# WHEN DOES EXTERNAL EVIDENCE HELP INDUCTIVE KNOWLEDGE GRAPH COMPLETION?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Knowledge graph foundation models (KGFMs) can reject implausible entities yet fail to distinguish the correct answer from relation-compatible alternatives. For both ULTRA and FLOCK, replacing random candidate pools with relation-compatible pools of matched pre-filter size reduces mean reciprocal rank by 0.238–0.261, exposing a fine-grained discrimination failure. We investigate candidate-conditioned evidence retrieval as a remedy: a KGFM proposes candidates, and an LLM verifies each hypothesized relation against externally acquired support. Holding the shortlists and verifier model fixed, we compare entity names, bounded graph context, and Web evidence, and observe two contrasting evidence regimes. On FB15k-237Ind v1, Web verification improves conditional mean reciprocal rank over graph verification by 0.077–0.154 and also outperforms the original KGFM ranking. On pooled NELL-995Ind v4 samples, graph context improves over names, whereas Web evidence yields no statistically reliable further gain. Retrieval analysis associates failures with missing answer support or competing evidence: evidence availability alone does not ensure discrimination. These findings demonstrate the conditional value of candidate-level evidence acquisition as a complement to structural ranking and motivate selective retrieval for resolving plausible alternatives under the evaluated protocols.

## 1 INTRODUCTION

Inductive knowledge graph completion (KGC) predicts missing links involving entities unseen during training. Knowledge graph foundation models (KGFMs) such as ULTRA (Galkin et al., 2024) and FLOCK (Kim et al., 2026) learn structural reasoning patterns that transfer across graphs. Their predictions are typically evaluated by ranking the held-out entity against the full entity vocabulary. This evaluation conflates two tasks: excluding candidates incompatible with the queried relation, and distinguishing the answer from alternatives that are already plausible. A model can perform well on the first task while remaining unreliable on the second, a distinction motivated by fine-grained evaluation and plausible-negative benchmarks (Meilicke et al., 2018; Safavi & Koutra, 2020).

We isolate this distinction by varying candidate composition while matching the pre-filter candidate-pool budget. The same model scores are evaluated against a random pool and against a relation-compatible pool. For both ULTRA and FLOCK, ranking performance drops substantially when the competing entities are relation-compatible. Because the comparison changes only which alternatives the answer must outrank, it exposes a difficulty that full-vocabulary ranking can obscure: selecting among plausible candidates is harder than rejecting unrelated entities.

A ranking failure alone does not reveal its cause. The model may fail to exploit a useful graph pattern, the evidence extractor may omit it, or the available context may offer little support for separating the remaining candidates. Figure 1 illustrates the last two possibilities. Observed facts about an actor's films and a city's neighbors do not directly verify the actor's birthplace, although each fact is relevant to one of the entities, whereas an external biography can support the specific candidate–relation pair. In another query, an observed employment relation already provides relevant support. These examples motivate our central question: *once a KGFM has produced plausible candidates, how does the information supplied for verification affect their ranking?*

We investigate candidate-conditioned external evidence as a remedy for this discrimination difficulty. A frozen KGFM supplies plausible entity IDs, and retrieval then seeks support for the specific



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

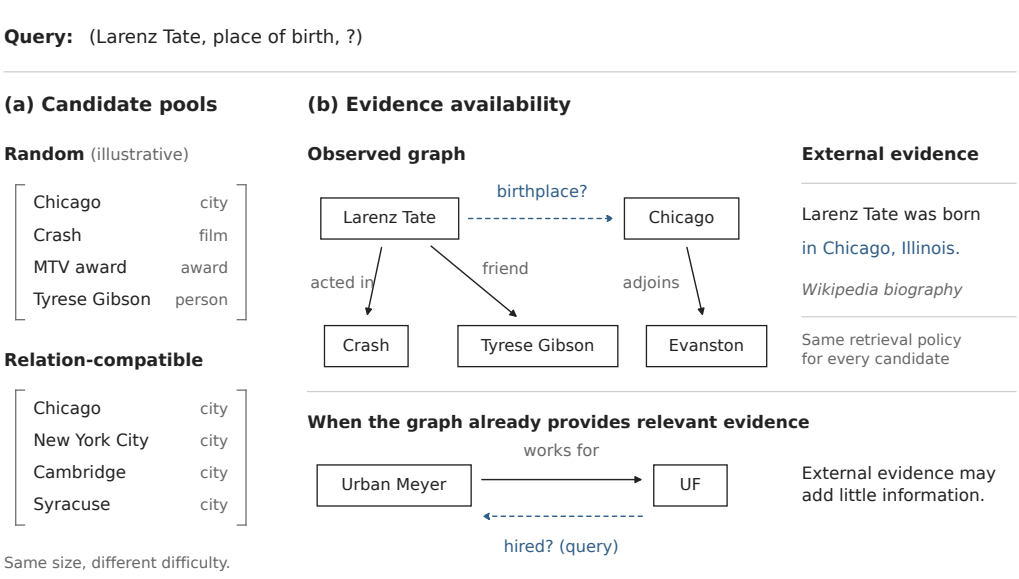



Figure 1: **Candidate discrimination and evidence availability.** (a) Illustrative equal-size pools contrast coarse filtering with ranking among relation-compatible entities; the cities are taken from a real shortlist. (b) Selected observed facts from an FB15k-237Ind query do not directly establish Larenz Tate's birthplace, whereas external biographical evidence supports Chicago. The lower NELL-995Ind example illustrates relevant graph support, not a measured gain over names-only verification. Solid arrows denote observed relations with shortened labels; dashed arrows denote query hypotheses. Only one external fact is displayed; acquisition applies the same candidate-wise eligibility policy, and retained-source limitations are documented in Appendix L.

relation between the known entity and each candidate. The same frozen LLM scores the shortlist under three conditions: NAME-ONLY, using entity names and parametric knowledge; GRAPH-ONLY, using bounded local facts and paths; and OPEN-WEB, using candidate-conditioned external evidence. Every comparison retains the same candidates, so improvements reflect reordering rather than increased candidate recall. Because the conditions use source-specific prompts, we compare verification protocols rather than estimate an isolated causal effect of the supplied text.

The remedy has a clear empirical boundary. On FB15k-237Ind v1, Web verification improves over both the graph condition and the original KGFM ranking for both backbones. On the pooled NELL-995Ind v4 samples, graph context improves over names, but Web verification yields no reliable further gain over graph verification. Missing answer support and support for competing candidates help characterize this contrast, although these retrospective groups do not themselves define a retrieval policy. All results concern queries whose answer is shortlisted; the candidate pools are benchmark-assisted, and evidence provenance precludes stronger claims.

Our contributions are threefold: we characterize the candidate-discrimination failure mode across two KGFM architectures, evaluate candidate-conditioned external evidence as a conditional remedy on fixed shortlists, and identify its benefits and limits on two factual benchmarks. The findings pertain to the supplied graph context and the complete external-evidence pipeline. They motivate selective retrieval without claiming that the full graph lacks useful information or that an adaptive retrieval gate has been validated.

## 2 RELATED WORK

**Inductive representations and graph reasoning.** GraIL (Teru et al., 2020) learns from enclosing subgraphs, while CoMPILE (Mai et al., 2021), RED-GNN (Zhang & Yao, 2022), and NBFNet (Zhu et al., 2021) develop message-passing or path-based reasoning. NodePiece (Galkin et al., 2022) represents entities through graph tokens, MorsE (Chen et al., 2022) transfers entity-independent

meta-knowledge, and InGram (Lee et al., 2023) further accommodates unseen relations through a relation graph. Earlier, DeepPath (Xiong et al., 2017) attributed differences in FB and NELL reasoning performance to the availability of paths. These studies motivate examining graph support; our experiments instead ask whether changing the verification context helps once candidate coverage is fixed.

**Knowledge graph foundation models.** ULTRA (Galkin et al., 2024) transfers reasoning through relation interactions, and FLOCK (Kim et al., 2026) uses anonymized random walks and probabilistic node–relation equivariance. We use both as frozen candidate generators. Semantic information can also enter the reasoner: SEMMA (Arun et al., 2025) incorporates textual relation semantics, and ProLINK (Wang et al., 2024) uses LLM-generated structural prompts for low-resource inductive reasoning. Whereas these approaches modify reasoning inputs or representations, we compare verification protocols for already generated shortlists.

**Candidate composition and evaluation.** Fine-grained KGC evaluation separates task categories that aggregate scores can obscure (Meilicke et al., 2018), and ranking metrics leave gaps in assessing completion quality (Pezeshkpour et al., 2020). CoDEx supplies plausible but verified-false triples (Safavi & Koutra, 2020). In homogeneous-graph link prediction, HeaRT (Li et al., 2023) replaces overly easy random negatives with heuristic hard candidates, and type-constrained KG learning makes explicit how plausible entities depend on the relation (Krompaß et al., 2015). Our diagnostic varies competitor composition for a fixed KGFM ranker while matching pre-filter pool budgets. Compatibility derives from benchmark-assisted relation roles rather than from an ontology or a learned hard-negative sampler, and no new ranker is trained.

**Language models and contextual KGC.** DKRL (Xie et al., 2016) and BLP (Daza et al., 2021) learn entity representations from descriptions, including for unseen entities. KG-BERT (Yao et al., 2019) scores textualized triples, and StATIK (Markowitz et al., 2022) combines entity descriptions with graph message passing for inductive KGC. CATS (Li et al., 2025b) uses latent type constraints and subgraph reasoning; KGR3 (Li et al., 2025a) combines context retrieval, reasoning, and reranking; and DrKGC (Xiao et al., 2025) integrates retrieved subgraph representations into LLM adaptation. KICGPT (Wei et al., 2023) pairs a KGC retriever with an LLM through structural in-context demonstrations, and search-based question answering has been applied to knowledge base completion (West et al., 2014). We focus instead on the benefit and boundary of external evidence for candidate discrimination, evaluated against names and bounded graph context on fixed shortlists. Names can elicit stored relational knowledge (Petroni et al., 2019), but factual probing may also reflect prompt artifacts and type guidance (Cao et al., 2021). This motivates an explicit names condition and caution in attributing gains solely to additional factual evidence.

