

More Evidence Is Not Always More Support: Controlled Diagnostics for Evidence Aggregation in Fact Verification

Shaoxin Zhong Yuchen Su Qian Liu Michael J. Witbrock

School of Computer Science, University of Auckland
Auckland, New Zealand

scho963@aucklanduni.ac.nz ysu132@aucklanduni.ac.nz

liu.qian@auckland.ac.nz m.witbrock@auckland.ac.nz

Abstract

Final-context accuracy in multi-passage fact verification can hide whether an added passage contributes a new verdict-relevant fact or merely repeats evidence already present. We introduce a controlled same-prefix diagnostic derived from HoVer supporting-fact annotations to measure marginal evidence contribution. For each claim, a shared prefix covers one fact’s evidence while leaving another required fact absent; the next passage either repeats the covered fact or supplies the missing fact. The diagnostic contains 1,180 same-prefix pairs and an exact-repetition ladder testing whether repeated copies of one evidence sentence can change a verifier’s label. Supervised verifiers with similar clean F1 scores, from 0.6324 to 0.6528, differ sharply in eight-repeat flip rate, from 8.5% to 30.3%. Conservative pooling and natural-language-inference (NLI) scoring reduce duplicate sensitivity but give near-zero missing-fact gain, while gain-oriented systems improve gain but remain repetition-sensitive. Latent-Gated Evidence Dynamics (LGED), a compact stateful reference model, shows that low duplicate sensitivity and positive missing-fact gain can coexist. Surface controls and decision-level analyses support this diagnostic interpretation. Redundancy and reader tests extend the duplicate-sensitivity finding beyond HoVer, while prompted open-weight LLMs reveal a gap between role recognition and verdict updating.

1 Introduction

A verifier can reach the right final verdict for the wrong evidential reason. Retrieval may add several passages that restate an established fact while the evidence required for another part of the claim remains absent (Lewis et al., 2020; Asai et al., 2024; Liu et al., 2024). The key question is what the next passage contributes, not how long the evidence list becomes.

One HOVER example (Jiang et al., 2020) makes the failure mode concrete. The claim requires two pieces of evidence: one identifying who played Latka Gravas’s wife on *Taxi*, and another connecting that actress to *My First Mister*. Repeating the Taxi-side evidence and supplying the film-side evidence are different marginal additions, although both look like relevant retrieved passages to a standard verifier.

From HoVer’s supporting-fact annotations, we derive *evidence families*, where each family corresponds to one verdict-relevant fact in the multi-hop evidence. We turn this structure into a same-prefix contrast: one next passage repeats a family already covered by the prefix, while the other supplies a missing family. The diagnostic asks whether systems keep the verdict margin stable under repetition and increase the gold-verdict margin when the missing family appears. From 1,972 held-out evidence configurations, the main HOVER construction retains 1,180 consistency-checked same-prefix pairs after all checks. Figure 1 illustrates the contrast.

The comparison separates stability from gain and makes the unit of evaluation the marginal role of the next passage, rather than the final label after an assembled evidence list. Clean supervised verifiers can remain sensitive to repeated evidence; conservative pooling and NLI-style scoring can be stable but flat on missing-family gain; gain-oriented objectives can still over-count repetitions. Our compact stateful reference model, Latent-Gated Evidence Dynamics (LGED), occupies the low-repetition, positive-gain region and shows that low duplicate sensitivity and positive missing-family gain can coexist. The LLM results add a second view: models often recognize which continuation carries new verdict-relevant evidence when asked directly, while the same distinction is expressed much less strongly through standard verdict prompts.

This paper makes three contributions:

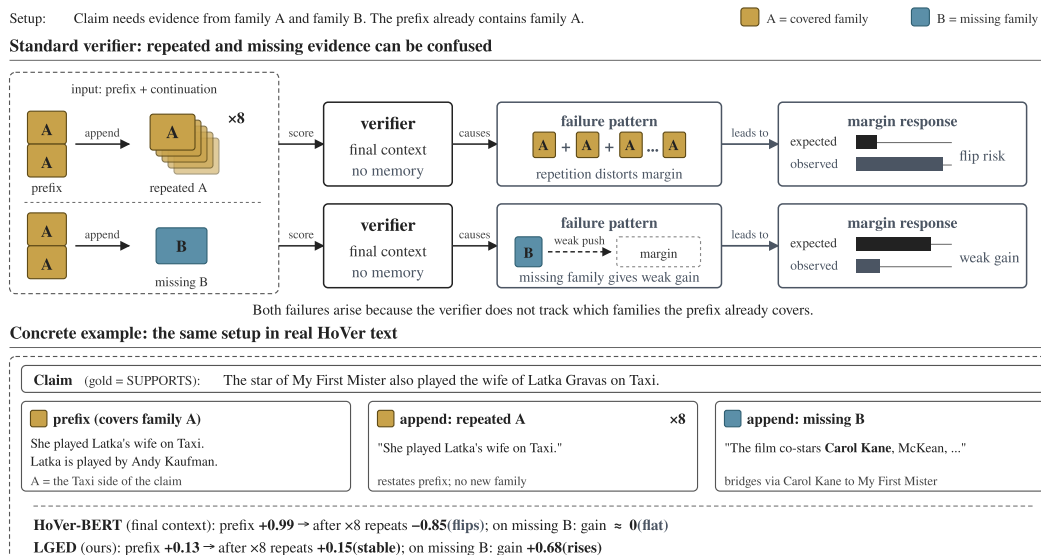


Figure 1: Same-prefix evidence aggregation diagnostic. The upper panel compares two next passages after the same claim and prefix: a repeated-family continuation and a missing-family continuation. The lower panel grounds the contrast in a HOVER example, where repeated evidence distorts the verifier margin and flips the predicted label. The metrics compare the verdict margin after each continuation with the margin at the shared prefix.

- A same-prefix diagnostic for marginal evidence aggregation, with paired probes for repeated-evidence stability and missing-family gain over 1,180 HOVER pairs (Section 3).
- A cross-system comparison showing that clean F1 in a narrow band, from 0.6324 to 0.6528, corresponds to repeated-evidence flip rates from 0.0854 to 0.3031 (Table 2).
- A compact stateful reference model (LGED) that occupies the low-flip, positive-gain region of the diagnostic, together with surface controls, reader stress tests, and LLM prompt analysis showing a separation between explicit role recognition and verdict updating (Sections 4, 5.2, and 5.4–5.6).

2 Related Work

Fact-verification research has broadened evidence settings, retrieval objectives, and behavioral tests for verifiers. Most evaluations, however, still score an assembled context rather than the marginal role of the next evidence passage.

Fact-verification benchmarks. Large-scale datasets such as FEVER and HOVER train and evaluate verifiers against textual evidence (Thorne et al., 2018; Jiang et al., 2020). HOVER is central here because its supporting-fact annotations let

us derive evidence families and then condition on which fact is already covered. Other benchmarks broaden fact verification to structured evidence, real-world question-answering evidence, and temporal or numerical evidence (Aly et al., 2021; Schlichtkrull et al., 2023; Strong and Vlachos, 2025). Our diagnostic instead holds the claim and prefix fixed and varies only the role of the next passage.

Evidence utility, sufficiency, and redundancy.

Evidence-utility work asks whether retrieved passages help the verifier rather than merely match the claim topically. Feedback-based retrieval shifts retrieval toward verifier utility (Zhang et al., 2023), and faithful revisiting trains verifiers to re-read evidence (Hu et al., 2023). Work on information aggregation, user-centric ranking, and redundancy-aware verification also shows that order, sufficiency, and repetition matter beyond final verdict accuracy (Hu et al., 2024; Alt et al., 2026; Lan et al., 2025; Liu et al., 2025). The same-prefix comparison targets a different distinction: deduplication may improve stability while weakening the gain assigned to useful missing-family evidence.

Behavioral perturbations and evidence-use diagnostics. Challenge sets, contrastive evidence, counterfactual examples, and CheckList-style tests reveal failures hidden by aggregate accuracy (Gardner et al., 2020; Schuster et al., 2021; Kaushik

et al., 2020; Ribeiro et al., 2020; Doddapaneni et al., 2024). Long-context and retrieval-augmented studies show sensitivity to irrelevant, repeated, or misplaced context (Liu et al., 2024; Asai et al., 2024; Yoran et al., 2024; Niu et al., 2024); long-context models may also contain relevant information without expressing it (Gao et al., 2024). Our diagnostic stays at the verifier interface and asks whether a margin change follows from the marginal role of the added passage. A label flip can reflect overcounting of repeated evidence or correct use of a missing fact; the paired construction separates these causes.

Thus, the evaluation unit is the next evidence passage: holding the claim and prefix fixed tests whether a system distinguishes new verdict-relevant evidence from evidence already present.

3 Evidence Aggregation Diagnostics

Using HOVER’s supporting-fact annotations, we define an *evidence family* as the set of sentences supporting one fact in the multi-hop evidence. Each diagnostic instance is a tuple $(c, y, \mathcal{F})$, where c is a claim, y its gold verdict, and $\mathcal{F}$ the evidence families. The prefix B covers at least one family while leaving another gold family missing. The tests keep c and B fixed while varying only the next passage’s role: a repeated-family continuation should leave the margin stable, whereas a missing-family continuation should increase the gold-verdict margin.

3.1 Evidence Family Episodes

For each held-out claim, we form a prefix B that contains evidence from at least one gold family and leaves another gold family missing. Each episode records covered and still-missing supporting facts. Construction checks preserve the original label, named entities, dates, numbers, and intended family assignment; Appendix F reports the full robustness analyses.

3.2 Redundancy Ladder

For a base evidence set B , we construct

$$B, \quad B + D_1, \quad B + D_2, \quad B + D_4, \quad B + D_8. \quad (1)$$

Here D_k contains k *exact repetitions* of a single sentence expressing the same covered fact. Powers-of-two depths cover mild to substantial repetition with four conditions. Shallow depths test mild duplication; D_8 tests whether eight repetitions of one fact can move the predicted label. We report the

eight step redundancy flip rate (FR-8), the fraction of examples whose predicted label changes between B and $B + D_8$. Shallower-depth checks are reported as robustness diagnostics in Appendix F.

