bi, Huawei Shen, and Xueqi Cheng. 2025. [Training a utility-based retriever through shared context attribution for retrieval-augmented language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 629–648, Suzhou, China. Association for Computational Linguistics.
- Fan Yang, Eduard Dragut, and Arjun Mukherjee. 2021. [Improving evidence retrieval with claim-evidence entailment](#). In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1553–1558, Held Online. INCOMA Ltd.
- Zeyu Zhang, Alessandro Moschitti, and Thuy Vu. 2025. [Retrieving support to rank answers in open-domain question answering](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 35098–35105, Suzhou, China. Association for Computational Linguistics.