**Selective retrieval.** Self-RAG (Asai et al., 2024) learns to retrieve on demand and to critique retrieved passages and generated answers, while Adaptive-RAG (Jeong et al., 2024) chooses among no, single-step, and iterative retrieval according to estimated question complexity. Selective retrieval is thus an established design principle. We contribute a KGC-specific empirical motivation: the value of external evidence relative to the supplied graph context differs across the evaluated samples. A policy that predicts this value before retrieval remains to be developed and tested.

## 3 Candidate-Conditioned Evidence Verification

We first expose the candidate-discrimination difficulty by comparing relation-compatible and random pools under a fixed ranker. To investigate an evidence-based remedy, we then decouple candidate generation from verification: a KGFM fixes the shortlist, candidate-conditioned retrieval collects support, and an LLM ranks the candidates. Names and observed-graph context serve as reference conditions for measuring the benefit of the external pipeline. No backbone or verifier parameters are updated.

### 3.1 Task and Candidate Composition

Let $G_{\text{obs}} = (V, R, E_{\text{obs}})$ be the observed inference graph, and let $E_{\text{eval}}$ contain the held-out evaluation triples. A query $q$ asks for the missing entity in $(h, r, ?)$ or $(?, r, t)$, with reference answer

$c^\star$. KGFM message passing and graph-evidence extraction use only $E_{\text{obs}}$. Evaluation covers both prediction directions and, following filtered ranking (Bordes et al., 2013), removes other benchmark-known answers to the same query.

Candidate-pool construction draws on a broader information set. The implemented benchmark diagnostic derives relation roles from $E_{\text{pool}} = E_{\text{obs}} \cup E_{\text{eval}}$; for tail prediction,

$$\mathcal{C}_r^{\text{rel}} = \{c \in V : \exists x,\ (x, r, c) \in E_{\text{pool}}\}, \tag{1}$$

with argument positions reversed for head prediction. We call these *benchmark-assisted relation-role pools*: held-out triples contribute role membership and filtering information, although their edges are absent from the inference graph. Relation compatibility therefore means occupying the relevant argument position, not being verified as a correct answer to the current query. This setup supports a filtered benchmark diagnostic rather than an answer-blind deployment protocol (Appendix M).

The composition diagnostic compares this pool (D1) with a uniformly sampled entity pool (D0) of the same pre-filter size. The reference answer is scored in both conditions, and only eligible competitors enter its filtered rank. Matching controls the initial pool budget, but filtering and answer membership can leave different final competitor counts. Random pools are fixed per relation and direction before evaluation. Appendix A details this distinction and the checkpoint settings.

### 3.2 Fixed Candidate Shortlists

For each evidence experiment, a frozen backbone scores entities on $G_{\text{obs}}$ and, after removing other benchmark-known answers, selects up to $K = 10$ entities from the relation-role pool:

$$\mathcal{S}_K(q) = \text{TopK}_{c \in \mathcal{C}_r^{\text{rel}} \setminus \mathcal{F}(q)}\, s_{\text{KGFM}}(q, c), \tag{2}$$

where $\mathcal{F}(q)$ excludes other known answers but retains the reference answer as a possible competitor. The reference answer is never inserted into the top-$K$ result. Queries whose relation-role pool contains fewer than two entities are excluded from the verification population, and a shortlist may contain fewer than ten entities after filtering. Closed candidate IDs also avoid the string-to-entity grounding errors of direct generation, examined in Appendix H.

All information conditions for a backbone/query pair share the same shortlist, and Condition A retains its KGFM order. Neither retrieval nor verification updates backbone parameters. Here, *frozen* describes this downstream use; it does not imply that every checkpoint was evaluated zero-shot. Specifically, the FB evidence experiments use split-trained ULTRA and FLOCK checkpoints, and the ULTRA NELL evidence experiment also uses a split-trained checkpoint, whereas FLOCK NELL uses an official pretrained checkpoint. Each backbone produces its own shortlist and eligible sample.

### 3.3 Three Verification Conditions

**Name-Only (B).** The LLM receives the known entity, relation description, candidate names, and prediction direction, and is asked to score the candidates using its general knowledge. This condition therefore includes parametric knowledge associated with entity identities; it is not a semantically empty control.

**Graph-Only (C).** Each candidate receives a bounded extraction from $G_{\text{obs}}$: at most six one-hop facts about the known entity, six about the candidate, and five connecting paths of length at most three. Facts and paths are ordered deterministically, with a 2,000-state search cap, and the target triple is checked to be absent before extraction. The prompt requests support from the supplied facts and paths and discourages external world knowledge. C thus tests one explicit graph-evidence representation and its use by the verifier, not all information recoverable from the observed graph (Appendix B).

**Open-Web (D).** Candidate-conditioned searches combine the known-entity label, relation phrase, and candidate label. Retrieved content is summarized and assigned one of four tiers: no connection (E0), topical relevance (E1), a specific indirect linkage (E2), or an explicit statement of the

queried relation (E3). Only E2/E3 summaries are passed as evidence; other candidates receive a no-passage marker (Appendix D). Collection instructions exclude answer identity and apply identical candidate-wise eligibility rules. Nevertheless, an audit found answer-referential comments in retained summaries; we report exclusion sensitivity in Appendix L and do not claim fully verified answer-blind execution. Benchmark targets also enter the role pools in Equation 1. D replaces, rather than augments, the graph-evidence block of C with external evidence.

### 3.4 Listwise Verification and Paired Comparisons

The verifier is `claude-haiku-4-5-20251001`, run at temperature zero with a 300-token output limit. A single call scores all candidates of a query, displayed in an order randomized with seed 42. Although evidence is candidate-specific, verification is listwise:

$$(s^X_{q,1}, \ldots, s^X_{q,k_q}) = f_\theta\big(q, \mathcal{S}_K(q), \{E^X_{q,i}\}_{i=1}^{k_q}; P_X\big), \qquad X \in \{B, C, D\}. \tag{3}$$

The names and graph prompts request integer scores from 0 to 10, and the Web prompt requests 0 to 100. A shared parser extracts candidate-indexed scores. Candidates are sorted by decreasing score, with ties broken by their stored KGFM order; a parsing failure falls back to the stored rank. Fourteen reused FB/FLOCK outputs retain the historical ordering described in Appendix A.

The model, decoding settings, and parsing procedure are held fixed, but $P_X$ contains condition-specific instructions and scales. The comparisons therefore measure performance under these implemented information conditions rather than under a pure intervention on evidence alone. Some retained Name-Only head-query templates also contain a directionally ambiguous arrow (exact variants in Appendix C). These differences limit the attribution of Graph$-$Names contrasts to graph content alone.

For sampled queries whose reference answer is shortlisted, let $r_X(q)$ denote its within-shortlist rank and $RR_X(q) = 1/r_X(q)$. We report

$$\widehat{\Delta}_{\text{graph}} = \frac{1}{n}\sum_q [RR_C(q) - RR_B(q)], \qquad \widehat{\Delta}_{\text{web}} = \frac{1}{n}\sum_q [RR_D(q) - RR_C(q)]. \tag{4}$$

The latter compares external-evidence verification with graph-evidence verification; it is not the effect of adding Web content to an otherwise identical graph prompt.

### 3.5 Sampling and Scope of Inference

FB uses 150 eligible queries per backbone. NELL uses 50 search-usable and 80 complementary-stratum queries, sampled separately with balanced head/tail counts. All three conditions are evaluated on the same sampled queries. The pooled NELL result describes this 130-query mixture rather than a population-weighted estimate. We report paired percentile bootstrap intervals with 10,000 resamples and seed 42, retaining all paired conditions in each draw. An interval that includes zero indicates an unresolved gain, not evidence of equivalence. Appendix E specifies the sampling design.

These are *conditional shortlist* metrics: they measure verification once candidate generation has retained the answer and must be interpreted alongside shortlist coverage. They are not full-entity KGC MRR. Likewise, contrasting dataset-level regimes motivates selective retrieval as a research direction but does not validate a query-level rule for when to retrieve.

## 4 Experiments

### 4.1 Experimental Setup

We use the GraIL inductive splits (Teru et al., 2020) FB15k-237Ind v1 and NELL-995Ind v4, derived from FB15k-237 (Toutanova & Chen, 2015) and NELL-995 (Xiong et al., 2017), respectively. Both factual datasets expose entity names for candidate-conditioned retrieval. FB v1 is the compact development setting in which the D0–D1 gap was established with split-trained backbones. NELL v4 provides a non-saturated counterpart: NELL v1 reaches D1 MRR 1.0 for both backbones,

Table 1: Filtered MRR under matched pre-filter candidate-pool budgets. D0 is random; D1 is relation-compatible. Comparisons are within rows. FB uses split-specific checkpoints; NELL uses official pretrained checkpoints. FLOCK uses one frozen test-time sample in this experiment.

| Dataset | Backbone | D0: random | D1: compatible | D1−D0 |
|---|---|---|---|---|
| FB15k-237Ind v1 | ULTRA | 0.806 | 0.568 | −0.238 |
| | FLOCK | 0.797 | 0.536 | −0.261 |
| NELL-995Ind v4 | ULTRA | 0.781 | 0.541 | −0.240 |
| | FLOCK | 0.818 | 0.574 | −0.244 |

whereas v4 retains a substantial gap. Earlier graph-only experiments covered FB and NELL v1–v4; Appendix N reports their results and the exploratory rationale for split selection. The evidence comparisons use these two splits for both backbones. Our primary evaluation follows the frozen pipeline's convention of merging inductive validation and test targets. The observed inference graph contains only `train_ind` edges; target edges are excluded from message passing and graph evidence. As specified in Section 3, relation-role pools and filtered evaluation also use held-out benchmark targets. This is a benchmark-informed diagnostic, not a simulation of candidate construction from observed edges alone.

**Backbones and verifier.** Checkpoint provenance differs across diagnostics and datasets; Appendix A lists each setting. The final FB/FLOCK results use clean split-specific candidate-generation ranks and shortlists. Verification is recomputed for 136 queries, and 14 existing verifier outputs are reused only where the candidate set is identical. For these 14, the A ranks come from the clean checkpoint, whereas B/C/D retain their historical display and tie orders. The remaining protocol settings follow Section 3; Appendix K reports a sensitivity check restricted to the 136 fresh outputs.