3.3 Comparison Under the Same Prefix

Whereas Section 3.2 stacks exact copies of one sentence, the same-prefix comparison contrasts two distinct continuations under the same prefix. Let d be evidence from a family already covered by B , and let u be evidence from a missing gold family. This is a family-role intervention rather than a lexical minimal-pair construction. Because missing-family continuations necessarily change factual content, surface-matched subsets and surface audits are reported as sensitivity analyses. We compute

$$G_{dup} = M_y(B + d) - M_y(B), \quad (2)$$

$$G_{uf} = M_y(B + u) - M_y(B). \quad (3)$$

The two gains above define paired metrics. **Uncovered Family Preference** (UF Pref.) is the preference for the missing-family continuation, $P(G_{uf} > G_{dup})$. **Median UF Gain** is the median of $G_{uf} - G_{dup}$. Reporting both separates rank preference from gain magnitude.

3.4 Scoring and Construction Checks

All system comparisons use a shared binary probability margin,

$$M_y(E) = p_y(E) - p_{-y}(E). \quad (4)$$

Probability margins bound the scoring scale to $[-1, 1]$ and apply uniformly to binary and three-way verifiers after REFUTES / Not Enough Information (NEI) collapse. Training objectives may use native logits; cross-system reporting uses probability margins only.

The main construction retains 1,180 consistency-checked same-prefix pairs from 1,972 held-out HOVER evidence configurations. The repeated-family continuation is drawn from a covered family, the missing-family continuation from an absent gold family, and labels and entity/date/number spans are preserved. At evaluation time, no model receives construction labels or family identifiers. Diagnostic pairs are held out from LGED training; construction-derived auxiliary signals are used only on the training portion, with a development subset for hyperparameter selection. Appendix F reports robustness analyses and paired effect statistics.

Setting	Final verdict	Perturb.	Same prefix	Family aware	Shared margin
FEVER / HoVer-style verdict benchmarks	yes	no	no	partial	no
Evidence-utility and retrieval-ranking work	yes	partial	no	partial	no
Long-context and RAG perturbation tests	yes	yes	no	no	no
NLI / factuality adapters	yes	partial	no	no	partial
This diagnostic	yes	yes	yes	yes	yes

Table 1: Positioning of the proposed diagnostic relative to nearby evaluation settings. The proposed test keeps the claim and prefix fixed, varies the evidential role of the next passage, and reports systems through a shared probability-margin interface.

Table 1 positions this protocol against nearby evaluation settings. Verdict benchmarks lack controlled perturbation; long-context and retrieval-augmented generation (RAG) tests vary context but not the prefix; and NLI or factuality adapters lack family-aware roles. The combination of all five columns defines this work.

4 LGED as a Stateful Reference Model

Latent-Gated Evidence Dynamics (LGED) is a compact stateful reference model for the proposed diagnostic. It tests whether a bounded learned state can merge repeated evidence while still allowing a missing family to change the verdict margin. It differs from generic stateful aggregators through learned slots and auxiliary objectives derived from the same-prefix construction. Appendix A gives the full update equations and training losses.

4.1 Evidence representation and state

Each evidence step e_t is encoded by a fine-tuned DeBERTa-v3 pair encoder applied to the claim-evidence pair (He et al., 2023). The representation combines the [CLS] vector, verifier logits, probabilities, and lightweight metadata features:

$$\begin{aligned} x_t &= [\text{CLS}_t; \ell_t; p_t; m_t], \\ h_t &= \text{LN}(\text{ReLU}(W_{\text{enc}}x_t + b_{\text{enc}})). \end{aligned} \quad (5)$$

Using the same encoder features as the trainable controls (Section 5) keeps the comparison focused on aggregation rather than encoder strength.

LGED receives no construction labels or family identifiers at test time.

LGED maintains K learned slots $S_t \in \mathbb{R}^{K \times d}$ that represent evidence already seen in the context. K is the state capacity, set independently of the annotated family count. The main experiments use $K = 6$. The capacity analysis shows the same operating-region pattern across settings: $K = 2$ and $K = 4$ are highly duplicate-stable but give median gains of 0.0068, while $K = 6$ is the smallest tested setting that raises median gain substantially (0.0185) while keeping FR-8 low (0.0120). The $K = 8$ and $K = 12$ results preserve the same qualitative pattern (Appendix Table 7).

4.2 State update

Before evidence step t , LGED reads the previous state with slot attention:

$$\begin{aligned} a_t &= \text{softmax}(q_t K_{t-1}^\top / \sqrt{d}), \\ \bar{s}_t &= a_t S_{t-1}. \end{aligned} \quad (6)$$

The attention a_t produces a soft read over the slots. A role head and an update gate estimate how strongly the incoming evidence should affect the state:

$$S_t = U(S_{t-1}, h_t, a_t, \rho_t, g_t), \quad (7)$$

where ρ_t is the auxiliary role distribution and g_t is the update-route gate. The final verdict head mean-pools the slots and applies a multilayer perceptron (MLP), so repeated evidence affects the verdict through state updates rather than raw sequence length.

4.3 Training objective

The canonical loss combines verdict supervision with auxiliary signals from the intervention construction:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{verdict}} + \lambda_{\text{dup}} \mathcal{L}_{\text{dup}} + \lambda_{\text{role}} \mathcal{L}_{\text{role}} \\ &\quad + \lambda_{\text{gain}} \mathcal{L}_{\text{gain}} + \lambda_{\text{pos}} \mathcal{L}_{\text{pos}} + \lambda_{\text{state}} \mathcal{L}_{\text{state}}. \end{aligned} \quad (8)$$

The duplicate loss discourages large state movement for repeated evidence. The gain and positive-gain losses encourage evidence from a missing family to receive margin gain, and the state loss encourages the missing-family continuation to move the state more than the repeated continuation. Training uses logit-space margins for stability; all reported diagnostics use the probability-margin space of Eq. 4. These auxiliary labels are available only

for training episodes constructed from the training split; at evaluation time, LGED receives only claim and evidence text, and role scores used by the update are model predictions rather than construction labels. Appendix A gives the remaining weights and training details.

5 Experiments

The experiments test the design space outlined in Table 1: whether verification systems separate repeated-evidence stability from missing-family gain, and whether prompted LLMs express the same distinction through a verdict interface. The main setting is HOVER, where supporting-fact annotations directly enable the same-prefix comparison. Additional experiments test semantic repetition in FEVER, secondary same-prefix comparisons in AVERITEC and FEVER, and downstream reader behavior under injected duplicates.

Compared systems. The system suite tests encoder strength, conservative pooling, gain objectives, graph verification, learned state, and prompting behavior. HOVER-BERT and DEBERTA-V3-LARGE represent supervised concatenation verifiers; HOVER-TRAINED KGAT tests in-domain graph verification (Liu et al., 2020). CLUSTER-WEIGHTED, verdict-only and objective-augmented verifiers, SETTRANSFORMER (Lee et al., 2019), CROSSATTENTIONPOOL, and LGED cover pooling, gain objectives, and learned state. We also evaluate zero-shot factuality and NLI adapters, plus 11 local Qwen, Llama, Mistral, and DeepSeek models.

Reporting. All systems receive the same claim-prefix-continuation inputs when the diagnostic applies. Outputs are mapped to the shared probability-margin interface when available; Table 2 also includes a native base score for within-family task context. Cross-system comparisons use the common FR-8, UF Preference, and Median UF Gain definitions on the same diagnostic inputs. Trainable systems are reported as mean $\pm$ standard deviation over seeds, while deterministic and prompted systems use single evaluations. Model identifiers, decoding settings, and full LLM rows appear in Appendix C.

5.1 Main HoVer Diagnostic and Failure Modes

Table 2 shows the central finding: similar clean evidence performance can mask large differences

in marginal aggregation behavior. Across the four supervised verifier rows, clean F1 runs from 0.6324 to 0.6528, while FR-8 runs from 0.0854 to 0.3031, a more than three-fold range within a narrow accuracy band. The modern DEBERTA-V3-LARGE verifier lowers FR-8 relative to HOVER-BERT (0.0854 vs. 0.1902), but it remains over six times as high as LGED’s 0.0137. Its UF gain (0.0165) is close to LGED’s (0.0195), making duplicate stability the main separation between the supervised verifier and the stateful reference.

Conservative aggregation fails differently. CLUSTER-WEIGHTED lowers FR-8 to 0.0450, but UF Pref. falls to 0.5034 and median gain to 0.0010; the zero-shot DEBERTA-NLI mean baseline follows the same pattern. Appendix Table 12 shows the effect on REFUTES cases, where the negative-gain rate is 0.5912 and median gain is -0.0209 .

Gain-oriented and generic stateful systems expose the other side of the trade-off. The objective-augmented verifier obtains larger UF gain than LGED (0.0675 vs. 0.0195), but remains sensitive to duplication (0.2493). CROSSATTENTIONPOOL and SETTRANSFORMER also produce larger gains than conservative baselines, with weaker duplicate stability. The HOVER-TRAINED KGAT graph reference is both duplicate-sensitive and nearly zero in median gain. Figure 2 shows the same pattern: conservative baselines lie in the low-FR-8, near-zero-gain region; gain-oriented systems increase gain but remain sensitive to repetition; and LGED lies in the low-FR-8, positive-gain region. Paired effect-size analysis in Appendix Table 24 confirms the stability advantage and the gain comparison with conservative aggregation.

Depth checks in Appendix Table 25 show the same qualitative pattern across FR-1–FR-8 and FR-AUC. Systems with similar clean accuracy continue to occupy different regions on both axes: conservative systems remain flat, gain-oriented systems remain duplicate-sensitive, and LGED occupies the low-FR-8, positive-gain region.

