

ETrACER: Polarity-Aware Evidence Grounding

Bohao Chu^{1,*} Qianli Wang² Hendrik Damm³
Shuning Zhang⁴ Hui Wang¹ Andrea Papenmeier⁵ Georg Lodde⁶
Elisabeth Livingstone⁶ Christoph M. Friedrich³ Norbert Fuhr¹

¹University of Duisburg-Essen, ²Technische Universität Berlin,

³University of Applied Sciences and Arts Dortmund,

⁴Tsinghua University, ⁵University of Twente, ⁶University Hospital Essen,

*Correspondence: bohao.chu@qq.com

Abstract

Evidence grounding makes generated content verifiable by linking statements to their sources. However, effective verifiability requires two properties that existing methods fail to jointly provide: *verification granularity*, the ability to localize fine-grained evidence for each statement, and *verification completeness*, the ability to capture both support and contradiction. Existing methods either operate at granularities too coarse for rapid inspection or are support-only. We formalize *polarity-aware evidence grounding*: linking each response sentence to its contextual evidence with a supporting, contradicting, or irrelevant label. We construct ETrACER-BENCH, a human-annotated benchmark with 7,905 supporting and contradicting sentence pairs across seven domains, covering both explicit and implicit contradictions. We further propose ETrACER, a post-hoc framework whose context-aware signed scorer recovers the full evidence map in a single forward pass per response, with localization and polarity emerging jointly from one mechanism. ETrACER achieves the best overall Citation F1 of 82.39, driven by a Contradict F1 of 89.20. A pre-registered within-subjects user study ($n=30$) shows that polarity-aware grounding nearly doubles contradiction detection over the document-level baseline (27.9%→51.9%, Cohen’s $d=0.82$, $p=0.0003$), and suggests a false-completeness bias under which support-only grounding fails to improve over this baseline (23.4% vs. 27.9%). Dataset and code: <https://github.com/chubohao/eTracer>.

1 Introduction

Retrieval-augmented generation (RAG) produces responses conditioned on retrieved documents, yet these responses remain difficult to verify: users cannot efficiently determine which statements are supported, contradicted, or ungrounded (Min et al., 2023; Gao et al., 2023). This problem is particularly acute when retrieved documents contain con-

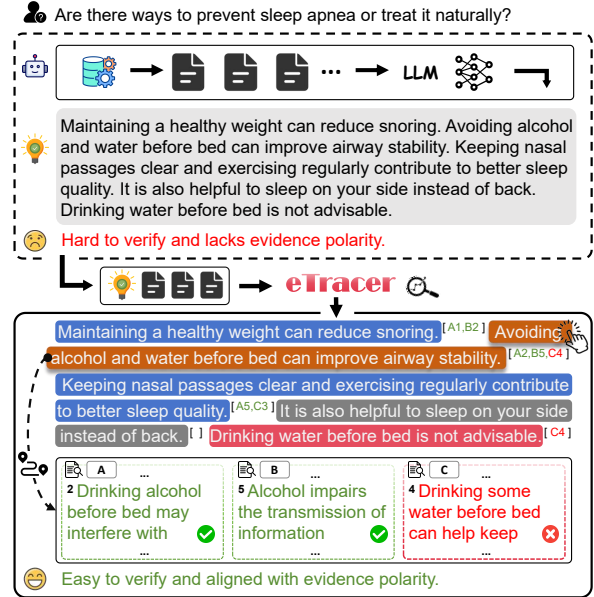


Figure 1: ETrACER grounds each response sentence in its supporting and contradicting context sentences.

flicting or outdated information (Gupta et al., 2024; Chu et al., 2026b). Effective verifiability requires two properties that existing methods fail to jointly provide: *verification granularity*, the ability to localize fine-grained evidence for each statement, and *verification completeness*, the ability to capture both supporting and contradictory evidence.