**Sampling and metrics.** For each backbone, FB uses a seed-42 simple random sample of 150 queries whose answer appears in the relation-compatible top-10 shortlist. NELL uses 50 search-usable queries (25 head and 25 tail predictions) plus 80 queries from the complementary label-quality stratum (40 head and 40 tail). Label quality is assessed before verification, and the complementary stratum can still yield usable evidence after retrieval. The 130-query NELL aggregate is an unweighted summary of this fixed sample, not a population-weighted benchmark estimate. Because the backbones have different eligible populations and are sampled separately, absolute differences between them are not paired model comparisons. Our primary outcome is conditional shortlist MRR. Within each backbone and sample, we report paired reciprocal-rank differences with percentile 95% confidence intervals from 10,000 bootstrap resamples (seed 42). These intervals capture query-sampling uncertainty, not training-seed or LLM-run variability. Appendices E and F give denominators and ranking conventions.

### 4.2 Candidate Composition Changes Ranking Difficulty

We first compare a uniformly sampled random pool (D0) with a relation-compatible pool (D1) of matched pre-filter budget. The answer's score serves as the ranking reference in both conditions. Because known alternative answers are filtered, the number of remaining competitors need not be identical. D1 budgets vary by relation and prediction direction: across evaluated queries, FB has mean 35.4, median 21, and range 1–98, including singleton pools, whereas NELL excludes singleton pools and has mean 193.2, median 262, and range 2–559. Both are substantially larger than the shortlists used in the subsequent verification experiment.

Table 1 shows a consistent pattern across all four settings. On FB, MRR falls from 0.806 to 0.568 for ULTRA and from 0.797 to 0.536 for FLOCK; on NELL, the corresponding drops are 0.781 to 0.541 and 0.818 to 0.574. The within-setting loss spans 0.238–0.261 MRR. Candidate composition thus changes difficulty even under matched pre-filter budgets: separating entities that already occupy the appropriate relation role is harder than rejecting random alternatives. This observation motivates the evidence study but neither explains the errors, measures the usable information in the full graph, nor establishes a ceiling for stronger graph-only rankers.

Table 2: Verification on fixed KGFM shortlists. Top: conditional MRR. Bottom: paired MRR differences with 95% bootstrap intervals; bold means the interval excludes zero. The verifier model is fixed, with condition-specific prompts. NELL combines 50 search-usable and 80 complementary-stratum queries without population weighting.

| Dataset | Backbone | $n$ | KGFM (A) | Names (B) | Graph (C) | Web (D) |
|---|---|---|---|---|---|---|
| FB-v1 | ULTRA | 150 | 0.710 | 0.780 | 0.743 | 0.897 |
| | FLOCK | 150 | 0.712 | 0.744 | 0.790 | 0.867 |
| NELL-v4 | ULTRA | 130 | 0.700 | 0.605 | 0.747 | 0.766 |
| | FLOCK | 130 | 0.753 | 0.598 | 0.744 | 0.729 |

| Dataset | Backbone | Graph−Names | Web−Graph | Web−Names |
|---|---|---|---|---|
| FB-v1 | ULTRA | $-0.036$ | $\mathbf{+0.154}$ | $\mathbf{+0.118}$ |
| | | $[-0.099, 0.025]$ | $[0.100, 0.209]$ | $[0.067, 0.169]$ |
| | FLOCK | $+0.046$ | $\mathbf{+0.077}$ | $\mathbf{+0.123}$ |
| | | $[-0.012, 0.103]$ | $[0.027, 0.129]$ | $[0.069, 0.176]$ |
| NELL-v4 | ULTRA | $\mathbf{+0.142}$ | $+0.019$ | $\mathbf{+0.161}$ |
| | | $[0.068, 0.217]$ | $[-0.040, 0.080]$ | $[0.091, 0.231]$ |
| | FLOCK | $\mathbf{+0.146}$ | $-0.016$ | $\mathbf{+0.131}$ |
| | | $[0.070, 0.221]$ | $[-0.085, 0.053]$ | $[0.066, 0.195]$ |

### 4.3 Two Evidence Regimes on Fixed Shortlists

We next fix each query's shortlist and compare the KGFM ranking (A), Name-Only verification (B), Graph-Only verification (C), and Open-Web verification (D). Table 2 reports condition means and paired contrasts. Names serve as a reference for entity identity and the LLM's parametric knowledge, not as an information-free baseline.

**FB: external evidence improves verification.** With ULTRA, supplying graph context changes MRR by $-0.036$ relative to names (95% CI $[-0.099, 0.025]$), whereas Web verification improves over graph verification by $+0.154$ ($[0.100, 0.209]$). The clean FLOCK experiment yields $+0.046$ for Graph$-$Names ($[-0.012, 0.103]$) and $+0.077$ for Web$-$Graph ($[0.027, 0.129]$). For both backbones, Web also improves over names and over the original KGFM ranking: Web$-$KGFM is $+0.187$ ($[0.133, 0.244]$) for ULTRA and $+0.155$ ($[0.095, 0.214]$) for FLOCK. The external pipeline thus improves candidate discrimination beyond both reference rankings. Graph$-$Names remains unresolved rather than equivalent to zero, while Graph$-$KGFM is positive for FLOCK ($+0.078$, $[0.037, 0.122]$). Appendix I reports all paired backbone comparisons.

**NELL: graph helps over names, with limited further Web benefit.** In the pooled NELL sample, graph verification improves over names by $+0.142$ for ULTRA ($[0.068, 0.217]$) and $+0.146$ for FLOCK ($[0.070, 0.221]$). The additional Web increments are small and uncertain: $+0.019$ ($[-0.040, 0.080]$) and $-0.016$ ($[-0.085, 0.053]$), respectively. This contrast holds for both backbones at the pooled-sample level, but the strata are heterogeneous. In FLOCK's search-usable stratum, Graph$-$Names is only $+0.034$ ($[-0.082, 0.151]$), compared with $+0.217$ ($[0.123, 0.311]$) in the complementary stratum. For ULTRA, both stratum-level graph increments are positive with intervals excluding zero. The pooled pattern therefore does not imply that every NELL relation or label-quality group is graph-informative (Appendix J).

This graph-over-names benefit largely recovers the performance lost when the KGFM ranking is replaced by names-only verification: neither backbone shows a reliable Graph$-$KGFM gain on NELL. Web$-$KGFM is positive for pooled ULTRA ($+0.066$, $[0.021, 0.116]$) but unresolved for FLOCK. The boundary is therefore the absence of a reliable Web-over-Graph advantage in these samples, not a claim that all NELL verification is ineffective. All gains remain conditional on shortlist coverage.

Table 3: Retrieval outcomes on FB and search-usable (SU) NELL queries. $P_T$: fraction with E2/E3 support for the target; $P_O$: fraction with E2/E3 support for at least one other candidate. Web−Graph gives paired MRR differences and 95% intervals. Evidence availability and competing support are descriptive correlates, not randomized interventions.

| Backbone | Sample | $n$ | $P_T$ | $P_O$ | Web−Graph [95% CI] |
|---|---|---|---|---|---|
| ULTRA | FB-v1 | 150 | 0.920 | 0.320 | $+0.154$ $[0.100, 0.209]$ |
| | NELL-v4 SU | 50 | 0.700 | 0.660 | $-0.017$ $[-0.125, 0.086]$ |
| FLOCK | FB-v1 | 150 | 0.893 | 0.467 | $+0.077$ $[0.027, 0.129]$ |
| | NELL-v4 SU | 50 | 0.780 | 0.780 | $+0.021$ $[-0.095, 0.142]$ |

### 4.4 Evidence Availability and Competing Support

We examine the saved retrieval records to characterize when Web evidence helps. A candidate has strong evidence if it receives an E2 or E3 record under the fixed evidence rubric, i.e., relation-relevant or direct support for the hypothesized fact. We distinguish support for the held-out answer from support for other shortlisted candidates. The latter constitutes an evidence collision relative to the benchmark target but need not be a false real-world statement, particularly for relations with multiple valid answers. For this retrospective analysis, we join answer identities to the retained evidence; execution-level provenance limits are examined separately below.

Table 3 compares FB with the search-usable NELL stratum. Strong answer evidence is found for 92.0% and 89.3% of FB queries under ULTRA and clean FLOCK, and for 70.0% and 78.0% of search-usable NELL queries. Nevertheless, Web−Graph remains uncertain on the latter stratum for both backbones, so availability alone does not fully account for the contrast. At least one non-target candidate receives strong evidence more often on search-usable NELL: 66.0% versus 32.0% for ULTRA, and 78.0% versus 46.7% for FLOCK. These differences are descriptive, since dataset, relation semantics, and evidence acquisition were not independently randomized.

Conditioning further on retrieval outcomes reveals both beneficial and adverse cases. In search-usable NELL queries without strong answer evidence, Web−Graph is $-0.259$ for ULTRA ($n = 15$; 95% CI $[-0.452, -0.066]$) and $-0.210$ for FLOCK ($n = 11$; $[-0.422, -0.015]$). In clean FB/FLOCK, queries with strong answer evidence and no competing strong evidence gain $+0.143$ ($n = 73$; $[0.079, 0.213]$), whereas collision cases gain $+0.042$ ($n = 61$; $[-0.046, 0.132]$). However, ULTRA's FB collision cases still improve by $+0.133$ ($n = 45$; $[0.019, 0.249]$). Collisions thus do not make retrieval uniformly unhelpful, and missing answer support is associated with risk rather than proven to be a necessary condition for failure. These post-hoc groups are small and uncorrected for multiple comparisons (Appendix E).

The analysis motivates, but does not implement, selective retrieval: a deployed gate cannot know whether the held-out answer received strong evidence and would instead require observable signals, such as disagreement or unresolved support among candidates.

### 4.5 Sensitivity and Scope of the Findings

Restricting the frozen samples to test-origin queries preserves the qualitative pooled contrast (Table 5, Appendix G). On FB, Web−Graph remains positive with intervals excluding zero for ULTRA ($n = 77$, $+0.196$) and clean FLOCK ($n = 73$, $+0.092$), while Graph−Names intervals include zero. On NELL, Graph−Names remains positive with intervals excluding zero for ULTRA ($n = 53$, $+0.119$) and FLOCK ($n = 60$, $+0.151$), whereas Web−Graph intervals include zero. These are descriptive subdomain checks within the original samples, not new test-population estimates.