Median gains in Table 2 are conservative statistics over retained pairs, including cases where both continuations have small effects. The qualitative cases in Appendix E show sharper contrasts: HOVER-BERT can rank the repeated-family continuation above the missing-family one, while LGED reverses that direction.

Family	System	Base score	FR-8 ↓	UF Pref. ↑	Med. UF Gain ↑
Supervised verifier	HoVer-BERT	0.6473 F1	0.1902	0.5932	0.0005
Supervised verifier	DeBERTA-v3-LARGE	0.6324 $\pm$ 0.036 F1	0.0854 $\pm$ 0.017	0.5839 $\pm$ 0.049	0.0165 $\pm$ 0.011
Supervised verifier	Verdict-only verifier	0.6439 $\pm$ 0.039 F1	0.3031 $\pm$ 0.043	0.5556 $\pm$ 0.021	0.0206 $\pm$ 0.006
Gain-oriented	Objective-augmented verifier	0.6528 $\pm$ 0.009 F1	0.2493 $\pm$ 0.033	0.6166 $\pm$ 0.006	0.0675 $\pm$ 0.020
Deduplication	CLUSTER-WEIGHTED	0.5846 $\pm$ 0.001 Acc.	0.0450 $\pm$ 0.004	0.5034 $\pm$ 0.002	0.0010 $\pm$ 0.000
Graph verifier	HoVer-TRAINED KGAT	0.5322 F1	0.3557	0.5017	0.0010
NLI adapter	DeBERTA-NLI mean	0.5867 adapter F1	0.0239	0.5059	0.0001
Stateful pooling	CROSSATTENTIONPOOL	0.6023 Acc.	0.0405	0.5463	0.0730
Stateful pooling	SETTRANSFORMER	0.5963 Acc.	0.0856	0.5361	0.0591
Stateful reference	LGED	0.5882 $\pm$ 0.004 Acc.	0.0137 $\pm$ 0.007	0.5554 $\pm$ 0.015	0.0195 $\pm$ 0.013

Table 2: Main HOVER diagnostic across system families. Base scores use the native metric available for each row: supervised verifier rows report clean macro-F1, diagnostic aggregation rows report base accuracy, and the NLI adapter reports adapter macro-F1. These scores provide within-family task context; the three diagnostic columns provide the common cross-system comparison through shared FR-8, UF Preference, and probability-margin definitions.

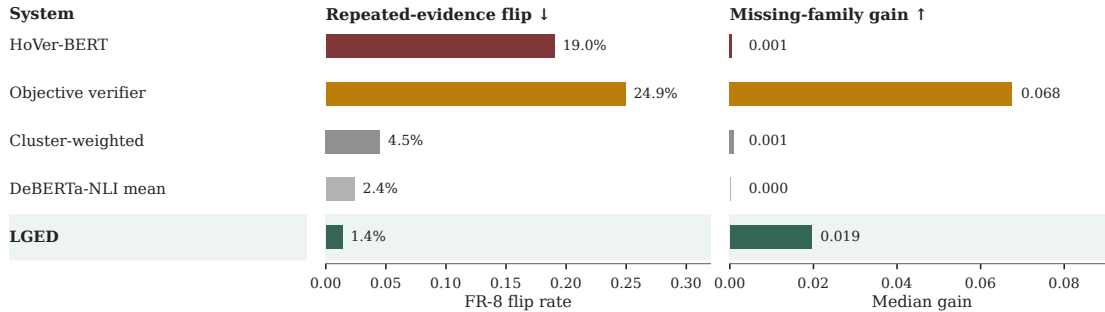


Figure 2: Systems occupy different regions of the trade-off between stability and gain. Conservative aggregation lowers repetition flips but gives little UF gain, whereas gain-oriented systems increase UF gain while remaining more sensitive to repetition. Lower FR-8 is better; higher median gain is better. Bars show mean values; seed standard deviations for stochastic rows are reported in Table 2 and Appendix B.

5.2 Decision Relevance and Surface Robustness

Let $\Delta = G_{uf} - G_{dup}$. We first examine whether the paired gain is large enough to affect final decisions. A *corrective flip* occurs when the repeated-family branch is incorrect and the missing-family branch is correct. For LGED, 48.22% of pairs have $\Delta > 0.03$, 11.44% produce corrective flips compared with 8.81% harmful flips, and 36.36% of near-boundary cases are corrective. Table 3 connects the probability-margin measure to concrete changes in predicted labels and shows why gain is most informative when considered together with duplicate stability.

We next test how the UF signal behaves relative to readily available surface cues. A leakage-controlled surface classifier reaches AU-ROC 0.8876. On the 250 pairs where this classifier selects the wrong branch, all five evaluated systems retain aggregate UF Preference above 0.5, ranging from 0.5840 to 0.6320, with positive median

System	FR-8 ↓	$P(\Delta > .03)$	Corrective	NB corrective
DeBERTA-v3-LARGE	0.0854	0.4475	0.0559	0.1515
Objective-augmented	0.2493	0.5407	0.1898	0.3333
LGED	0.0137	0.4822	0.1144	0.3636

Table 3: Decision-level audit. “NB” denotes near-boundary cases; corrective flips change an incorrect repeated-family decision into a correct missing-family decision.

paired gain (Table 4). Across the full set, surface confidence is non-positively correlated with verifier gain.

The same pattern remains when exact repetition is removed. Replacing the repeated branch with a non-identical passage from the same covered family preserves positive UF for DeBERTa-v3, the objective-augmented verifier, and LGED. We also examine 208 naturally occurring pairs where the missing-family continuation has at least as much prefix overlap as the covered-family continuation. DeBERTa-v3 remains responsive to the missing-

family evidence in this setting (UF 0.6154; median gain 0.0195). Together, these controls show that the family-role signal persists across settings in which prominent surface cues are weakened or point in the opposite direction.

5.3 What Drives the Trade-off

A sweep over LGED objective weights traces the trade-off within the model family. Without duplicate or gain penalties, LGED becomes duplicate-sensitive (FR-8 0.1853); over-regularizing duplication collapses gain to zero. The main LGED configuration in Table 2 ($\lambda_{dup} = 0.01$, $\lambda_{gain} = 0.05$) is the frozen development-selected setting used for the five-seed main comparison, balancing low FR-8 with positive median gain. Appendix Table 10 reports additional exploratory three-seed operating points.

Same-backbone ablations identify which signals support this region (Table 5). Removing the duplicate-stability loss raises FR-8 from 0.0165 to 0.0755. Removing the paired-gain objective removes rank coupling; positive-gain, state-separation, and duplicate-stability losses preserve missing-family pressure and duplicate suppression, so median gain can rise without improving UF preference. Removing the role-diagnosis head drops median gain from 0.0219 to 0.0029, a roughly $7.5\times$ reduction. The model therefore needs to identify whether the incoming passage has a new family role, not only to push the margin upward whenever useful evidence appears. The same distinction appears again in the LLM interface results below under direct role prompts.

The sweep also explains why stability and gain separate. Penalties that discourage state movement stabilize repeated-family evidence but can flatten the response to a missing family; gain rewards move the model in the opposite direction. In these settings, the two behaviors separate most cleanly when the model receives role or paired supervision before updating state. Generic stateful alternatives produce larger median gains (CROSSATTENTIONPOOL 0.0730, SETTRANSFORMER 0.0591), but their duplicate stability is weaker (FR-8 0.0405 and 0.0856).

5.4 LLM Prompting for Evidence Roles

The LLM experiments separate three stages of evidence use: recognizing the evidence role, explicitly identifying which continuation contributes new verdict-relevant information, and reflecting that dis-

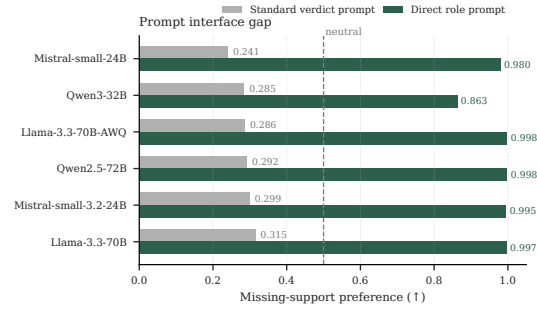


Figure 3: Evidence-role information is much more visible under direct role-oriented prompting than in assembled-context verdict updates. The dashed line marks neutral preference. Appendix C reports all 11 models and the probability-source audit.

inction in the final verdict margin. Verdict scores are computed from requested probability fields when available; Appendix C.2 reports the corresponding probability-source audit. The main analysis uses the Qwen, Llama, and Mistral rows, for which verdict probabilities are consistently available.

Table 6 shows a clear separation across these stages. The representative models identify the evidence roles accurately (0.8797–0.9695) and strongly prefer the missing-family continuation when asked explicitly (0.9644–1.0000). The same distinction appears in the verdict margin for only 0.2466–0.3263 of pairs. The models therefore expose substantially more information about marginal evidence value through direct role-oriented judgments than through the assembled-context verdict interface.

Table 6 reports the dedicated three-stage audit. Appendix Table 16 complements it with the broader prompt suite, including standard verdict, role-aware verdict, pairwise, stepwise, and order-control protocols. These evaluations use their corresponding prompt protocols, so the role and preference columns serve complementary views rather than duplicate measurements. For the main Qwen and Llama rows, order effects are small, while the DeepSeek rows show substantially larger protocol sensitivity. Figure 3 summarizes the same interface gap across the broader prompt comparison.

These results localize the behavioral gap to the transition from explicit role judgments to the final verdict interface.

System	Surface-classifier error subset ($n = 250$)		Non-identical same-family control	
	UF Pref.	Med. Δ	Pairs	UF Pref. / Med. Δ
HoVer-BERT	0.6200	0.0010	—	—
DeBERTA-v3-LARGE	0.6320	0.0268	406	0.5936 / 0.0151
Objective-augmented	0.6320	0.0678	406	0.5665 / 0.0266
CLUSTER-WEIGHTED	0.5840	0.0260	99	0.6162 / 0.0429
LGED	0.6280	0.0636	406	0.5640 / 0.0181

Table 4: Surface controls. Left: pairs where the surface classifier predicts the wrong branch. Right: non-identical same-family evidence replaces exact repetition. All reported median paired gains are positive.