Regarding *verification granularity*, existing methods are too coarse on at least one side of the grounding link. On the evidence side, both response-to-passage methods (Perplexity AI, 2026; Gao et al., 2023) and claim-to-passage methods (Min et al., 2023; Tang et al., 2024) ground statements against entire passages, forcing users to scan long text for the relevant sentences. On the response side, claim-to-passage methods additionally decompose responses into atomic claims, a step that introduces decomposition errors and produces units misaligned with how users read (Wanner et al., 2024; Gunjal and Durrett, 2024). Sentences are the natural unit of both writing and verification, offer-

ing a balance between precision and human readability. We therefore adopt *sentence-to-sentence* grounding, linking each response sentence directly to the contextual sentences that evidence it.

Regarding *verification completeness*, the dominant paradigms are *support-only* grounding (Honovich et al., 2022; Min et al., 2023; Xie et al., 2024) and polarity-collapsed consistency scoring (Laban et al., 2022; Zha et al., 2023); neither separates ungrounded statements from *actively contradicted* ones. This blind spot is consequential when sources conflict, as in evidence-based medicine (Sackett et al., 1996) and multi-outlet news synthesis (Fabri et al., 2019). Natural language inference (NLI) models do classify polarity, but only on isolated sentence pairs (Williams et al., 2018), and this pair-wise paradigm faces two issues in RAG. First, it presupposes that the relevant context sentences have already been localized. Second, scoring pairs independently discards two cross-sentence signals critical for polarity classification: the *contrastive cue* between co-occurring supporting and contradicting sentences of similar surface form, and *cross-sentence aggregation*, where evidence is distributed across multiple sentences that jointly establish support or imply a contradiction none states directly.

We present ETRACER, a post-hoc framework whose context-aware signed scorer recovers the full evidence map in a single forward pass per response, with localization and polarity emerging jointly from one mechanism. As illustrated in Figure 1, the resulting signed map grounds each response sentence in its supporting and contradicting context sentences, directly addressing the granularity and completeness gaps identified above. The main contributions are as follows (C1–4):

C1: We formalize *polarity-aware evidence grounding*, a sentence-to-sentence task that makes RAG responses verifiable through fine-grained evidence alignment with their sources, jointly localizing evidence and classifying its polarity (§3).

C2: We construct ETRACER-BENCH, a human-annotated dataset of 7,905 sentence-to-sentence evidence pairs across seven domains and two contradiction difficulties (explicit and implicit), filling a gap in existing benchmarks that lack three-way polarity at this fine-grained granularity (§4).

C3: We propose ETRACER, whose context-aware scorer achieves the best overall Citation F1 of 82.39, driven by a Contradict F1 of 89.20 (+19.16 over its matched pair-wise counterpart) (§5–§7).

C4: A pre-registered user study ($n=30$) confirms our primary hypothesis that polarity-aware grounding nearly doubles contradiction identification over the document-level baseline ($d=0.82$), and shows signs of a *false-completeness bias* under which support-only grounding offers no such gain (§8).

2 Related Work

2.1 Evidence Granularity

Finer-grained evidence grounding enables faster human verification (Slobodkin et al., 2024; Chu et al., 2026a). Existing attribution methods vary along two orthogonal axes: the response unit being verified and the evidence unit it is grounded against. Traditional RAG systems such as Perplexity and Microsoft Copilot link response sentences to entire web pages (Perplexity AI, 2026; Microsoft, 2026). FActScore (Min et al., 2023) and MiniCheck (Tang et al., 2024) decompose responses into atomic claims but verify them against entire passages, forcing users to scan the full source to confirm any single claim. LongCite (Zhang et al., 2025), TracSum (Chu et al., 2025), and AGREE (Slobodkin et al., 2024) achieve sentence-to-sentence grounding but remain exclusively support-oriented. In contrast, ETRACER adopts the same granularity and additionally resolves evidence polarity, explicitly distinguishing support from contradiction.