**Support selection and tie breaking.** The FB Web−Graph advantage remains positive under uniform random tie breaking ($+0.141$/$+0.079$ for ULTRA/FLOCK; both intervals exclude zero). On NELL, Web scores depend substantially on the KGFM tie prior: under random tie breaking, Web−Graph becomes negative ($-0.140$/$-0.238$). Moreover, in none of the four settings does the final Web verifier reliably outperform a rule that ranks the existing support tiers. The benefit therefore belongs to the complete acquisition-and-verification pipeline and cannot be credited solely to the final LLM judge. Appendix K describes these controls and their annotation requirements.

**Evidence provenance.** Two retained ULTRA summaries contain explicit answer-referential comments. Excluding their queries preserves the Web−Graph pattern: $+0.155$ ($[0.101, 0.211]$, FB $n = 149$) and $+0.023$ ($[-0.038, 0.085]$, NELL $n = 129$). This check and the source-metadata subsets (Appendix L) qualify the observed benefit but do not certify that all remaining summaries are answer-blind or independent of benchmark-derived material.

## 5 Discussion and Limitations

**From plausible candidates to supported facts.** The failure mode separates two requirements of completion: identifying entities compatible with a relation and determining which candidate satisfies the particular query. This suggests a division of labor in which the KGFM supplies structurally plausible candidates and retrieval seeks support for their hypothesized facts. The FB results demonstrate the value of this combination on fixed shortlists. The support-tier controls locate this benefit at the pipeline level: the final LLM judge does not reliably improve on ranking the acquired support annotations.

**Evidence availability is not enough.** A retrieved passage may mention both entities without resolving their relation, or may support several competing candidates. Such cases distinguish retrieval coverage from useful discrimination. NELL illustrates this boundary: Web evidence is often available, yet it provides no reliable advantage over the graph condition. Missing answer support and competing evidence are descriptive correlates, not independently established causes, of this contrast. Moreover, a competing fact may be valid in the world but absent from the benchmark, so disagreement with the reference answer does not necessarily indicate faulty retrieval.

**What the comparison establishes.** The measured benefit concerns complete verification protocols. Bounded graph extraction may omit useful facts, so a weak graph condition does not establish that the full graph lacks information. Source-specific prompts, score scales, and the Name-Only head-direction ambiguity limit attribution to evidence alone; shared prompts, corrected directions, and richer graph contexts would sharpen this attribution.

**Scope and provenance.** Two factual splits and a single verifier do not establish broad generalization. Benchmark-assisted pools, answer-in-shortlist sampling, and the unweighted NELL mixture limit population and deployment claims. Names can invoke parametric knowledge, and private or anonymized entities may offer little retrievable support. Incomplete source attribution and answer-referential comments in retained summaries further prevent certifying fully independent, answer-blind acquisition. The reported sensitivity checks preserve the main pattern without resolving every provenance concern.

**Implications for selective retrieval.** A natural next step is to predict the value of acquisition from observable graph context, candidate disagreement, or verifier uncertainty. Such a policy must be evaluated on unseen queries and must account for acquisition cost, which grows with shortlist size. Our post-hoc support groups cannot serve directly as gate inputs because they require answer identity. The present study establishes neither an adaptive gate nor computational savings; rather, it provides evidence of both the opportunity and the boundary that such a policy must address.

## 6 Conclusion

Relation-compatible candidates expose a discrimination difficulty in ULTRA and FLOCK that is less apparent under size-matched random pools. Candidate-conditioned external evidence offers a conditional remedy: on the evaluated FB15k-237Ind v1 shortlists, the Web pipeline improves over both graph-based verification and the original KGFM ranking. On pooled NELL-995Ind v4 samples, graph context improves over names, whereas the further Web advantage remains unresolved. Evidence availability and competing support help delineate where this remedy succeeds or falters.

The supported finding concerns a complete evidence pipeline on fixed shortlists and is subject to prompt, sampling, and provenance limitations. It motivates evaluating evidence acquisition alongside ranking capacity, and developing selective retrieval policies that test whether the added information justifies its cost.

## AI USE STATEMENT

Generative AI tools assisted with manuscript drafting and revision, code inspection, analysis scripting, interpretation of recorded results, feedback on methodological descriptions, and figure preparation. The frozen LLM verifier is also an explicit component of the research method; its model, prompts, and scoring procedure are described in Section 3 and Appendix C. AI-generated explanations were not used as substitutes for recorded experimental results or source evidence.

## ETHICS STATEMENT

This study uses existing knowledge graph benchmarks and publicly accessible Web material, and reports no new human-subject experiments. Retrieved statements may be inaccurate, ambiguous, or outdated, and a verifier score does not constitute independent confirmation of a real-world fact. The findings are limited to the reported benchmark setting and do not justify unrestricted retrieval of information about private entities.

## REPRODUCIBILITY STATEMENT

The manuscript specifies the candidate pools, graph extraction limits, verification prompts, score parsing, sampling design, and checkpoint settings. The supplementary materials describe the stored artifacts needed to reproduce the reported summaries, and the reported tables summarize saved model outputs and evidence records. Because live Web results and hosted-model behavior may change, exact replay requires the evidence and model outputs retained from the original runs.

## REFERENCES

Arvindh Arun, Sumit Kumar, Mojtaba Nayyeri, Bo Xiong, Ponnurangam Kumaraguru, Antonio Vergari, and Steffen Staab. SEMMA: A semantic aware knowledge graph foundation model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31825–31848, 2025. doi: 10.18653/v1/2025.emnlp-main.1621. URL https://aclanthology.org/2025.emnlp-main.1621/.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2787–2795, 2013.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1860–1874, 2021. doi: 10.18653/v1/2021.acl-long.146. URL https://aclanthology.org/2021.acl-long.146/.

Mingyang Chen, Wen Zhang, Yushan Zhu, Hongting Zhou, Zonggang Yuan, Changliang Xu, and Huajun Chen. Meta-knowledge transfer for inductive knowledge graph embedding. In *Proceedings of the 45th International ACM SIGIR Conference*, pp. 927–937, 2022. doi: 10.1145/3477495.3531757.

Daniel Daza, Michael Cochez, and Paul Groth. Inductive entity representations from text via link prediction. In *Proceedings of The Web Conference*, 2021. doi: 10.1145/3442381.3450141.

Mikhail Galkin, Etienne Denis, Jiapeng Wu, and William L. Hamilton. NodePiece: Compositional and parameter-efficient representations of large knowledge graphs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xMJWUKJnFSw.

Mikhail Galkin, Xinyu Yuan, Hesham Mostafa, Jian Tang, and Zhaocheng Zhu. Towards foundation models for knowledge graph reasoning. In *International Conference on Learning Representations (ICLR)*, 2024.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050, 2024. doi: 10.18653/v1/2024.naacl-long.389. URL https://aclanthology.org/2024.naacl-long.389/.

Jinwoo Kim, Xingyue Huang, Krzysztof Olejniczak, Kyungbin Min, Michael M. Bronstein, Seunghoon Hong, and İsmail İlkan Ceylan. Flock: A knowledge graph foundation model via learning on random walks. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.01510.

Denis Krompaß, Stephan Baier, and Volker Tresp. Type-constrained representation learning in knowledge graphs. In *International Semantic Web Conference (ISWC)*, volume 9366, pp. 640–655. Springer, 2015. doi: 10.1007/978-3-319-25007-6_37.

Jaejun Lee, Chanyoung Chung, and Joyce Jiyoung Whang. InGram: Inductive knowledge graph embedding via relation graphs. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 18796–18809, 2023. URL https://proceedings.mlr.press/v202/lee23c.html.

Juanhui Li, Harry Shomer, Haitao Mao, Shenglai Zeng, Yao Ma, Neil Shah, Jiliang Tang, and Dawei Yin. Evaluating graph neural networks for link prediction: Current pitfalls and new benchmarking. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Muzhi Li, Cehao Yang, Chengjin Xu, Xuhui Jiang, Yiyan Qi, Jian Guo, Ho-fung Leung, and Irwin King. Retrieval, reasoning, re-ranking: A context-enriched framework for knowledge graph completion. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4349–4363, 2025a. doi: 10.18653/v1/2025.naacl-long.221.

Muzhi Li, Cehao Yang, Chengjin Xu, Zixing Song, Xuhui Jiang, Jian Guo, Ho-fung Leung, and Irwin King. Context-aware inductive knowledge graph completion with latent type constraints and subgraph reasoning. In *AAAI Conference on Artificial Intelligence*, pp. 12102–12111, 2025b. doi: 10.1609/aaai.v39i11.33318.

Sijie Mai, Shuangjia Zheng, Yuedong Yang, and Haifeng Hu. Communicative message passing for inductive relation reasoning. In *AAAI Conference on Artificial Intelligence*, pp. 4294–4302, 2021. doi: 10.1609/aaai.v35i5.16554.

Elan Markowitz, Keshav Balasubramanian, Mehrnoosh Mirtaheri, Murali Annavaram, Aram Galstyan, and Greg Ver Steeg. StATIK: Structure and text for inductive knowledge graph completion. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 604–615, 2022. doi: 10.18653/v1/2022.findings-naacl.46. URL https://aclanthology.org/2022.findings-naacl.46/.

Christian Meilicke, Manuel Fink, Yanjie Wang, Daniel Ruffinelli, Rainer Gemulla, and Heiner Stuckenschmidt. Fine-grained evaluation of rule- and embedding-based systems for knowledge graph completion. In *The Semantic Web – ISWC 2018*, volume 11136, pp. 3–20. Springer, 2018. doi: 10.1007/978-3-030-00671-6_1.

Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H. Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2463–2473, 2019. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. Revisiting evaluation of knowledge base completion models. In *Automated Knowledge Base Construction*, 2020. URL https://www.akbc.ws/2020/papers/1uufzxsxfL.

Tara Safavi and Danai Koutra. CoDEx: A comprehensive knowledge graph completion benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 8328–8350, 2020. doi: 10.18653/v1/2020.emnlp-main.669.

Komal K. Teru, Etienne Denis, and William L. Hamilton. Inductive relation prediction by subgraph reasoning. In *International Conference on Machine Learning (ICML)*, 2020.