Ablation	FR-8 $\downarrow$	UF Pref. $\uparrow$	Med. UF Gain $\uparrow$
LGED full	0.0165	0.5653	0.0219
w/o duplicate-stability loss	0.0755	0.5593	0.0320
w/o paired-gain objective	0.0204	0.5483	0.0305
w/o role-diagnosis head	0.0199	0.5458	0.0029
flat / pool max	0.0000	0.4534	0.0000

Table 5: LGED ablations on the main HoVer diagnostic. Removing the duplicate-stability loss raises duplicate sensitivity; removing the paired-gain objective removes rank coupling; removing the role-diagnosis head sharply reduces missing-family gain. Ablations use a single controlled comparison setting, so absolute values can differ slightly from the five-seed results in Table 2. Full architecture controls appear in Appendix Table 14.

Model	Role Acc.	Explicit UF Pref.	Verdict Update
QWEN2.5-72B	0.9695	1.0000	0.3263
LLAMA-3.3-70B	0.9178	0.9949	0.2466
MISTRAL-SMALL-24B	0.8797	0.9644	0.2771

Table 6: Dedicated three-stage LLM interface audit. Role Acc. measures recognition of repeated versus missing evidence; Explicit UF Pref. measures direct preference for the continuation contributing new verdict-relevant evidence; Verdict Update measures whether that distinction appears in the gold-label verdict margin.

5.5 Boundary Checks Beyond HoVer

The transfer experiments examine how the two diagnostic components behave beyond the supporting-fact structure available in HoVer. The redundancy ladder transfers directly as a duplicate-sensitivity stress test. On FEVER, we evaluate exact, compressed, paraphrased, and mixed repetitions; conservative clustering and QWEN3-32B remain comparatively stable across these perturbations. These results extend the repeated-evidence effect beyond the main HoVer construction.

We also test whether evidence-family roles can be recovered when gold family annotations are hidden. TF-IDF grouping yields overall role agreement of 0.9085–0.9903 and recovers both branches correctly in 0.8169–0.9805 of cases, showing that

the same-prefix roles can be approximated automatically from passage-level evidence. The AVERITEC and FEVER experiments then provide secondary same-prefix checks under their coarser evidence structure. Appendix D reports the complete transfer results.

5.6 Reader Stress Test

The reader experiment tests whether duplicate evidence affects downstream answer generation when retrieved contexts are already provided. We inject exact, compressed, or paraphrased duplicates into dataset-provided HOTPOTQA (Yang et al., 2018) and 2WIKIMULTIHOPQA (Ho et al., 2020) contexts and evaluate QWEN3-32B and LLAMA-3.3-70B-AWQ readers. Duplicate injection can change answers even when F1 moves little, so flip rate captures sensitivity missed by aggregate accuracy.

Filtering with the trained LGED scorer reduces flip rates across reader and dataset combinations. HOTPOTQA with QWEN3-32B drops from 0.1140 to 0.0720 under exact injection, and 2WIKIMULTIHOPQA with paraphrased duplicates drops from 0.1660 to 0.0780 (Appendix Table 20). The reductions hold with reader LLMs different from the encoder used to train the scorer, showing that the diagnostic-derived redundancy signal transfers to downstream answer stability across reader architectures.

6 Conclusion

More evidence is not always more support. The same-prefix tests introduced here separate stability under repeated-family evidence from gain under missing-family evidence. Systems with similar clean evidence accuracy occupy widely different regions on both diagnostic axes: conservative systems can be stable by flattening marginal gain, while gain-oriented systems can remain sensitive to repeated evidence. LGED provides a compact

stateful reference point with low repeated-evidence sensitivity and positive missing-family gain. The LLM results add a complementary view in which explicit evidence-role recognition is substantially stronger than the corresponding verdict update.

These results support reporting repeated-evidence stability and missing-family gain alongside clean accuracy. Surface-error, semantic same-family, and high-overlap controls show that the missing-family response persists across substantial changes in lexical cues, while the decision-level audit connects paired gain to final-label changes. The same perspective carries to downstream readers and prompted LLMs: evidence redundancy affects answer stability, and explicit evidence-role recognition can be much stronger than the corresponding verdict update. Marginal support contribution is therefore a useful evaluation target in its own right.

7 Limitations

The exact same-prefix construction uses HOVER’s supporting-fact structure. Automatic TF-IDF grouping recovers evidence-family roles with high agreement, and the FEVER and AVERITEC experiments provide evidence under coarser passage-level annotation, while broader validation in weakly annotated and open-web settings remains useful. Surface properties are also predictive of branch identity. The error-subset, semantic same-family, and converse high-overlap analyses cover several important surface alternatives; the converse analysis currently uses naturally occurring relative-overlap pairs and one verifier. A deliberately constructed high-overlap set would extend this stress test further.

LGED is designed as a compact reference model for the diagnostic. Its strongest empirical result is duplicate stability, while missing-family gain is comparable to several supervised alternatives rather than uniformly larger. The LLM analysis is behavioral and localizes the gap between explicit evidence-role judgments and verdict updating. Internal representation analysis would provide an additional view of this transition. Finally, the main diagnostic uses English Wikipedia-based fact-verification data, leaving multilingual, multimodal, and end-to-end retrieval settings for broader evaluation.

Ethical considerations

We use public fact-verification and QA datasets (HOVER, FEVER, AVERITEC, HOTPOTQA, and 2WIKIMULTIHOPQA) through released research splits and do not collect new human annotations, run crowdworker studies, or conduct manual human evaluation. The interventions could be misused to probe deployed verifiers; we present them as evaluation tools for robustness analysis, not as guidance for evading moderation or fact-checking systems. All external datasets and models are used through their released research versions and associated access terms; any released diagnostic artifacts from this work are intended for research evaluation and robustness analysis only.

References

- Guy Alt, Eran Hirsch, Serwar Basch, Ido Dagan, and Oren Glickman. 2026. [User-centric evidence ranking for attribution and fact verification](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7215–7237.
- Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. [FEVEROUS: Fact extraction and VERification over unstructured and structured information](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. [Self-RAG: Learning to retrieve, generate, and critique through self-reflection](#). In *International Conference on Learning Representations*.
- Sumanth Doddapaneni, Mohammed Safi Ur Rahman Khan, Sshubam Verma, and Mitesh M. Khapra. 2024. [Finding blind spots in evaluator LLMs with interpretable checklists](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16279–16309.
- Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. [FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952.
- Muhan Gao, TaiMing Lu, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024. [Insights into LLM long-context failures: When transformers know but don’t tell](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7611–7625.

- Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. [Evaluating models’ local decision boundaries via contrast sets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *International Conference on Learning Representations*.
- Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. [Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.
- Xuming Hu, Zhaochen Hong, Zhijiang Guo, Lijie Wen, and Philip S. Yu. 2023. [Read it twice: Towards faithfully interpretable fact verification by revisiting evidence](#). In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2319–2323. Association for Computing Machinery.
- Yebowen Hu, Kaiqiang Song, Sangwoo Cho, Xiaoyang Wang, Wenlin Yao, Hassan Foroosh, Dong Yu, and Fei Liu. 2024. [When reasoning meets information aggregation: A case study with sports narratives](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4293–4308.
- Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. [HoVer: A dataset for many-hop fact extraction and claim verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460.
- Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. 2020. [Learning the difference that makes a difference with counterfactually-augmented data](#). In *International Conference on Learning Representations*.
- Yuqing Lan, Zhenghao Liu, Yu Gu, Xiaoyuan Yi, Xiaohua Li, Liner Yang, and Ge Yu. 2025. [Multi-evidence based fact verification via a confidential graph neural network](#). *IEEE Transactions on Big Data*, 11(2):426–437.
- Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosior, Seungjin Choi, and Yee Whye Teh. 2019. [Set transformer: A framework for attention-based permutation-invariant neural networks](#). In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Yuxuan Liu, Hongda Sun, Wenya Guo, Xinyan Xiao, Cunli Mao, Zhengtao Yu, and Rui Yan. 2025. [BiDeV: Bilateral defusing verification for complex claim fact-checking](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 541–549.
- Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. [Fine-grained fact verification with kernel graph attention network](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. [RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878.
- Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. [Beyond accuracy: Behavioral testing of NLP models with CheckList](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912.
- Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. [AVeriTeC: A dataset for real-world claim verification with evidence from the web](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. [Get your vitamin C! robust fact verification with contrastive evidence](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643.
- Marek Strong and Andreas Vlachos. 2025. [TSVer: A benchmark for fact verification against time-series evidence](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29906–29926.
- Liyan Tang, Philippe Laban, and Greg Durrett. 2024. [MiniCheck: Efficient fact-checking of LLMs on grounding documents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847.

- James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. [FEVER: A large-scale dataset for fact extraction and VERification](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A dataset for diverse, explainable multi-hop question answering](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. [Making retrieval-augmented language models robust to irrelevant context](#). In *International Conference on Learning Representations*.
- Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. [From relevance to utility: Evidence retrieval with feedback for fact verification](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6373–6384.

A Implementation and Protocol Details

This appendix collects implementation details, extended tables, qualitative examples, and robustness analyses for the main diagnostic.

Canonical scoring. All evaluations use a shared binary probability margin adapter. Equation 4 is the scoring interface used by all main diagnostics. The adapter provides a bounded common scale for binary and collapsed three-way verifiers. Binary models use their two output probabilities directly. Three-way verifiers map REFUTES and NEI into NOT-SUPPORTS. LLM analyses use prompt behavior for preference and role metrics; when probability margins are unavailable, those rows are omitted from margin-gain comparisons rather than mixed with classifier probabilities.

