AKBC 2026 / EMNLP WORKSHOP

BUDAPEST · 28 OCT 2026

When Relevance Is Not Evidence:

A Controlled Diagnostic of Claim-Conditioned Support Ranking

Ke Lu Independent Researcher look970726@gmail.com

18–34%

TIE-AWARE FOCAL-PAIR ERROR

High relevance ≠ evidential support

Five query-only rankers · 175 constructed development examples

Not a real-world prevalence estimate

WHAT THIS STUDY ADDS

A seven-cue controlled diagnostic of support-ordering failures, with explicit information-access distinctions and role-free, tie-aware evaluation.

01 A concrete contrast

NEGATION · FICTIONAL MINI-WORLD

TARGET CLAIM

Under the Meridian Data Standard, the CX-2 cipher is not permitted for storing customer records.

SUPPORTS

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 · CONTRADICTION

Under the Meridian Data Standard, CX-2 is a permitted cipher for storing customer records.

Similar words. Opposite support relation.

02 Controlled diagnostic

175 development examples | **7 cues × 25 examples**

TASK DEFINITION

Query q

Target claim c

→

5-passage shortlist

1 support / 1 targeted non-support
3 neutral distractors

Probe: does non-support outrank support?

Condition A scores each candidate passage against query q, without the gold claim. Condition C additionally uses the gold answer and an NLI signal.

BM25 uses shortlist-local IDF (N = 5); not corpus-scale retrieval.

03 Read the comparison correctly

A **QUERY-ONLY RANKERS**
query + passage; no gold claim

C **GOLD-CLAIM NLI REFERENCE**
q + gold answer + NLI signal; diagnostic

D **GOLD-LABEL ORACLE**
gold support labels; upper bound

SWR = P(score(non-support) > score(support)) + ½ P(tie)

SWR compares the focal support/non-support pair. E[Support@1] evaluates the full five-candidate shortlist, averaging over uniform within-tie orderings. A vs C is not input-equivalent.

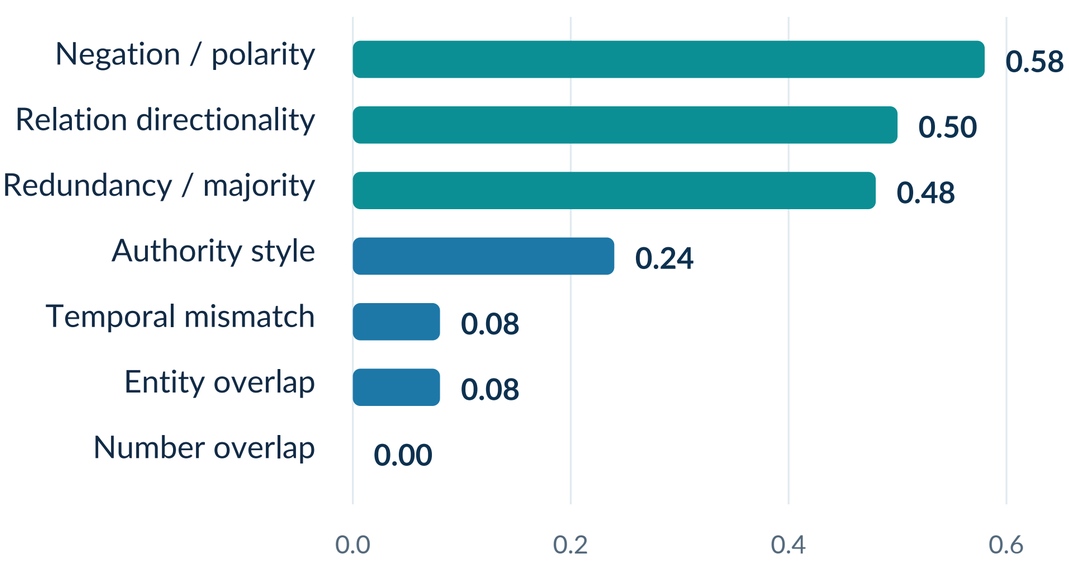
04 Systems on the same shortlists

SYSTEM	SWR ↓	E[Support@1] ↑
Lexical overlap	**0.246**	0.744
BM25	**0.280**	0.697
Dense E5-small	**0.274**	0.726
Dense MiniLM	**0.337**	0.663
CE relevance	**0.177**	0.823
BM25 + NLI	**0.086**	0.909
Oracle	**0.000**	1.000

A: no gold claim | C: gold answer | D: oracle

05 BM25 support-ordering errors

Cue-wise tie-aware SWR · n = 25 each



Largest BM25 errors: negation 0.58, directionality 0.50, redundancy 0.48. Cue strata are descriptive—not isolated causal effects.

06 Evaluation direction

PROPOSED FRAMEWORK (NOT IMPLEMENTED)

01 Query relevance

02 **Claim support**

03 **Claim contradiction**

04 Source attributes & provenance

SCOPE & LIMITATIONS

- 175 constructed development examples; same-project labels
- Five-passage shortlists, not corpus-scale retrieval
- Cue/relation confounds; no set-level majority test
- No downstream generation results or deployable method

Code, data & reproduction

Public controlled_175 diagnostic · scripts, metrics, model versions, results & tests

github.com/kelu-look/when-relevance-is-not-evidence

Frozen release: v1.0-akbc2026