Kristina Toutanova and Danqi Chen. Observed versus latent features for knowledge base and text inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*, pp. 57–66, 2015. doi: 10.18653/v1/W15-4007. URL https://aclanthology.org/W15-4007/.

Kai Wang, Yuwei Xu, Zhiyong Wu, and Siqiang Luo. LLM as prompter: Low-resource inductive reasoning on arbitrary knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3742–3759, 2024. doi: 10.18653/v1/2024.findings-acl.224. URL https://aclanthology.org/2024.findings-acl.224/.

Yanbin Wei, Qiushi Huang, Yu Zhang, and James Kwok. KICGPT: Large language model with knowledge in context for knowledge graph completion. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8667–8683, 2023. doi: 10.18653/v1/2023.findings-emnlp.580.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd International Conference on World Wide Web*, pp. 515–526, 2014. doi: 10.1145/2566486.2568032. URL https://research.google/pubs/knowledge-base-completion-via-search-based-question-answering/.

Yongkang Xiao, Sinian Zhang, Yi Dai, Huixue Zhou, Jue Hou, Jie Ding, and Rui Zhang. DrKGC: Dynamic subgraph retrieval-augmented LLMs for knowledge graph completion across general and biomedical domains. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 16432–16445, 2025. doi: 10.18653/v1/2025.findings-emnlp.892. URL https://aclanthology.org/2025.findings-emnlp.892/.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. doi: 10.1609/aaai.v30i1.10329.

Wenhan Xiong, Thien Hoang, and William Yang Wang. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 564–573, 2017. doi: 10.18653/v1/D17-1060. URL https://aclanthology.org/D17-1060/.

Liang Yao, Chengsheng Mao, and Yuan Luo. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*, 2019.

Yongqi Zhang and Quanming Yao. Knowledge graph reasoning with relational digraph. In *The Web Conference (WWW)*, pp. 912–924, 2022. doi: 10.1145/3485447.3512008.

Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. Neural bellman-ford networks: A general graph neural network framework for link prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

# A Evaluation Protocol and Checkpoint Provenance

## A.1 Observed graph, query set, and filtering

The observed inference graph contains the triples in `train_ind.txt`, augmented with reverse edges for KGFM message passing. The loaders merge `valid_ind.txt` and `test_ind.txt` into the primary evaluation query set. These held-out triples are added neither to the message-passing graph nor to the graph-evidence extractor. The benchmark evaluator uses them to filter other known answers and, in the historical composition and shortlist implementations, to define relation-role pools. A test-only sensitivity analysis restricts the scored queries to the `test_ind.txt` portion of this merged population; it does not reconstruct a new observed graph or candidate-pool protocol.

For a tail query $(h, r, ?)$, filtering removes every entity $c \neq c^\star$ for which $(h, r, c)$ appears in the union of observed and held-out triples; head queries use the corresponding reversed argument pattern. Full-entity ranks are computed over the inference entity vocabulary. The KGFM ranking implementation counts any remaining competitor whose score equals the answer's score as outranking it:

$$r_{\mathrm{KGFM}}(q) = 1 + \sum_{c \in \mathcal{N}(q)} \mathbf{1}\{s_{\mathrm{KGFM}}(q, c) \geq s_{\mathrm{KGFM}}(q, c^\star)\}, \tag{5}$$

where $\mathcal{N}(q)$ is the relevant filtered competitor set. This pessimistic convention differs from the stable-order tie breaking used when reranking discrete LLM scores.

The verification populations additionally require a relation-role pool of at least two entities before query-specific filtering. The resulting stored populations contain 813 directional FB queries for each backbone, 2,892 NELL queries for ULTRA, and 2,893 NELL queries for FLOCK. These artifact-specific denominators should not be silently replaced by the total number of directional benchmark triples. In particular, the one-query NELL difference means that cross-backbone coverage figures do not share exactly identical denominators.

## A.2 What is matched in the composition diagnostic?

For each relation and prediction direction, let $m_r$ denote the cardinality of its benchmark-assisted relation-role pool. D0 draws $m_r$ entities uniformly without replacement from the inference vocabulary, whereas D1 uses the relation-role pool itself. Each sampled D0 pool is shared by all queries with that relation and direction rather than resampled per query, using seeds 42 and 43 for tail and head pools, respectively. The ranker computes the answer score separately and then counts filtered competitors whose scores are at least as large.

The comparison thus matches *pre-filter pool cardinality*, not final negative counts. The answer belongs to D1 by construction but need not belong to the random draw, and known-answer filtering can remove different numbers of entities from the two pools. The reported composition effect should be read with this qualification. The NELL official-checkpoint diagnostic excludes relation-role pools smaller than two, whereas the original FB diagnostic includes singleton pools. Neither diagnostic removes held-out relation-role membership from pool construction.

## A.3 Checkpoint settings differ between diagnostics

Table 4 separates the checkpoint settings used for candidate composition from those used for evidence verification. Each checkpoint remains frozen after its original training stage throughout the verification experiments. This neither turns a split-trained model into a zero-shot pretrained model nor justifies comparing absolute MRR across rows with different training histories.

The ULTRA evidence checkpoints are `ultra_v1.pth` and `ultra_nell_v4.pth`. The former was saved by the FB split-training script; the latter was produced by the NELL split-training script with 1,000 optimization steps. The NELL composition diagnostic instead loads the official `ultra_3g.pth` checkpoint. Because these are distinct models, the composition and verification studies establish related empirical patterns rather than a single checkpoint-identical causal sequence.

The original FB/FLOCK evidence study loaded an official checkpoint whose pretraining mixture included transductive FB15k-237, creating an inappropriate overlap with its inductive derivatives.

Table 4: Checkpoint provenance by experiment. "Split-trained" means trained from scratch on the training graph of the indicated inductive split. All checkpoints are frozen during the reported diagnostics.

| Dataset | Backbone | Composition | Evidence verification |
|---|---|---|---|
| FB v1 | ULTRA | Split-trained | Split-trained |
| FB v1 | FLOCK | Split-trained | Split-trained (corrected) |
| NELL v4 | ULTRA | Official `ultra_3g` | Split-trained |
| NELL v4 | FLOCK | Official pretrained | Official pretrained |

The paper therefore reports the completed correction based on the split-trained `flock_v1.pth` checkpoint, whose eligible population contains 615 of 813 queries. A fixed sample of 150 was drawn from this population: 136 rows received fresh verification, and 14 reused stored verifier ranks because the query and unordered candidate set matched a previous record. The KGFM reference rank was replaced with that of the corrected checkpoint. The canonical result is the complete 150-row artifact, not any intermediate, partially completed result.

This correction removes the inappropriate checkpoint from candidate selection. The 14 reused rows retain their historical listwise verifier outputs and were not regenerated under new display permutations or tie orders; the correction is therefore not an independent rerun of all 150 verifier calls. The clean-checkpoint label refers to this identified pretraining-overlap correction, not to a guarantee against every possible source of leakage or memorization in the LLM.

# B Graph Evidence Construction

## B.1 Candidate-specific extraction

For a known entity $x$ and candidate $c$, the graph extractor returns three blocks: incident facts about $x$, incident facts about $c$, and paths from $x$ to $c$, all constructed exclusively from the observed inference triples. Before construction, the implementation checks that the query's held-out target triple is absent from that graph and aborts upon detecting a violation. Answer metadata is used for this audit and for later evaluation, but is never included in the evidence blocks sent to the verifier.

Incident facts are sorted with outgoing before incoming edges, then by relation identifier and neighboring-entity identifier, and each endpoint retains its first six facts. This deterministic rule gives every candidate the same maximum budget, but it is not a learned relevance ranking and can omit a useful fact appearing later in the order. The known-entity facts are repeated in each candidate's evidence block, allowing the verifier to assess the candidate-specific relation without consulting a separate, unlabeled context.

Paths are simple paths, with no repeated entity and at most three edges. Traversal may follow either edge direction, and each rendered hop records its actual relation and orientation. The implementation uses depth-first search with at most 2,000 visited search states per endpoint pair, sorts the discovered paths by length, relation/direction sequence, and entity identifiers, and retains the first five. Although the original written specification describes bounded breadth-first search, the executable implementation, which generated the reported evidence, is bounded depth-first search. In high-degree regions, the state cap can terminate search before all eligible paths are discovered.

The maximum block therefore contains 12 incident facts and five short paths. Missing facts or paths are rendered explicitly rather than replaced with external information, and the block is shorter when fewer than the cap exist. Only the extraction caps, not the token counts, are equal across candidates. KGFM scores, reference-answer flags, and relation-pool sizes are excluded from this Graph-Only prompt.

## B.2 Interpretation of a weak graph condition

A small Graph-Only improvement admits several explanations: the observed graph may contain little useful information, the deterministic extractor may omit it, the textual representation may obscure it, or the LLM may fail to use it. The experiment does not disentangle these possibilities.

Accordingly, the evidence regimes characterize the supplied graph representation under the chosen verifier. They establish neither an information-theoretic insufficiency of the full graph nor that a stronger graph retriever or reasoning model would fail.

# C Verifier Prompts, Direction Handling, and Parsing

## C.1 Shared call configuration

Every condition uses the same API model identifier, `claude-haiku-4-5-20251001`, with temperature zero and a maximum of 300 output tokens. The user-message prompt lists all candidates and requests a single JSON object keyed by candidate index strings; no explanation or chain of reasoning is requested. Candidate order is randomized with seeded streams, constituting a fixed order policy rather than an average over repeated permutations. The corresponding experiment scripts preserve both the streams and the query traversal order. The retained outputs represent one model run per condition/query, except for single-candidate cases that some scripts resolve directly as rank one.

The shared parser first attempts JSON decoding; if that fails, a regular expression extracts numeric index–score pairs, and parsing succeeds only if every candidate index is present. Scores are converted to floating-point numbers without enforcing the requested integer scale or clipping to a fixed interval. A parsing failure retains the stored backbone rank. On success, scores are mapped back to the original candidate indices and sorted in descending order, with ties broken by stored KGFM order. The 14 reused FB/FLOCK outputs retain their historical tie order, as specified in Appendix A. A score tie thus inherits a structural prior from the KGFM. Because names and graph use a coarser requested scale than Web, tie frequency can differ across conditions even though raw scores are never compared between them.