LGED details. The canonical model uses $K = 6$ slots, hidden dimension $d = 128$, and input feature dimension 779. Each input vector is $x_t = [\text{CLS}_t; \ell_t; p_t; m_t]$, where $\text{CLS}_t \in \mathbb{R}^{768}$, $\ell_t \in \mathbb{R}^3$, $p_t \in \mathbb{R}^3$, and $m_t \in \mathbb{R}^5$. The auxiliary role labels are base, duplicate, independent, near-miss, and conflict. The update gates are merge, create, conflict, and suppress. The active objective uses $\lambda_{dup} = 0.01$, $\lambda_{role} = 0.10$, $\lambda_{gain} = 0.05$, $\lambda_{pos} = 0.02$, and $\lambda_{state} = 0.02$, with $\gamma_{gain} = 0.05$, $\gamma_{pos} = 0.01$, and $\gamma_{state} = 0.02$. Only λ_{dup} and λ_{gain} vary in the sweep reported in Section 5.3; the other auxiliary weights are fixed small values relative to the verdict loss. The main setting uses development data to minimize FR-8 subject to positive median missing-family gain. Training uses AdamW for 10 epochs with learning rate 3×10^{-4} , batch size 16, and weight decay 0.01. We use $K = 6$ as the compact capacity setting for the main row; Table 7 reports the slot-capacity sensitivity analysis. The diagnosis logits are supervised as an auxiliary signal, while the soft role scores support the update mechanism at test time.

Detailed LGED equations. The equations below expand Eqs. 6 through 8 and define the update used by Algorithm 1.

$$\begin{aligned} q_t &= W_q h_t + b_q, \\ k_{t,j} &= W_k s_{t-1,j} + b_k, \\ a_t &= \text{softmax}_j(q_t^\top k_{t,j} / \sqrt{d}), \\ \bar{s}_t &= \sum_j a_{t,j} s_{t-1,j}. \end{aligned} \quad (9)$$

Let $u_t = [h_t; \bar{s}_t]$. The role and update-route distributions are

$$\begin{aligned} \rho_t &= \text{softmax}((W_{role} u_t + b_{role}) / \tau_{role}), \\ r_t &= W_{g1} [u_t; \rho_t] + b_{g1}, \\ g_t &= \text{softmax}(W_{g2} \text{ReLU}(r_t) + b_{g2}). \end{aligned} \quad (10)$$

Create and conflict routes use learned slot selectors:

$$\begin{aligned} c_t &= \text{softmax}(W_c h_t + b_c), \\ f_t &= \text{softmax}(W_f h_t + b_f), \\ \beta_t &= g_{t,merge} a_t + g_{t,create} c_t + g_{t,conflict} f_t. \end{aligned} \quad (11)$$

The shared update vector is $u_t^{upd} = \tanh(W_u u_t + b_u)$. Each slot is updated by

$$s_{t,j} = \text{LN}(s_{t-1,j} + \beta_{t,j} u_t^{upd}), \quad (12)$$

with the previous slot retained unchanged when $|\beta_{t,j}| < \epsilon$, where ϵ is fixed across runs. Suppress is not subtractive; a high suppress gate reduces the mass assigned to the other update routes.

For duplicate-role steps, let $\Delta_t = \sum_j \|s_{t,j} - s_{t-1,j}\|_2$. Then

$$\begin{aligned} \mathcal{L}_{dup} &= \mathbb{E}_{t:role_t=dup} \Delta_t, \\ \mathcal{L}_{role} &= \text{CE}(W_{role} u_t + b_{role}, role_t). \end{aligned} \quad (13)$$

For a same-prefix pair (B, d, u) , define the logit margin

$$m(z, y) = z_y - \log \sum_{c \neq y} \exp z_c. \quad (14)$$

The paired gains are

$$\begin{aligned} gain_d &= m(z_{B+d}, y) - m(z_B, y), \\ gain_u &= m(z_{B+u}, y) - m(z_B, y). \end{aligned} \quad (15)$$

The gain losses are

$$\begin{aligned} \mathcal{L}_{gain} &= \mathbb{E} \text{ReLU}(\gamma_{gain} - gain_u + gain_d), \\ \mathcal{L}_{pos} &= \mathbb{E} \text{ReLU}(\gamma_{pos} - gain_u), \\ \mathcal{L}_{state} &= \mathbb{E} \text{ReLU}(\gamma_{state} - \Delta_u + \Delta_d). \end{aligned} \quad (16)$$

Algorithm 1 gives the step-level update used in implementation.

Cluster-weighted baseline. CLUSTER-WEIGHTED groups evidence steps using encoder-space cosine similarity, pools each cluster before verdict scoring, and records cluster size as a feature. Thresholds are selected on a development split and held fixed for evaluation. Table 12 tests whether this conservative pooling also suppresses missing-family gain.

Algorithm 1 LGED slot update for one evidence step

Require: evidence feature x_t , previous slots $S_{t-1} \in \mathbb{R}^{K \times d}$

- 1: $h_t \leftarrow \text{LayerNorm}(\text{ReLU}(W_{enc}x_t + b_{enc}))$
- 2: $q_t \leftarrow W_q h_t + b_q$; $k_{t,j} \leftarrow W_k s_{t-1,j} + b_k$
- 3: $a_t \leftarrow \text{softmax}_j(q_t^\top k_{t,j} / \sqrt{d})$
- 4: $\bar{s}_t \leftarrow \sum_j a_{t,j} s_{t-1,j}$; $u_t \leftarrow [h_t; \bar{s}_t]$
- 5: $\rho_t \leftarrow \text{softmax}((W_{role}u_t + b_{role}) / \tau_{role})$
- 6: $g_t \leftarrow \text{Gate}(u_t, \rho_t)$
- 7: $c_t \leftarrow \text{softmax}(W_c h_t + b_c)$; $f_t \leftarrow \text{softmax}(W_f h_t + b_f)$
- 8: $\beta_t \leftarrow \text{Mix}(a_t, c_t, f_t; g_t)$
- 9: $u_t^{upd} \leftarrow \tanh(W_u u_t + b_u)$
- 10: **for** $j = 1$ to K **do**
- 11: **if** $|\beta_{t,j}| > \epsilon$ **then**
- 12: $s_{t,j} \leftarrow \text{LayerNorm}(s_{t-1,j} + \beta_{t,j} u_t^{upd})$
- 13: **else**
- 14: $s_{t,j} \leftarrow s_{t-1,j}$
- 15: **end if**
- 16: **end for**
- 17: **return** $S_t = [s_{t,1}, \dots, s_{t,K}]$

LLM prompting and verdict UF metric. The LLM suite uses prompts for standard verdict, role-aware verdict, pairwise preference, role classification, and stepwise update. All LLM evaluations use temperature 0 and each model’s prompt format. Verdict prompts request JSON output containing label, prob_support, and prob_refute. When both probability fields are parsed, we normalize them and compute a binary gold-label probability margin; when only a valid label is recovered, a deterministic one-hot fallback assigns 1.0 to the parsed label and 0.0 to the complement. Verdict UF is therefore a prompted verdict-margin diagnostic rather than a calibrated probability estimate. Pairwise UF is computed separately from a parsed preferred_candidate field and does not rely on self-reported probabilities. Appendix C.2 reports the verdict probability-source audit used to separate probability-field rows from label-fallback rows.

Reader filtering procedure. The reader analysis starts from dataset-provided contexts, then injects exact, compressed, or paraphrased duplicates. The trained LGED scorer is applied to question-passage pairs without updating the reader model. In the duplicate-injected condition, the filter ranks injected candidates by redundancy score and removes the highest-redundancy candidates while retaining the original context. Gold answers are used only for evaluation.

Computational budget and implementation record. All trainable diagnostic models were run on NVIDIA A100-SXM4-80GB GPUs under Python 3.10.19, using PyTorch 2.11.0 with CUDA

13.0, Hugging Face Transformers 4.57.6, NumPy 2.2.6, SciPy 1.15.3, and scikit-learn 1.7.2. The main LGED experiments used the compact setting described above with $K = 6$ slots and hidden dimension $d = 128$, trained for 10 epochs with AdamW, learning rate 3×10^{-4} , batch size 16, and weight decay 0.01. Hyperparameters were selected on development data to minimize FR-8 subject to positive median missing-family gain. Prompted LLM evaluations used temperature 0 and fixed prompt templates. We report stochastic trainable systems as mean $\pm$ standard deviation over seeds and deterministic or prompted systems as single evaluations. The total compute budget was approximately 200 A100-GPU hours, including model training, diagnostic evaluation, prompted LLM evaluation, and reader stress tests. Model scale is identified through the public model names used in the experimental tables, including DeBERTa-v3 encoders and local LLMs ranging from 8B to 72B parameters.

B Extended HoVer Results and Mechanism Checks

This section expands the main HoVer diagnostic along three axes: capacity, comparator coverage, and mechanism checks. Figure 4 and Tables 11 and 25 examine whether repeated evidence mainly affects confidence and repetition depth; Tables 8, 9, 13, and 14 place the main systems beside additional controls. Bold values mark entries emphasized in the main text.

B.1 Slot-capacity sensitivity analysis

The number of slots K controls the capacity of the learned evidence state. K is set independently of the gold evidence-family count, and slots are learned without family-identifier supervision. Table 7 varies K while keeping the rest of the training protocol fixed. Very small capacities preserve duplicate stability but give weak missing-family gain. $K = 6$ is the smallest tested setting that substantially increases median missing-family gain while keeping repeated-evidence flips low, so we use it as the canonical compact setting in the main experiments.