2.2 Polarity Classification

Distinguishing whether evidence supports or contradicts a statement is essential for complete verification, especially when sources conflict. TRUE (Honovich et al., 2022), DocLens (Xie et al., 2024), and LongCite (Zhang et al., 2025) reduce the output to binary support labels, whether via NLI or LLM-based consistency judgments, collapsing contradiction into the unsupported category. SummaC (Laban et al., 2022) and AlignScore (Zha et al., 2023) can in principle detect contradiction, but aggregate pair-wise scores into a single consistency judgment, discarding sentence-level localization. ETRACER jointly localizes evidence and resolves polarity in a single forward pass, recovering both signals existing methods leave behind.

2.3 Evaluation Benchmarks

Evaluating polarity-aware evidence grounding requires benchmarks that provide both fine-grained evidence annotations and three-way polarity labels. Table 1 summarizes existing RAG benchmarks

Benchmark	Resp.	Evid.	Polarity	Annot.
LFRQA	sent.	doc.	S/U	human
ALCE	sent.	pass.	S/U	NLI
HAGRID	sent.	pass.	S/U	human+LLM
LongCite	sent.	sent.	S/U	GPT-4o
ETRACER-BENCH	sent.	sent.	S/C/I	human

Table 1: Benchmark comparison. **Resp./Evid.:** granularity of the response/evidence unit (sent.=sentence, doc.=document, pass.=passage). **Polarity:** S=support, C=contradict, I=irrelevant, U=unsupported. **Annot.:** annotation source.

along these two dimensions; none satisfies both. LFRQA (Han et al., 2024) is human-annotated but offers only document-level citations without contradiction labels. ALCE (Gao et al., 2023) and HAGRID (Kamalloo et al., 2023) ground responses at passage granularity and remain support-only. LongCite (Zhang et al., 2025) reaches sentence-to-sentence granularity but still lacks contradiction labels, and its annotations are produced by GPT-4o rather than humans. ETRACER-BENCH (§4) closes both gaps with human-annotated sentence-to-sentence evidence links spanning two contradiction difficulty levels (explicit and implicit).

3 Task Formulation

We formalize *polarity-aware evidence grounding* at sentence-to-sentence granularity. Given a generated response and its source context, the task is to (i) localize the context sentences that serve as evidence for each response sentence, and (ii) classify each link as supporting, contradicting, or irrelevant. Below we define the grounding function (§3.1).

3.1 Polarity-Aware Evidence Grounding

Let $\mathcal{R} = \{r_i\}_{i=1}^n$ denote a generated response segmented into n sentences, and let $\mathcal{S} = \{s_j\}_{j=1}^m$ denote its source context segmented into m sentences. Polarity-aware evidence grounding determines, for every response–context sentence pair, whether the context sentence provides supporting evidence, contradicting evidence, or no relevant evidence. Formally, a *polarity-aware grounding function* $\phi : \mathcal{R} \times \mathcal{S} \rightarrow \{-1, 0, +1\}$ assigns a signed label to each pair (r_i, s_j) according to their evidential relationship:

$$\phi(r_i, s_j) = \begin{cases} +1, & \text{if } s_j \text{ supports } r_i, \\ -1, & \text{if } s_j \text{ contradicts } r_i, \\ 0, & \text{if } s_j \text{ is irrelevant to } r_i. \end{cases}$$

Applying ϕ to all $n \times m$ sentence pairs yields a matrix $\Phi \in \{-1, 0, +1\}^{n \times m}$, which we call the

signed evidence map. Its rows correspond to response sentences and its columns to context sentences. Nonzero entries localize the evidence associated with each response sentence, while their signs distinguish supporting from contradicting evidence. The map therefore represents evidence localization and polarity within a single sentence-level structure.

4 ETRACER-BENCH

As shown in Table 1, no existing benchmark jointly provides sentence-to-sentence granularity and three-way polarity annotation. We therefore construct ETRACER-BENCH, a human-annotated benchmark of 7,905 sentence-to-sentence evidence pairs across seven domains with two levels of contradiction difficulty (explicit and implicit). Full construction details are provided in Appendix A.