## C.2 Name-Only instructions

The common Name-Only instruction is:

> You are evaluating candidate answers for a knowledge graph completion task. Use only your own general knowledge. Do not explain your reasoning.

It is followed by the known entity, a relation pattern or directional statement, and the numbered candidates. The prompt requests a score from 0 (definitely incorrect) to 10 (definitely correct) for each candidate and a JSON-only response. Entity names therefore expose both lexical identity and any relevant knowledge already stored in the verifier.

Two Name-Only template variants were retained. ULTRA FB uses a statement builder that renders a tail query as "known entity – relation – [candidate]" and a head query as "[candidate] – relation – known entity." The builder used for NELL and the FLOCK studies instead displays "known entity –[relation]→ candidate" in both directions, while labeling the candidate role as head or tail; in head queries, this arrow can conflict with the role label. The reported results retain the outputs of this historical prompt, and its impact has not been isolated. This implementation detail precludes treating the Names comparison as a perfectly matched, direction-controlled intervention.

## C.3 Graph-Only instructions

The graph prompt begins:

> You are verifying a candidate fact using only evidence from an observed knowledge graph. For the query relation and candidate below, assess how strongly the provided graph evidence supports this candidate compared with the other candidates. Do not rely on external world knowledge. Do not assume the candidate is correct because it appears in the list. Use only the provided graph evidence.

Subsequent fields identify the known entity, the relation phrase, and the head or tail role being predicted. Each numbered candidate is followed by its known-entity facts, candidate facts, and connecting paths. The final instruction requests an integer score from 0 (no support) to 10 (strong, specific support) for every candidate, returned as a JSON object.

### C.4 OPEN-WEB INSTRUCTIONS

The Web prompt begins:

> You are judging whether retrieved evidence supports a SPECIFIC knowledge-graph relation between a known entity and a candidate entity. Use ONLY the supplied passage for each candidate.

It supplies a direction-specific statement and, for each candidate, either one passage or a no-passage marker. Additional instructions reject generic topical association, support for a different relation, and the use of general world knowledge to override the passage. The scoring instruction is:

> For EACH candidate, rate 0–100 how strongly its SUPPLIED PASSAGE supports the SPECIFIC queried relation with the known entity (0 = passage gives no support / no passage found, 100 = passage explicitly confirms this exact relation).

This requested scale differs from that of Names and Graph. Furthermore, Graph and Web restrict the use of parametric knowledge, whereas Names explicitly invites it. These are fixed information conditions with different instructions: only the verifier model and API configuration, not the full prompt, are identical.

## D EXTERNAL EVIDENCE AND PROVENANCE

### D.1 LABELS, SEARCHES, AND EVIDENCE TIERS

FB entities use the stored entity-name mapping when available and otherwise fall back to the raw identifier. NELL labels are derived from their `concept:type:slug` identifiers, retaining type context. Relation phrases are deterministic verbalizations of the benchmark relation labels; notably, a Freebase relation path need not yield a fluent or unambiguous natural-language predicate. Label quality and relation verbalization are therefore part of the evaluated acquisition pipeline.

The candidate-conditioned search pattern combines the quoted known label, relation phrase, and candidate label. Collection instructions exclude the answer indicator from the retrieval task. Nevertheless, retained summaries contain answer-referential comments in identified cases (Appendix L), so intended policy must be distinguished from verified execution. Each summary consists of one or two sentences describing what the returned content supports. E3 denotes an explicit statement of the queried fact; E2 a concrete but indirect relational linkage, such as a shared named award, cast, team, or event context; E1 topical material without such a linkage; and E0 no retrieved connection. These are support categories, not calibrated probabilities that a benchmark answer is correct.

Only E2/E3 content is passed to the Web verifier; E0/E1 entries receive "no passage found linking these two entities for this relation." D therefore evaluates retrieval, summarization, tiering, and verification jointly, without isolating the search engine's contribution from evidence selection or summarization. An E2 linkage may still be insufficient for the precise relation, and evidence supporting another valid real-world answer need not support the benchmark's single reference label.

For the corrected FB/FLOCK collection, the retained protocol budgets one search call with the fixed query string for every non-self-referential candidate. It prohibits reformulation, retries of successful searches, and benchmark-dump sources, and assigns E0 to self-referential candidates. If a search fails or is skipped because of a tool limit, the query is left incomplete rather than assigned fabricated E0 evidence; the final 150-query artifact contains completed D ranks and excludes such placeholders. These constraints describe the retained collection; the artifacts provide no common provider-level latency or monetary-cost measurement across all earlier runs. Although the retained protocol identifies the search interface as `WebSearch`, it does not pin a common underlying search-engine version, index snapshot, or retrieval-assistant model across collections. Recomputing ranks from saved evidence is therefore more reproducible than repeating live acquisition.

### D.2 NELL LABEL QUALITY AND MISSING EVIDENCE

The label classifier assigns GOOD, USABLE, or BAD/OPAQUE based on identifier form. It recognizes interpretable year literals, parses alphabetic slugs, strips trailing disambiguation digits for

the quality check, and rejects unparseable or generic tokens. A query is search-usable only if the known entity and every shortlisted candidate pass the usability threshold. This definition treats all candidates equally rather than inspecting whether the reference answer is readable.

The complementary stratum is not discarded. Its sampled queries are evaluated with the same verifier, whereas candidate identifiers classified as unresolvable are not searched and are recorded as E0 with failed entity resolution. Hence, "not search-usable" does not mean that no candidate in a query can yield evidence: a single opaque distractor suffices to place an otherwise interpretable query in this stratum. The two strata therefore reflect label-level acquisition conditions and do not directly test graph sufficiency.

### D.3 What the source audit establishes

The stored audit distinguishes identifiable independent pages, benchmark/KG-derived material, and entries with unclear or synthesized provenance. The historical FLOCK collection contains 2,381 candidate-evidence records: 356 with an identifiable independent source, 2,025 classified as unclear or synthesis, and zero flagged as benchmark/KG-derived. These counts refer to that historical collection, including its superseded FB shortlist, and do not constitute a complete provenance census of the corrected FB experiment.

A zero count of flagged benchmark sources means only that none was detected in the recorded evidence. It does not exclude contamination in sources with missing URLs, in synthesized records, or in the verifier's pretraining data. Some records preserve only a domain or synthesis label rather than a full URL and retrieval timestamp. The retained evidence thus supports an audit of recorded sources, not a fully replayable or independently certified contamination-free retrieval process.

## E Sampling and Statistical Interpretation

### E.1 Eligible populations and sample sizes

Shortlist eligibility is determined before verifier outcomes are inspected. On FB, ULTRA retains the reference answer in 634 of 813 stored queries, and corrected FLOCK in 615 of 813. Each evidence experiment uses a uniform sample of 150 eligible queries drawn with seed 42. Sampling is backbone-specific: replication refers to the direction of the information-condition effects, not to paired comparisons between ULTRA and FLOCK on identical query sets.

On NELL, ULTRA has 2,171 eligible queries (1,584 search-usable and 587 in the complementary stratum), and FLOCK has 2,489 (2,033 search-usable and 456 in the complementary stratum). Within each backbone, 50 search-usable queries are sampled with 25 per prediction direction, and 80 complementary-stratum queries with 40 per direction. For example, the latter ULTRA population contains 299 head and 288 tail queries, yielding inclusion probabilities $40/299$ and $40/288$; FLOCK has 255 head and 201 tail queries in that stratum, with probabilities $40/255$ and $40/201$. Balanced sampled directions therefore do not imply equal sampling probabilities in the underlying population.

The pooled NELL analysis combines all 130 observations with equal sample weight, so its stratum weights are $50/130$ and $80/130$ rather than the eligible population shares; it also weights head and tail equally within each stratum. This secondary sample summary is useful for checking the broad pattern but must not be interpreted as the expected performance on a uniformly sampled NELL query. Separate stratum results are necessary because effects can differ substantially across strata.

### E.2 Paired intervals and uncertainty

For each displayed comparison, the bootstrap resamples query indices with replacement and retains all compared conditions for each selected index. We generate 10,000 draws with seed 42 and take the 2.5th and 97.5th percentiles of the mean reciprocal-rank difference. The same procedure applies to the canonical corrected FB/FLOCK result. The pooled NELL intervals resample the pooled observations without reweighting them to the population strata.

These intervals describe resampling uncertainty in the retained query sample. They do not incorporate uncertainty from new model seeds, repeated LLM calls, candidate-order permutations, or

a changing Web index. Nor does the query-level bootstrap cluster head and tail predictions of the same triple or queries sharing an entity. No multiplicity correction is applied across the reported contrasts. We therefore emphasize effect sizes, intervals, and directional consistency across backbones, rather than treating every subgroup interval as an independent confirmatory test. A nonsignificant Web−Graph difference provides no evidence of a reliable gain under this design, but it does not establish equivalence.

### E.3 Post-hoc evidence diagnostics

After retrieval is complete, answer identity can be joined to the saved evidence tiers for analysis. "Strong answer evidence" means that the reference candidate has E2 or E3 content. A selective case has strong answer evidence and no competing candidate with E2/E3 content, whereas a collision case includes a competitor with strong evidence. These group labels are assigned retrospectively for analysis, independently of the answer-referential comments identified in some retained summaries. They describe the acquired evidence and cannot serve directly as deployable retrieval-gating features, because the correct candidate is unknown at inference time. Associations between these groups and ranking gains support mechanistic hypotheses but do not identify causal effects of tier assignment or source quality.

## F Conditional Metrics and Population Accounting

The principal evidence metric is conditional shortlist MRR,

$$\widehat{\mathrm{MRR}}_X^{\mathrm{cond}} = \frac{1}{n} \sum_{q \in \mathcal{Q}_{\mathrm{sample}}} \frac{1}{r_X(q)}, \qquad c^\star(q) \in \mathcal{S}_K(q). \tag{6}$$

It measures whether verification improves the answer's position once the candidate generator has retained it. Conditional Hits@10 is necessarily uninformative when every retained list contains at most ten candidates; the relevant coverage statistic is shortlist Recall@10, measured before this eligibility conditioning. In the evaluated setup, neither a high conditional MRR nor evidence that improves it can recover a candidate-generation miss.