The extended diagnostic table keeps the base-score convention visible for every row and adds the NLI pooling variants. Max pooling is the most stable of these references and has near-zero missing-family gain, preserving the qualitative pattern in

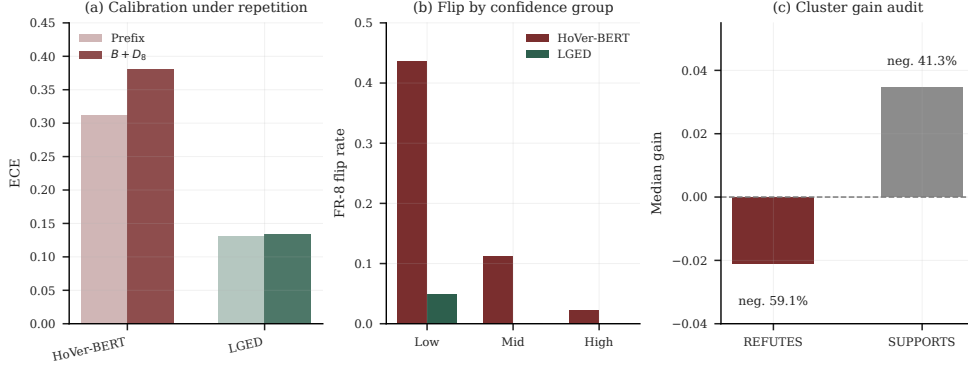


Figure 4: Mechanism analyses for repeated-evidence sensitivity. Repetition worsens HOVER-BERT calibration as measured by Expected Calibration Error (ECE) and flips low-confidence examples; LGED has lower flip rates under the same conditions.

Slots K	Base Acc.	FR-8 ↓	UF Pref. ↑	Med. Gain ↑
2	0.5850 $\pm$ 0.0020	0.0052 $\pm$ 0.0040	0.5356 $\pm$ 0.0061	0.0068 $\pm$ 0.0008
4	0.5797 $\pm$ 0.0046	0.0065 $\pm$ 0.0059	0.5512 $\pm$ 0.0121	0.0068 $\pm$ 0.0039
6	0.5868 $\pm$ 0.0071	0.0120 $\pm$ 0.0035	0.5515 $\pm$ 0.0199	0.0185 $\pm$ 0.0138
8	0.5881 $\pm$ 0.0043	0.0122 $\pm$ 0.0056	0.5458 $\pm$ 0.0167	0.0130 $\pm$ 0.0083
12	0.5827 $\pm$ 0.0041	0.0093 $\pm$ 0.0057	0.5695 $\pm$ 0.0097	0.0164 $\pm$ 0.0055

Table 7: Slot-capacity sensitivity analysis over five seeds. K is the number of learned latent state slots and is set independently of the annotated evidence-family count. Small capacities ($K = 2, 4$) are duplicate-stable but give weak missing-family gain. $K = 6$ is the compact default because it first raises median gain substantially while keeping FR-8 low; larger capacities preserve the same qualitative operating-region pattern.

Table 2.

The modern-verifier addendum isolates encoder scaling within the standard concatenation-verifier family. The DeBERTa-v3-large row improves duplicate stability relative to HoVer-BERT and remains above the stateful reference in repeated-evidence flip rate.

The objective sweep places the reported LGED setting within the broader stability–gain trade-off. Some settings reduce repetition flips further, while higher seed variance or collapsed missing-family gain makes their operating behavior less balanced.

The calibration check asks where repeated evidence changes decisions. The contrast is concentrated in low-confidence examples for HoVer-BERT, whereas LGED has lower calibration error and fewer flips across confidence bands.

The cluster-weighted baseline provides a useful conservative point, so we also test where its stability comes from. The split by gold label shows that conservative pooling can suppress missing-family gain, especially in REFUTES examples.

MiniCheck and FactKB adapters (Tang et al., 2024; Feng et al., 2023) plus NLI adapters provide conservative zero-shot references. Their near-zero

gains reinforce the distinction between duplicate stability and missing-family response.

The final table in this section combines loss ablations with architecture controls. The ablation rows indicate which LGED components support the main operating region, while the generic pooling rows show that larger gain can come with weaker duplicate stability.

Model	Base score	FR-8 ↓	UF Pref. ↑	Med. UF Gain ↑
HoVER-BERT	0.6473 F1	0.1902	0.5932	0.0005
Verdict-only verifier	0.6439 ±0.0388 F1	0.3031 ±0.0431	0.5556 ±0.0208	0.0206 ±0.0059
Objective-augmented verifier	0.6528 ±0.0086 F1	0.2493 ±0.0333	0.6166 ±0.0061	0.0675 ±0.0203
CLUSTER-WEIGHTED	0.5846 ±0.0012 Acc.	0.0450 ±0.0035	0.5034 ±0.0017	0.0010 ±0.0004
LGED	0.5882 ±0.0044 Acc.	0.0137 ±0.0073	0.5554 ±0.0150	0.0195 ±0.0129
HoVER-TRAINED KGAT	0.5322 F1	0.3557	0.5017	0.0010
DeBERTA-NLI mean	0.5867 adapter F1	0.0239	0.5059	0.0001
DeBERTA-NLI max	0.5873 adapter F1	0.0000	0.4653	0.0000
DeBERTA-NLI sum	0.5867 adapter F1	0.0239	0.5059	0.0003

Table 8: Full HoVER diagnostic table. Base scores use the native available interface for each row and provide context rather than a single cross-system ranking: supervised verifier rows report clean macro-F1, diagnostic aggregation rows report base accuracy, and NLI rows report adapter macro-F1. The diagnostic columns use the shared probability-margin interface.

Model	Encoder	Seeds	Dev macro-F1	FR-8 ↓	Med. Gain ↑
DeBERTA-v3-LARGE verifier	microsoft/deberta-v3-large	3	0.6324 ±0.0362	0.0854 ±0.0169	0.0165 ±0.0114
RoBERTa-LARGE verifier	roberta-large	1	0.5150	0.1227	0.0009

Table 9: Modern supervised verifier addendum. DeBERTa-v3-large reduces duplicate sensitivity relative to HoVER-BERT while remaining more duplicate-sensitive than LGED; RoBERTa-large provides an additional scaling reference.

Point	λ_{dup}	λ_{gain}	Base Acc.	FR-8 ↓	UF Pref. ↑	Med. UF Gain ↑
Main	0.01	0.05	0.5882 ±0.0044	0.0137 ±0.0073	0.5554 ±0.0150	0.0195 ±0.0129
No penalties	0.0	0.0	0.5826 ±0.0051	0.1853 ±0.1254	0.5534 ±0.0088	0.0398 ±0.0209
Low dup+gain	0.05	0.05	0.5832 ±0.0033	0.0053 ±0.0023	0.5638 ±0.0063	0.1386 ±0.0696
Balanced point	0.3	0.3	0.5896 ±0.0012	0.0002 ±0.0003	0.5653 ±0.0060	0.0304 ±0.0069
High-dup gain	1.0	0.05	0.5877 ±0.0021	0.0006 ±0.0000	0.5621 ±0.0092	0.1443 ±0.0389
Over-regularized	10.0	0.05	0.5881 ±0.0005	0.0000 ±0.0000	0.5201 ±0.0020	0.0000 ±0.0000

Table 10: Representative points from the within-LGED sweep over stability and gain. The first row is the frozen main LGED configuration selected on development data and evaluated over five seeds; the remaining rows are exploratory three-seed operating points.

Model	ECE prefix	ECE after D_8	Low-conf. flip	Mid-conf. flip	High-conf. flip
HoVER-BERT	0.3119	0.3814	0.4361	0.1124	0.0221
LGED	0.1310	0.1343	0.0494	0.0000	0.0000

Table 11: Expected Calibration Error (ECE) and confidence-stratified repeated-evidence flips. This table provides the calibration and confidence evidence used in Section 5.1.

Subset	Negative-gain rate	Median gain	Interpretation
SUPPORTS	0.4131	0.0347	missing-family evidence usually helps
REFUTES	0.5912	-0.0209	missing-family evidence often under-rewarded

Table 12: Cluster-weighted negative-gain analysis. Conservative pooling is stable under duplication, but the values show where it can suppress gain from missing-family evidence.

Model	Family	Diagnostic base macro-F1	FR-8 ↓	UF Pref. ↑	Med. Gain ↑
MINICHECK-DeBERTA-v3-LARGE	factuality	0.3635	0.0153	0.5136	0.0003
FACTKB	factuality	0.4070	0.4708	0.5237	0.0009
DeBERTA-NLI mean	NLI	0.5867	0.0239	0.5059	0.0001
DeBERTA-NLI max	NLI	0.5873	0.0000	0.4653	0.0000
DeBERTA-NLI sum	NLI	0.5867	0.0239	0.5059	0.0003

Table 13: Expert factuality and NLI baselines. Diagnostic base macro-F1 is computed through the probability-margin adapter, while supervised HoVer verifier clean F1 uses the native task metric. These systems provide zero-shot aggregation references.

Model / ablation	FR-8 ↓	UF Pref. ↑	Med. UF Gain ↑	Interpretation
LGED full (ablation setting)	0.0165	0.5653	0.0219	full model; base Acc. 0.5906
w/o duplicate-stability loss	0.0755	0.5593	0.0320	duplicate stability weakens
w/o paired-gain objective	0.0204	0.5483	0.0305	gain not sole driver
w/o role-diagnosis head	0.0199	0.5458	0.0029	role head supports gain
flat / pool max	0.0000	0.4534	0.0000	stable with near-zero gain
Objective-augmented verifier	0.2215	0.6127	0.0585	higher gain with duplicate sensitivity
CROSSATTENTIONPOOL	0.0405	0.5463	0.0730	higher gain with weaker stability; base Acc. 0.6023
SETTRANSFORMER	0.0856	0.5361	0.0591	higher gain with weaker stability; base Acc. 0.5963
HOVER-TRAINED KGAT	0.3557	0.5017	0.0010	in-domain graph training leaves gain near zero
DEBERTA-NLI mean	0.0239	0.5059	0.0001	zero-shot factuality baseline

Table 14: Extended LGED ablations and architecture controls. This table extends Table 5 with full architecture controls. Ablations use a single controlled comparison setting, so absolute values can differ slightly from the five-seed main results in Table 2; comparisons should be read within this table.

C LLM Prompting Results

This appendix reports the LLM prompt and parsing controls, the verdict probability-source audit, the full local LLM table, and a few-shot role-prompt check.