Base corpus. We extend LFRQA (Han et al., 2024), a RAG-QA dataset of human-written long-form answers with document-level citations. We sample 100 responses per domain, yielding 700 responses across seven domains: biomedical (BIOASQ), finance (FIQA), and the LoTTE subsets TECH, SCI, LIFE, WRITING, and REC.

Annotation protocol. For each response sentence, two trained annotators independently identify SUPPORT and CONTRADICT sentences from the cited documents, with all remaining sentences treated as IRRELEVANT. A third annotator adjudicates the 7.3% disagreements. Inter-annotator agreement (Cohen, 1960) is $\kappa=0.87$ on evidence identification (evidence vs. irrelevant) and $\kappa=0.92$ on polarity classification among the sentences both annotators marked as evidence (Appendix A.4). This yields 4,565 supporting and only 94 naturally occurring contradicting pairs (see Table 2), motivating the adversarial augmentation below.

Dataset split. We split the 700 responses at the response level, stratified by domain, into train/dev/test with a 70/10/20 per-domain ratio. Response-level splitting ensures each response and its source documents appear in exactly one split, preventing context leakage. The dev split is used exclusively for model selection, hyperparameter tuning, and few-shot demonstration sampling; all reported results are on the held-out test split.

Contradiction construction. To compensate for the scarcity of natural contradictions, we adversarially augment the corpus at two difficulty levels: **explicit** contradictions use surface-level alterations

Domain	Resp.	Ctx	S	C _n	C _e	C _i
BioASQ	163	2,701	412	1	144	129
FiQA	285	2,753	749	16	179	393
LoTTE-Life	315	2,601	821	21	200	398
LoTTE-Rec	214	1,899	585	12	154	238
LoTTE-Sci	233	2,344	686	16	159	300
LoTTE-Tech	266	2,174	634	12	173	327
LoTTE-Writ	260	2,322	678	16	196	256
Total	1,736	16,794	4,565	94	1,205	2,041

Table 2: Statistics of ETRACER-BENCH. Resp.: response sentences. Ctx: context sentences. S: supporting pairs. C_n: naturally contradicting pairs. C_e/C_i: adversarially constructed explicit/implicit contradicting pairs.

(negation, antonym substitution, factual replacement), while **implicit** ones introduce logically incompatible statements that preserve the topic but require multi-step inference. Each constructed paragraph is placed in a new document and appended to the context, which is then shuffled to simulate conflicting retrieval while eliminating positional bias. To prevent stylistic leakage, three disjoint annotators construct contradictions for the train/dev/test splits, with a fourth performing blind verification (Appendix A.6). This yields 1,205 explicit and 2,041 implicit contradicting pairs (Table 2).

5 ETRACER Framework

ETRACER rests on a structural correspondence: the task output is an $n \times m$ signed matrix over response–context sentence pairs, and a transformer encoder over the concatenated sequence already produces $n \times m$ boundary-token interactions under full self-attention. We exploit this directly. Rather than classifying each pair (r_i, s_j) in isolation, ETRACER encodes the full response and full context in a single forward pass and reads off all pairs from their boundary-token representations. This preserves two cross-sentence signals that pair-wise scoring discards by construction. The first is the *contrastive cue* between co-occurring supporting and contradicting sentences of similar surface form; the second is *cross-sentence aggregation* of evidence distributed across multiple sentences. Both emerge from vanilla attention within the encoded sequence (validated in §7.2 and §7.5), while inference collapses from $n \times m$ forward passes to one.