Earlier accounting artifacts additionally define a hybrid rank:

$$\widetilde{r}_X(q) = \begin{cases} r_X(q), & c^\star(q) \in \mathcal{S}_K(q), \\ r^{\mathrm{full}}_{\mathrm{KGFM}}(q), & \text{otherwise.} \end{cases} \tag{7}$$

For an eligible answer, this convention treats the within-shortlist rank as a global rank; otherwise, it retains the original full-entity rank. This is an explicit accounting choice, not the result of simply inserting a reranked block into the original full-entity ordering, since entities outside the relation-role pool may originally outrank the shortlist. The paper does not use this hybrid summary to claim standard full-entity KGC performance or superiority over graph-only baselines.

Population inference would additionally require weights matching the sampling design. Let $\mathcal{Q}_0$ be the census of queries whose answer lies outside the shortlist, and let $\mathcal{Q}_j$ be mutually exclusive eligible strata, including prediction direction when its sampling fraction differs. For a chosen rank convention, a stratified estimate would take the form

$$\widehat{\mu}_X = \frac{1}{N} \left[ \sum_{q \in \mathcal{Q}_0} \frac{1}{r^{\mathrm{full}}_{\mathrm{KGFM}}(q)} + \sum_j |\mathcal{Q}_j| \, \overline{RR}_{X,j} \right]. \tag{8}$$

The stored historical NELL estimators weight searchability strata but do not fully adjust their balanced head/tail samples to the separate population direction counts; we therefore do not present those summaries as unbiased population estimates. Multiplying shortlist recall by the pooled conditional MRR is likewise inadequate, as it both changes the ranking universe and discards the contribution of candidate-generation misses.

Finally, all conclusions remain restricted to the tested graph representations, verifier, datasets, and checkpoint settings. Stronger graph retrieval, alternative prompt scales, unambiguous direction rendering, multiple verifier models, and a fixed-prompt evidence ablation could separate evidence acquisition from utilization more sharply. These remain open tests rather than capabilities demonstrated by the current framework.

Table 5: Sensitivity to retaining only test-origin queries from the frozen verification samples. These are descriptive subdomain summaries; NELL is pooled without population weighting. Entries are paired MRR differences with 95% bootstrap intervals. FB/FLOCK uses the clean checkpoint.

| Dataset | Backbone | $n$ | Graph−Names [95% CI] | Web−Graph [95% CI] |
|---|---|---|---|---|
| FB-v1 | ULTRA | 77 | $-0.073$ $[-0.160, 0.013]$ | $+0.196$ $[0.112, 0.283]$ |
| | FLOCK | 73 | $+0.046$ $[-0.036, 0.128]$ | $+0.092$ $[0.025, 0.160]$ |
| NELL-v4 | ULTRA | 53 | $+0.119$ $[0.002, 0.235]$ | $+0.022$ $[-0.063, 0.105]$ |
| | FLOCK | 60 | $+0.151$ $[0.043, 0.257]$ | $-0.002$ $[-0.108, 0.099]$ |

Table 6: Answer Recall@10 on the same 50 queries per dataset (25 head, 25 tail). Direct generation uses open-ended names; KGFM shortlists use relation-compatible benchmark pools. The comparison diagnoses the implemented generation protocols, with different candidate spaces.

| Candidate-generation protocol | FB-v1 | NELL-v4 |
|---|---|---|
| DIRECT-GRAPH-LLM | 0.38 | 0.22 |
| ULTRA relation-compatible shortlist | 0.82 | 0.78 |
| FLOCK relation-compatible shortlist | 0.78 | 0.78 |

## G TEST-ORIGIN SENSITIVITY ANALYSIS

Table 5 restricts each frozen verification sample to queries originating in `test_ind.txt`. This is a deterministic reanalysis of saved query identities and ranks: no backbone inference, retrieval, or verifier call is repeated. FB retains 77 ULTRA and 73 corrected FLOCK queries, and NELL retains 53 ULTRA and 60 FLOCK queries. On FB, Graph−Names intervals include zero and Web−Graph intervals are positive for both backbones. In the pooled NELL subsets, Graph−Names is positive while the Web−Graph interval includes zero. The direction of the principal sample-level contrast is thus preserved after removing validation-origin observations.

The subset sizes arose from the original sampling process; they were neither chosen nor powered as independent test-only experiments. Candidate pools, filtering, and evidence remain those of the original merged protocol, so these results test sensitivity to which observations are scored, not to rebuilding every component without validation labels. NELL also remains an unweighted mixture of its two sampled searchability strata. This analysis supports the stability of the reported sample pattern, subject to these protocol and sampling limitations.

## H CANDIDATE GENERATION AND ENTITY GROUNDING

A separate diagnostic examines what happens when the verifier receives no KGFM shortlist. DIRECT-GRAPH-LLM uses the same model at temperature zero to generate up to ten ranked answer names from the known entity, the relation, and at most six one-hop graph facts. A fresh seed-42 sample contains 50 queries per dataset (25 head and 25 tail predictions), drawn without conditioning on shortlist success. Generated names are mapped to benchmark IDs by fixed normalized label matching, and ambiguous or unmatched names remain unresolved.

Direct generation recalls the answer within the top ten for 38% of FB and 22% of NELL queries (Table 6), and only 16.2% and 45.2% of generated strings, respectively, map to the corresponding entity vocabulary. On the same queries, the relation-compatible ULTRA shortlists achieve 82% and 78% recall, and FLOCK achieves 78% on both datasets. The query denominators are matched, but the candidate spaces and contexts are not: the KGFM benefits from a benchmark-informed role pool, whereas the LLM generates open-ended strings from a small neighborhood. The result therefore demonstrates a grounding limitation of this particular direct-generation protocol, not the necessity of KGFMs or a general capacity ranking between graph models and LLMs. Verification on a closed shortlist avoids this string-to-entity bottleneck by construction.

## I VERIFICATION RELATIVE TO THE KGFM RANKING

Table 7 complements the evidence-condition contrasts in the main text; all comparisons use paired reciprocal ranks on the same conditional shortlists. On FB, Web improves over the original KGFM for both backbones. On NELL, Names reduces performance relative to the KGFM, and Graph's improvement over Names does not yield a reliable Graph−KGFM gain. ULTRA's pooled NELL Web−KGFM gain is positive, but its test-origin subset shows a $+0.043$ difference with interval $[-0.032, 0.120]$. For FLOCK on NELL, the Graph−KGFM and Web−KGFM intervals establish neither a gain nor a loss.

Table 7: Paired conditional MRR differences from the original KGFM ranking, with 95% bootstrap intervals. NELL uses the unweighted 50+80 mixture.

| Dataset | Backbone | Names−KGFM | Graph−KGFM | Web−KGFM |
|---|---|---|---|---|
| FB | ULTRA | $+.070$ $[.005, .137]$ | $+.033$ $[-.005, .073]$ | $+.187$ $[.133, .244]$ |
| | FLOCK | $+.032$ $[-.035, .101]$ | $+.078$ $[.037, .122]$ | $+.155$ $[.095, .214]$ |
| NELL | ULTRA | $-.095$ $[-.169, -.022]$ | $+.047$ $[-.008, .104]$ | $+.066$ $[.021, .116]$ |
| | FLOCK | $-.155$ $[-.220, -.089]$ | $-.009$ $[-.073, .055]$ | $-.024$ $[-.077, .027]$ |

## J NELL STRATA AND EVIDENCE BENEFITS

Table 8 disaggregates the samples pooled in the main results. Search usability is a label-quality classification made before verification; it neither guarantees that useful evidence will be retrieved nor indicates graph sufficiency. FLOCK's graph-over-names improvement is concentrated in the complementary stratum, whereas both ULTRA strata have positive Graph−Names intervals. No stratum has a Web−Graph interval excluding zero.

Table 8: NELL conditional paired contrasts by label-quality stratum.

| Backbone | Stratum | $n$ | Graph−Names (95% CI) | Web−Graph (95% CI) |
|---|---|---|---|---|
| ULTRA | Search-usable | 50 | $+0.131$ $[0.009, 0.253]$ | $-0.017$ $[-0.125, 0.086]$ |
| ULTRA | Complementary | 80 | $+0.149$ $[0.055, 0.241]$ | $+0.042$ $[-0.032, 0.117]$ |
| FLOCK | Search-usable | 50 | $+0.034$ $[-0.082, 0.151]$ | $+0.021$ $[-0.095, 0.142]$ |
| FLOCK | Complementary | 80 | $+0.217$ $[0.123, 0.311]$ | $-0.038$ $[-0.122, 0.047]$ |

## K SUPPORT SELECTION, SCORING, AND TIE SENSITIVITY

The four-tier control (Table 9) sorts E3>E2>E1>E0; the binary control sorts E2/E3 ahead of E0/E1; both break ties by the current KGFM order. Because these existing annotations already encode acquisition and support assessment, the controls measure the final verifier's ranking benefit beyond the annotations, not a cost-free alternative to acquisition. The binary rule mirrors passage availability, whereas the four-tier rule additionally distinguishes E0 from E1, although both are masked in the Web prompt. On the 136 newly scored FB/FLOCK queries, Web−Graph is $+.086$ ($[.031, .141]$) and Web−KGFM is $+.160$ ($[.099, .225]$). Web minus four-tier MRR is $-.010$ ($[-.041, .022]$), consistent with the full sample. The 14 reused verifier outputs retain their original tie order, whereas the tier control uses the corrected shortlist order.

Table 9: Support-annotation controls: conditional MRR and paired 95% intervals.

| Dataset | Backbone | Four-tier | Binary | Web | Web−Four-tier (95% CI) |
|---|---|---|---|---|---|
| FB | ULTRA | .9147 | .8869 | .8974 | $-.0172$ $[-.0444, .0094]$ |
| FB | FLOCK | .8815 | .8659 | .8673 | $-.0142$ $[-.0443, .0145]$ |
| NELL | ULTRA | .7829 | .7609 | .7659 | $-.0170$ $[-.0599, .0244]$ |
| NELL | FLOCK | .6978 | .7645 | .7285 | $+.0307$ $[-.0145, .0778]$ |

Web minus the binary rule on FLOCK/NELL is $-.036$ ($[-.074, -.001]$). Neither support control therefore establishes the necessity of a final LLM judge. Both controls share the original annotations and remain subject to their provenance limitations.