C.1 Prompt and parsing controls

All LLM prompt templates serialize the same claim, prefix, and two candidate continuations. No prompt exposes evidence-family names, gold labels, construction identifiers, or diagnostic metric definitions. The standard verdict prompt asks for a fact-checking label and two probability fields; the pairwise prompt asks which candidate contributes more new verdict-relevant evidence under the shared prefix; and the role prompt asks whether each candidate is repeated-family or missing-family evidence. The dedicated audit in Table 6 evaluates these three stages separately—role recognition, explicit candidate preference, and verdict update—while Table 16 reports the broader prompt suite and its corresponding protocol-specific metrics. JSON outputs are parsed by exact field names. Deterministic fallbacks are used only when a valid label or preferred-candidate field is recoverable, and malformed or incomplete probability fields are not mixed with classifier probability margins.

C.2 Verdict probability-source audit

Verdict prompts request JSON fields `label`, `prob_support`, and `prob_refute`. When both probability fields are present and valid, the evaluator normalizes them and computes the gold-label probability margin. When a model returns a valid verdict label without valid probabilities, the evaluator uses a deterministic one-hot fallback. Invalid or unparsable outputs are excluded from the corresponding verdict-margin calculation. Pairwise UF and role accuracy are computed separately from parsed preference or role fields and do not rely on verdict-probability fields.

Across the two verdict-prompt protocols, covering 22 model-protocol rows over 11 local LLMs, we obtain 290,958 valid verdict predictions. Of these, 264,740 contain parsed probability fields and 26,218 use label-only fallback; 14,942 additional predictions are invalid or parsing errors. Table 15 reports the audit by model family and verdict-prompt protocol. The Qwen, Llama, and Mistral verdict rows are probability-field based and form the main verdict-interface analysis. The DeepSeek-R1 distilled rows rely primarily on label fallback and are retained as protocol-stress cases. Table 15 makes this distinction explicit across model families and protocols.

C.3 Full local LLM results

Section 5.4 reports a dedicated three-stage audit, while Table 16 gives the broader prompt suite under the corresponding standard-verdict, role-aware, pairwise, stepwise, and order-control protocols. The main text highlights representative Qwen, Llama, and Mistral rows, whose verdict UF values are probability-field based according to Table 15. DeepSeek rows provide an additional protocol-stress case because their verdict outputs rely much more heavily on label fallback. A value of 0.0000 is an observed score. Order sensitivity is the absolute change under the order-control condition, so lower values are better.

C.4 Few-shot role-prompt check

Table 17 reports a few-shot role-prompt check for representative local LLMs.

The few-shot role-prompt check tests whether the missing-family role is accessible when the prompt directly asks for it. These rows support the interpretation that role recognition and verdict updating are different interfaces.

Model family	Protocol	Model rows	Valid rate	Parsed-prob share	Label-fallback share	Invalid rate
DeepSeek-R1 distilled	Vanilla verdict	2	0.2046 to 0.8549	0.0000 to 0.0020	0.9980 to 1.0000	0.1451 to 0.7954
DeepSeek-R1 distilled	Role-aware verdict	2	0.8531 to 0.9585	0.0027 to 0.0058	0.9942 to 0.9973	0.0415 to 0.1469
Llama	Vanilla verdict	2	1.0000	1.0000	0.0000	0.0000
Llama	Role-aware verdict	2	1.0000	1.0000	0.0000	0.0000
Mistral	Vanilla verdict	2	1.0000	1.0000	0.0000	0.0000
Mistral	Role-aware verdict	2	1.0000	1.0000	0.0000	0.0000
Qwen	Vanilla verdict	5	1.0000	1.0000	0.0000	0.0000
Qwen	Role-aware verdict	5	1.0000	1.0000	0.0000	0.0000

Table 15: Verdict probability-source audit for local LLM verdict prompts. The table reports the 22 model-protocol rows covering 11 local LLMs. Ranges give model-level values within each family and protocol. Parsed-prob and label-fallback shares are computed among valid predictions; invalid rate is computed over all predictions. Qwen, Llama, and Mistral return valid probability fields for all verdict-prompt predictions and provide the main verdict-interface rows. DeepSeek-R1 distilled rows often omit valid probability fields or produce invalid outputs and are reported as protocol-stress cases based largely on label fallback.

Family	Model	FR-8 ↓	Standard UF	Role-aware UF	Role acc.	Pairwise UF	Stepwise UF	Order sens. ↓
DeepSeek	R1-DISTILL-LLAMA-70B	0.0000	0.0566	0.1241	0.9774	0.2297	0.8516	0.7262
DeepSeek	R1-DISTILL-QWEN-32B	0.0000	0.0000	0.1255	0.7997	0.2667	0.8564	0.7003
Llama	LLAMA-3.3-70B	0.2589	0.3153	0.2466	0.9589	0.9966	0.9949	0.0025
Llama	LLAMA-3.3-70B-AWQ	0.2447	0.2864	0.2339	0.9648	0.9983	0.9941	0.0153
Mistral	MISTRAL-SMALL-24B	0.3282	0.2407	0.2771	0.9398	0.9797	0.9644	0.0619
Mistral	MISTRAL-SMALL-3.2-24B	0.2158	0.2992	0.1703	0.9233	0.9949	0.9737	0.0475
Qwen	QWEN2.5-72B	0.0829	0.2924	0.3263	0.9847	0.9983	1.0000	0.0093
Qwen	QWEN3-14B	0.1692	0.3500	0.3483	0.9538	0.9568	0.9644	0.0254
Qwen	QWEN3-30B-A3B	0.1034	0.2271	0.2669	0.9208	0.9424	0.8712	0.0551
Qwen	QWEN3-32B	0.2135	0.2847	0.3831	0.9097	0.8627	0.9229	0.0110
Qwen	QWEN3-8B	0.1777	0.2847	0.3271	0.8992	0.9983	0.9941	0.0992

Table 16: Local LLM evidence-role prompting results. Standard verdict prompting stays below neutral preference for the comparable Qwen, Llama, and Mistral rows, while direct pairwise prompting is high for those families. The DeepSeek rows provide a contrasting protocol-stress case with high order sensitivity, weaker pairwise preference, and extensive label fallback according to Table 15.

Model	Pairs	Role macro acc.	Both roles correct	UF pref. when present
QWEN2.5-72B	1180	0.9381	0.8763	0.8797
LLAMA-3.3-70B	1180	0.9169	0.8398	0.8439
MISTRAL-SMALL-24B	1180	0.8367	0.6735	0.6735

Table 17: Few-shot role-prompt evaluation for representative local LLMs. The results show that explicit role examples improve role identification, reinforcing that the main gap lies between role recognition and verdict-margin updating.

D Transfer and Reader Results

The boundary and reader analyses are separated by diagnostic role. Semantic redundancy tests whether systems are stable under repeated or paraphrased evidence without requiring family structure. Secondary same-prefix comparisons test how far the family-role idea travels beyond HoVer. Reader experiments test whether a redundancy signal learned from the diagnostic can reduce answer flips in downstream QA contexts.

The first transfer table isolates semantic redundancy independently of evidence-family roles and serves as a duplicate-sensitivity stress test.

The next table applies secondary same-prefix comparisons outside HoVer. Under the coarser evidence structure available in these datasets, the family-role signal is weaker and more model-dependent.

The reader table asks whether the diagnostic-derived redundancy signal helps after retrieval, when a separate LLM reader consumes the context. The consistent flip-rate reductions show practical value even when answer F1 changes are small.

Model	Base Acc.	Exact FR-8	Compression FR-8	Paraphrase FR-8	Mixed FR-8
CLUSTER-WEIGHTED	0.8603	0.0000	0.0098	0.0515	0.0081
LGED	0.7721	0.0535	0.0980	0.0588	0.0589
QWEN3-32B	0.9706	0.0134	0.0098	0.0074	0.0122
Verdict-only verifier	0.9412	0.1245	0.0588	0.0515	0.1076
Objective-augmented verifier	0.9485	0.1165	0.0588	0.0809	0.1056

Table 18: FEVER semantic-redundancy variants. These rows evaluate general duplicate, compression, paraphrase, and mixed stability; UF metrics do not apply to this perturbation.

Dataset	Model	FR-8 ↓	UF Pref. ↑	Med. Gain ↑
AVERITEC	CLUSTER-WEIGHTED	0.0000	0.5969	0.0955
AVERITEC	LGED	0.0560	0.4923	−0.0049
AVERITEC	Objective-augmented verifier	0.3280	0.5046	0.0286
FEVER full dev	CLUSTER-WEIGHTED	0.0000	0.4728	−0.0027
FEVER full dev	LGED	0.0573	0.4853	−0.0003
FEVER full dev	Objective-augmented verifier	0.1103	0.8635	0.0857

Table 19: Secondary same-prefix family comparisons outside HOVER. The table shows where the HoVer-centered family-structure finding weakens.

Dataset	Reader	Injection	Dup. ΔF1	Dup. Flip	LGED ΔF1	LGED Flip
HOTPOTQA	QWEN3-32B	compression	−0.0066	0.1080	−0.0002	0.0720
HOTPOTQA	QWEN3-32B	exact	−0.0065	0.1140	0.0003	0.0720
HOTPOTQA	QWEN3-32B	paraphrase	−0.0109	0.1240	−0.0002	0.0740
2WIKIMULTIHOPQA	QWEN3-32B	compression	0.0100	0.1120	0.0143	0.0820
2WIKIMULTIHOPQA	QWEN3-32B	exact	0.0176	0.1300	0.0137	0.0780
2WIKIMULTIHOPQA	QWEN3-32B	paraphrase	−0.0139	0.1660	0.0146	0.0780
HOTPOTQA	LLAMA-3.3-70B-AWQ	compression	−0.0105	0.0660	−0.0081	0.0420
HOTPOTQA	LLAMA-3.3-70B-AWQ	exact	−0.0104	0.0800	−0.0054	0.0400
HOTPOTQA	LLAMA-3.3-70B-AWQ	paraphrase	−0.0067	0.0920	−0.0063	0.0400
2WIKIMULTIHOPQA	LLAMA-3.3-70B-AWQ	compression	0.0015	0.0780	−0.0026	0.0480
2WIKIMULTIHOPQA	LLAMA-3.3-70B-AWQ	exact	−0.0143	0.0920	−0.0011	0.0500
2WIKIMULTIHOPQA	LLAMA-3.3-70B-AWQ	paraphrase	−0.0093	0.1100	−0.0006	0.0500

Table 20: Full retrieval-context reader results for duplicate-injected contexts and the trained LGED filter. Exact, compressed, and paraphrased injections affect answer stability; LGED filtering generally lowers flip rates even when F1 changes vary by reader and dataset.