5.1 Context-Aware Signed Scoring

Given a response $\mathcal{R} = \{r_i\}_{i=1}^n$ and its context $S = \{s_j\}_{j=1}^m$, the context-aware scorer $\mathcal{M}_{con}$ produces a signed evidence map $\Phi \in \{+1, -1, 0\}^{n \times m}$ in a single forward pass. The response and context are

packed into one input sequence with learned separators [R_SEP] and [S_SEP] (initialized from [SEP] and fine-tuned) marking sentence boundaries, and the native [SEP] delimiting the two blocks:

$$[\text{CLS}] \underbrace{r_1 [\text{R_SEP}] \cdots r_n [\text{R_SEP}]}_{\text{response}} [\text{SEP}] \underbrace{s_1 [\text{S_SEP}] \cdots s_m [\text{S_SEP}]}_{\text{context}}$$

Full self-attention conditions every boundary token on the entire input, so each h_i^r and h_j^s already encodes its sentence in the context of all others. The signed label for any pair (r_i, s_j) then reduces to a two-token readout:

$$p(\cdot \mid r_i, s_j) = \text{softmax}(\text{MLP}([h_i^r; h_j^s])),$$

with $\phi(r_i, s_j) = \arg \max p(\cdot \mid r_i, s_j) \in \{+1, -1, 0\}$ at inference. Stacking the readouts over all $n \times m$ pairs recovers the full evidence map Φ in a single forward pass: the cross-sentence reasoning that the task fundamentally demands and the cross-sentence attention that the encoder already performs are the same underlying computation.

5.2 Training

Training data. The context-aware scorer $\mathcal{M}_{con}$ is trained on ETRACER-BENCH (train split), which provides sentence-to-sentence polarity annotations between multi-sentence responses and their full source contexts. Each training instance is a tuple $(\mathcal{R}, S, \mathbf{Y})$ where $\mathcal{R} = [r_1, \dots, r_n]$ is the full response, $S = [s_1, \dots, s_m]$ is the concatenation of all source document sentences, and $\mathbf{Y} \in \{+1, -1, 0\}^{n \times m}$ is the polarity label matrix, matching the inference-time input exactly. To mirror the variable document orderings returned by retrieval at deployment, we augment the train split (Appendix B.4): document-order shuffling (8 seeds) combined with two contradiction-type views, expanding $490 \rightarrow 5,386$ training instances.

Backbone and objective. We initialize $\mathcal{M}_{con}$ from ModernBERT-large (Warner et al., 2025) (395M parameters), whose bidirectional attention and native 8,192-token context window accommodate the concatenated response and context of every ETRACER-BENCH instance within a single sequence, preserving the cross-sentence signals that motivate our design (§5.1).¹ $\mathcal{M}_{con}$ is optimized with class-weighted cross-entropy loss summed over all $n \times m$ response–context sentence pairs:

¹Longer contexts are handled by a document-level sliding-window extension, which stays within 0.8 Citation F1 of the native-window result even at $T=1,024$ (Appendix C.8).

$$\mathcal{L} = \sum_{i=1}^n \sum_{j=1}^m w_{y_{ij}} \text{CE}(\hat{y}_{ij}, y_{ij})$$

where $w_{y_{ij}} \in \{w_{\text{con}}, w_{\text{irr}}, w_{\text{sup}}\}$ weighs each class according to its label y_{ij} ; the specific values are selected on the dev split (see §6.3).

6 Experiments

We evaluate ETRACER on the ETRACER-BENCH test split (§4) and compare it against strong baselines spanning two paradigms: NLI-based pairwise scoring and end-to-end LLM prompting. Implementation details are reported in Appendix B.

6.1 Baselines

Our task requires three-way polarity labels at the sentence-to-sentence level (§3), which rules out two classes of prior methods. Support-only methods (Min et al., 2023; Tang et al., 2024; Honovich et al., 2022; Zhang et al., 2025; Xie et al., 2024) cannot separate contradiction from irrelevance. Aggregation-based methods (Laban et al., 2022; Zha et al., 2023) compute a pair-wise NLI matrix internally but collapse it into a single consistency score, discarding the sentence-level polarity our task demands; their upstream pair-wise NLI scoring is conceptually similar to our NLI baselines. The remaining applicable baselines fall into two families: pair-wise scoring and end-to-end LLM. Checkpoints and access details are in Table 13.