**Parsing and tie coverage.** All 1,587 retained nonempty score vectors can be reparsed with the original parser. Thirteen singleton Web queries required no call, and 80 ULTRA/NELL Names records lack raw responses, so their parsing and ties cannot be reconstructed. Table 10 reports counts with explicit denominators. A final rank equal to the KGFM rank does not by itself indicate a parse fallback. An all-tied query with multiple candidates inherits the full KGFM ordering.

Table 10: Score audit. Tie counts use parsed-output denominators. Missing raw responses are not counted as parsing successes or failures.

| Dataset | Backbone | Condition | Parsed | Truth tied | All tied | No call | Missing |
|---|---|---|---|---|---|---|---|
| FB | ULTRA | Names | 150 | 46 | 6 | 0 | 0 |
| | | Graph | 150 | 49 | 19 | 0 | 0 |
| | | Web | 144 | 24 | 11 | 6 | 0 |
| FB | FLOCK | Names | 150 | 54 | 5 | 0 | 0 |
| | | Graph | 150 | 51 | 14 | 0 | 0 |
| | | Web | 146 | 31 | 12 | 4 | 0 |
| NELL | ULTRA | Names | 50 | 36 | 3 | 0 | 80 |
| | | Graph | 130 | 53 | 6 | 0 | 0 |
| | | Web | 127 | 77 | 59 | 3 | 0 |
| NELL | FLOCK | Names | 130 | 91 | 17 | 0 | 0 |
| | | Graph | 130 | 49 | 5 | 0 | 0 |
| | | Web | 130 | 99 | 60 | 0 | 0 |

**Exact random-tie expectation.** With $h$ candidates strictly above the answer and $t$ sharing its score,

$$\mathbb{E}[RR \mid h, t] = \frac{1}{t} \sum_{j=h+1}^{h+t} \frac{1}{j}.$$

This quantity differs from the reciprocal of the expected rank. Table 11 recomputes it from retained outputs without new inference. Random tie breaking exposes the backbone's contribution; it is not a proposed deployment policy.

Table 11: Exact random-tie expected MRR and paired Web−Graph intervals.

| Dataset | Backbone | Graph | Web | Web−Graph (95% CI) |
|---|---|---|---|---|
| FB | ULTRA | .7342 | .8752 | $+.1409$ $[.0882, .1936]$ |
| FB | FLOCK | .7600 | .8387 | $+.0787$ $[.0200, .1358]$ |
| NELL | ULTRA | .7058 | .5657 | $-.1400$ $[-.2099, -.0693]$ |
| NELL | FLOCK | .6989 | .4612 | $-.2377$ $[-.3108, -.1641]$ |

## L Source-Metadata and Answer-Annotation Sensitivity

The initial ULTRA/FB source screen retains 49 queries with identifiable answer-evidence sources and 12 without supplied answer evidence. Recovering URL/title metadata from retained search logs increases the first count to 98; this recovery joins exact search strings and checks lexical correspondence, not page-level entailment. A stricter post-hoc screen requires complete metadata for every supplied candidate (96 queries). All subsets retain the 12 queries without answer evidence, including nine with no supplied evidence for any candidate. Table 12 reuses fixed ranks rather than rerunning the full sample with uncertain evidence removed.

A further lexical audit found two supplied-tier summaries that explicitly refer to the reference candidate. The FB county-disambiguation record is E2 and names the "truth candidate (Glens Falls)";

Table 12: ULTRA/FB metadata sensitivity. Complete metadata does not certify source content.

| Subset | $n$ | Web−Names (95% CI) | Web−Graph (95% CI) |
|---|---|---|---|
| Full sample | 150 | $+.118$ $[.067, .169]$ | $+.154$ $[.100, .209]$ |
| Initial answer-source screen | 61 | $+.054$ $[-.016, .124]$ | $+.170$ $[.090, .255]$ |
| Recovered answer metadata | 110 | $+.113$ $[.053, .173]$ | $+.161$ $[.097, .227]$ |
| All supplied candidates complete | 96 | $+.123$ $[.064, .183]$ | $+.166$ $[.103, .231]$ |

the NELL sports-team record is E3 and refers to the "truth candidate (Man Utd)." The retained evidence files identify these queries as 202_tail and 11_tail, respectively. Both tiers qualify for passage inclusion under the original Web protocol. This concrete answer-annotation concern, distinct from missing URLs, prevents certifying the collection as answer-blind. All other detected answer references in the scanned ULTRA and historical FLOCK collections occur in masked E0/E1 records. The scan constitutes neither exhaustive semantic certification nor a reconstruction of every historical API input.

Excluding the two identified queries from the saved ULTRA results leaves FB Web−Graph $+.155$ ($[.101, .211]$, $n = 149$) and NELL $+.023$ ($[-.038, .085]$, $n = 129$). These subset results preserve the average pattern but neither certify undetected records nor substitute for verification on cleaned evidence. The main table retains the original saved outputs, and no corrected full-sample rerun is claimed.

## M Observed-Only Role Membership

Reconstructing role sets from `train_ind.txt` alone yields a necessary condition for observed-only retrieval. The target occupies its queried role for 420/813 FB queries (51.7%), 2,168/2,892 ULTRA/NELL queries (75.0%), and 2,168/2,893 FLOCK/NELL queries (74.9%). These are role-pool recall ceilings, not newly measured Recall@10. Enforcing this membership removes 253/634 old ULTRA/FB hits, 230/615 FLOCK/FB hits, 443/2,171 ULTRA/NELL hits, and 551/2,489 FLOCK/NELL hits. Because reranking the smaller pool can also promote previously low-ranked eligible answers, the retained old hits do not equal the recall of a rebuilt shortlist. Held-out role membership thus materially affects eligibility in the reported diagnostic.

## N Split Selection and Development Scope

**FB v1: a compact setting with a demonstrated failure mode.** FB v1, with 1,093 observed entities and 1,993 observed triples, was the initial development setting. There, the split-trained ULTRA/FLOCK composition diagnostic establishes D0–D1 drops of .238/.261 (Table 1). Retaining this setting links the evidence study to an already demonstrated discrimination problem and to existing split-specific checkpoints. This choice does not imply that other FB splits lack the problem: graph-only development also covered v2–v4, whose frozen ULTRA D1 MRRs range from .544 to .621 (Table 13).

**NELL v4: room to measure candidate discrimination.** The official-checkpoint composition diagnostic on NELL v1 reaches D1 MRR 1.0 for both backbones in both directions, even after excluding pre-filter role pools smaller than two; it retains 71 tail and 160 head queries. Reconstructing filtering from the raw triples leaves no eligible competitor for 41 of those tail queries and 130 of the head queries. The ceiling is thus partly a structural property of this evaluation setting rather than evidence that general NELL completion is solved. With no D1 errors in that diagnostic, v1 leaves no ranking headroom for the proposed remedy.

NELL v4, in contrast, exhibits the D0–D1 gap in Table 1. The separate all-split, tail-only comparison in Table 13 also finds lower D1 accuracy on v4 than on v2/v3 for both official backbones. The observed graph of v4 is smaller than that of v3, although v2 is smaller still; we do not claim that v4 is the smallest non-saturated split. Together, these observations support v4 as a tractable setting with substantial remaining discrimination difficulty. For the official-checkpoint diagnostic, NELL

also avoids the FB15k-237 benchmark-family overlap in the three-graph pretraining mixture; the evidence-stage checkpoint choices remain those described in Appendix A.

Table 13: Graph-only development across splits. All D1 results here are filtered tail-only MRR with pre-filter role pools of at least two; $n$ counts eligible tail queries. Edges are original observed triples, before adding reverse edges. FB uses split-trained checkpoints and NELL uses official pretrained checkpoints. Dashes denote unavailable entries in these records. These are not the conditional LLM-verification results.

| Dataset | Split | Entities | Edges | $n$ | ULTRA D1 | FLOCK D1 |
|---|---|---|---|---|---|---|
| FB | v1 | 1,093 | 1,993 | 404 | 0.595 | 0.554 |
| FB | v2 | 1,660 | 4,145 | 916 | 0.621 | — |
| FB | v3 | 2,501 | 7,406 | 1,689 | 0.605 | — |
| FB | v4 | 3,051 | 11,714 | 2,768 | 0.544 | — |
| NELL | v1 | 225 | 833 | 71 | 1.000 | 1.000 |
| NELL | v2 | 2,086 | 4,586 | 931 | 0.567 | 0.600 |
| NELL | v3 | 3,566 | 8,048 | 1,616 | 0.592 | 0.640 |
| NELL | v4 | 2,795 | 7,073 | 1,446 | 0.523 | 0.568 |

**Scope of the selection evidence.** The FB v2–v4 entries come from the frozen-backbone structural-reranker experiments, and the FB v1 and NELL v1–v4 entries from the matched-backbone comparison. Because the composition diagnostic additionally evaluates head queries, its aggregated values need not equal the tail-only entries here. Separate FLOCK runs each use a single sampled walk realization, permitting small run-to-run differences. The broader development sequence also explored structural objectives, gates, and WN18RR splits. The final candidate-conditioned factual Web study focuses on FB v1 and NELL v4 and is not an all-split Web benchmark. These were exploratory choices informed by graph-only results, neither a preregistered selection nor proof that the two splits represent every inductive graph. In particular, v4's graph-only difficulty does not by itself predict a Web gain. Moreover, because NELL's composition and ULTRA evidence checkpoints differ, these stages do not measure the repair of a single checkpoint-identical D0–D1 gap.

**Dependence and multiplicity.** A sensitivity bootstrap resamples original triples, keeping their head/tail observations together. FB has 136/137 triples (ULTRA/FLOCK); NELL has 126/120. Graph$-$KGFM intervals are respectively $[-.007, .075]$, $[.036, .125]$, $[-.008, .104]$, and $[-.073, .055]$; Web$-$KGFM intervals are $[.132, .245]$, $[.093, .219]$, $[.019, .115]$, and $[-.079, .029]$. These contrasts agree qualitatively with query-level resampling. The check does not account for all shared-entity dependence, development-stage selection, or multiple subgroup comparisons, so subgroup intervals remain descriptive rather than family-wise guarantees.