E Qualitative Examples

Table 21 and Table 22 show illustrative cases from the HOVER same-prefix diagnostic. The examples make the paired construction concrete; the aggregate claims rely on the main paired results. In each case, the repeated-family continuation restates evidence already present in the prefix, while the missing-family continuation supplies a held-out gold family. The reported gains show how a small median effect can still contain decisive pair-level reversals.

Claim	Peter Julian Curtin appeared in <i>The Plough and the Stars</i> with an Australian actress whose biggest role was in the 1982 film <i>Looking for Alibrandi</i> .
Prefix evidence	Peter Julian Curtin was an Australian television and stage actor whose career began when he joined the Melbourne Theatre Company, appearing in <i>The Plough and the Stars</i> with Wendy Hughes.
Repeated-family continuation	Peter Julian Curtin was an Australian television and stage actor whose career began when he joined the Melbourne Theatre Company, appearing in <i>The Plough and the Stars</i> with Wendy Hughes.
Missing-family continuation	A separate evidence sentence states that Wendy Hughes’s biggest role was in <i>Lonely Hearts</i> , played in 1982, which began a long collaboration with director Paul Cox.

Model	G_d	G_u	$G_u - G_d$
HOVER-BERT	+0.0805	-1.8790	-1.9595
LGED	-0.0042	+0.9557	+0.9599

Table 21: Qualitative REFUTES example. The repeated-family continuation restates the Peter Curtin evidence in the prefix; the missing-family continuation supplies the Wendy Hughes-side evidence.

Claim	Eystein Orre died in 1066 at a battle in the East Riding of Yorkshire. It is where Hastings defeated King Harald of Norway.
Prefix evidence	The Battle of Stamford Bridge took place at the village of Stamford Bridge, East Riding of Yorkshire, on 25 September 1066, between an English army under King Harold Godwinson and an invading Norwegian force led by King Harald Hardrada.
Repeated-family continuation	The Battle of Stamford Bridge took place at the village of Stamford Bridge, East Riding of Yorkshire, on 25 September 1066, between an English army under King Harold Godwinson and an invading Norwegian force led by King Harald Hardrada.
Missing-family continuation	In September 1066, Harold defeated and killed King Harald Hardrada of Norway at the Battle of Stamford Bridge, and Harold was himself defeated and killed the following month by William the Conqueror at the Battle of Hastings.

Model	G_d	G_u	$G_u - G_d$
HOVER-BERT	+0.0430	-1.9133	-1.9563
LGED	-0.0042	+0.8886	+0.8928

Table 22: Qualitative REFUTES example. The repeated-family continuation repeats the Stamford Bridge location and combatant evidence; the missing-family continuation contributes the relation between Stamford Bridge, Harald Hardrada, and Hastings.

F Statistical Tests and Robustness Checks

This section reports statistical and construction checks behind the main comparisons: sample size, paired contrasts, redundancy depth, controls, and consistency checks for family, label, and surface constraints.

The sample-size table gives scale for resolving same-prefix effect sizes.

The paired statistical contrasts target the paper’s main diagnostic claims: duplicate stability, missing-family gain, and the LLM rank comparison under the same paired construction.

The redundancy-depth table checks that the qualitative pattern holds beyond FR-8.

Construction consistency. Across the 1,180 HOVER same-prefix pairs, the repeated-family step is drawn from an already covered family and the missing-family step is drawn from a gold family absent from the prefix. The claim, prefix, and presentation position are shared across continua-

tions. Table 26 reports the corresponding consistency checks. Matched-subset, order, leakage, and covariate analyses probe properties beyond the deterministic construction checks.

Diagnostic controls. Table 27 reports direction and sensitivity controls for the same-prefix probe. The oracle and swapped-branch rows verify the expected direction, while near-miss replacement weakens the signal. Permutation preserves the marginal gain distribution while breaking pair alignment and provides a complementary sensitivity check on the paired construction.

Surface-confound robustness. We construct a strict surface-matched subset of 157 pairs by bounding token length, entity count, number/date count, and claim overlap; LGED’s UF Pref. drops from 0.5653 to 0.5096. A relaxed matched subset of 324 pairs recovers a positive signal (0.5710, median gain 0.0326). A leakage-controlled surface classifier with claim/episode-grouped five-fold

Diagnostic quantity	Same-prefix pairs	Approx. minimum detectable effect
UF preference	1,180	0.04 absolute proportion
Paired margin gain	1,180	0.004 probability-margin units

Table 23: Sample-size calculation for the same-prefix construction at $\alpha = 0.05$. The paired-gain calculation assumes a standard deviation of 0.05 in probability-margin units.

Comparison	Metric	Difference	95% CI	Holm p
LGED vs. Verdict-only verifier	FR-8 flip	-0.2894	[-0.3040, -0.2750]	< 0.0001
LGED vs. Objective verifier	FR-8 flip	-0.2355	[-0.2482, -0.2233]	< 0.0001
LGED vs. HoVer-BERT	FR-8 flip	-0.1738	[-0.1931, -0.1556]	< 0.0001
LGED vs. HoVer-BERT	median UF gain	0.0015	[-0.0110, 0.0154]	n.s.
LGED vs. CLUSTER-WEIGHTED	median UF gain	0.0185	[0.0080, 0.0278]	< 0.001
LGED vs. HoVer-TRAINED KGAT	median UF gain	0.0208	[0.0100, 0.0341]	< 0.0001
LGED vs. QWEN2.5-72B	UF rank	0.2729	[0.2356, 0.3102]	< 0.0001

Table 24: Primary paired statistical contrasts. FR-8 rows use paired flip-rate differences, while median-gain and rank rows report same-prefix paired effects. Differences are reported as LGED minus the comparator; negative FR-8 differences and positive gain/rank differences are favorable to LGED. For stochastic trainable systems, paired contrast rows use the aligned five-seed diagnostic package used in Table 2. LLM comparisons use rank rather than margin-gain contrasts because prompt-interface scores are not pooled with classifier probabilities.

out-of-fold prediction and symmetric candidate ordering reaches AUROC 0.8876. On its 250 error pairs, HoVer-BERT, DeBERTa-v3, the objective-augmented verifier, cluster-weighted aggregation, and LGED all retain UF Preference above 0.5 with positive median paired gain. Agreement between the classifier-preferred branch and verifier preference is 0.4653–0.5568, and surface confidence has non-positive Spearman correlation with verifier gain. Replacing the exact duplicate with non-identical evidence from the same covered family also preserves positive UF for multiple verifier designs. Finally, on 208 naturally occurring converse pairs where missing-family evidence has at least as much prefix overlap as covered-family evidence, DeBERTa-v3 obtains UF 0.6154 and median paired gain 0.0195. A deliberately constructed absolute-high-overlap set would extend this control beyond naturally occurring relative-overlap pairs.

Evidence order and duplicate placement. Order and placement analyses compare random prefix shuffles and duplicate append at end versus interleaved placement. These analyses probe whether the main redundancy and family-comparison results depend on presentation order.

Role-guidance analysis. The role diagnosis head is reliable for same-family steps but has low hard missing-family recall. The mechanism analysis therefore centers on soft role scores and state updates, with hard role predictions reported as a supplementary diagnostic.

KGAT mapping and graph-verifier settings.

All KGAT scoring uses the same three-way-to-binary probability adapter as the main tables. KGAT settings distinguish released FEVER-trained references from three-seed in-domain HoVer retraining in the appendix tables.

Model	FR-1	FR-2	FR-4	FR-8	FR-AUC	UF Pref.	Med. Gain
HoVer-BERT	0.0914	0.1346	0.1670	0.1902	0.1468	0.5932	0.0005
Objective verifier	0.2044	0.1988	0.2141	0.2215	0.1985	0.6127	0.0585
CLUSTER-WEIGHTED	0.0244	0.0318	0.0369	0.0426	0.0335	0.5008	0.0005
LGED	0.0040	0.0068	0.0085	0.0165	0.0091	0.5653	0.0219

Table 25: Metric-robustness results over redundancy depths. FR-AUC aggregates FR-1, FR-2, FR-4, and FR-8. The same qualitative pattern holds across shallow and deeper redundancy conditions.

Check	Pass rate	Fail count
Duplicate family already covered by prefix	1.0000	0
Missing-family continuation absent from prefix	1.0000	0
Missing-family continuation belongs to gold families	1.0000	0
Repeated-family continuation preserves covered-family assignment	1.0000	0
Repeated-family entity spans preserved from prefix	1.0000	0
Repeated-family number/date spans preserved from prefix	1.0000	0
Gold label unchanged	1.0000	0
Hidden construction markers absent from model inputs	1.0000	0

Table 26: Construction consistency checks over the same-prefix pairs. The checks report whether the construction satisfies the intended family, label, and surface-form constraints; lexical similarity is evaluated separately through the surface-control analyses.

Control	UF Pref.	Med. Gain	Role
Oracle	1.0000	n/a	positive
Original	0.5653	0.0219	signal
Swap	0.4347	−0.0219	reverses
Permute	0.5661	0.0215	marginal
Near-miss	0.5475	0.0167	weaker

Table 27: Diagnostic controls for the LGED reference. Oracle and swapped-branch rows check direction, near-miss replacement weakens the signal, and permutation tests sensitivity to pair alignment while preserving marginal gain distributions.