Pair-wise scoring. These methods score each response-context sentence pair (r_i, s_j) independently. We evaluate DeBERTa-v3-large (He et al., 2023) and BART-large (Lewis et al., 2020) in two configurations: off-the-shelf (fine-tuned on MNLI) and fine-tuned on the ETRACER-BENCH train split under the same class-weighted cross-entropy objective as ETRACER. We additionally train $\mathcal{M}_{\text{pair}}$, a pair-wise variant sharing ETRACER’s backbone, training data, and loss, and differing only in input construction, isolating cross-sentence context as the axis of variation (verified in §7.5).

LLM prompting. As representative LLM baselines, we evaluate the closed-source GPT-4o (Hurst et al., 2024) and the open-source Qwen3-235B-Instruct (Yang et al., 2025) under zero-shot and 7-shot settings. These two models cover both proprietary and open-source families and serve to characterize the difficulty of the task at scale, rather than to enumerate the current LLM frontier exhaustively. Each model receives the full response and

full context, and outputs structured sentence-level polarity labels. Prompts are listed in Appendix E.

6.2 Metrics

We evaluate pair-level polarity labels against gold annotations. Support F1 and Contradict F1 score each (r_i, s_j) prediction, and Citation F1 is their macro average. We further decompose Contradict F1 into Contradict_e F1 (explicit) and Contradict_i F1 (implicit) over surface-level and inference-requiring contradictions (§4). Confidence intervals on the main metrics are 95% paired bootstrap over response sentences (10,000 replicates) (Koehn, 2004); for the small natural-contradiction subset (Appendix C.6), we additionally report Wilson 95% CIs (Wilson, 1927).

6.3 Implementation

All fine-tuned models are initialized from their respective backbones and trained on the ETRACER-BENCH train split with identical class-weighted cross-entropy loss and optimization hyperparameters; $\mathcal{M}_{\text{con}}$ and $\mathcal{M}_{\text{pair}}$ differ only in input construction (§5.1). Class weights $[w_{\text{sup}}, w_{\text{irr}}, w_{\text{con}}]$ are selected via grid search on the dev split for $\mathcal{M}_{\text{con}}$, sweeping $w_{\text{irr}} \in \{0.4, 0.5, 1, 2, 4, 6, 8, 10, 12\}$ with $w_{\text{sup}} = w_{\text{con}} = 1$; the resulting optimum $[1, 10, 1]$ up-weights the dominant irrelevant class, a counterintuitive choice reflecting that the bottleneck is reliable non-evidence rejection rather than minority-class recall (Appendix C.3). The same value is applied uniformly to all pair-wise baselines; this choice never disadvantages any baseline by more than 0.8 dev Citation F1 relative to its own optimum, and in two of three cases yields improved or comparable results (Appendix C.4). Final checkpoints are selected by dev Citation F1.

7 Results and Analysis

7.1 Main Results

Table 3 reports polarity-aware grounding performance on the ETRACER-BENCH test set, from which three findings emerge.

1) Context-awareness drives contradiction detection. ETRACER reaches 89.20 Contradict F1 (95% CI [87.31, 91.16]), +19.16 over the matched pair-wise variant $\mathcal{M}_{\text{pair}}$ (70.04, [66.91, 72.96]); the two intervals are disjoint by nearly 15 points, so the gap is statistically separable and attributable to context-aware scoring alone since backbone, data, and loss are shared. By contrast, Support F1 improves by

Category	Method	Arch.	Params	Infer.	S.P	S.R	S.F1	C.P	C.R	C.F1	Cit.F1
Prompt	GPT-4o (0-shot) [†]	Dec-only	—	1*	88.60	68.96	77.56	<u>91.02</u>	48.64	63.40	70.48
	GPT-4o (7-shot) [†]	Dec-only	—	1*	<u>85.39</u>	71.95	<u>78.10</u>	87.31	64.96	74.50	76.30
	Qwen3-235B (0-shot) [†]	Dec-only	235B	1*	79.18	72.95	75.94	92.42	70.24	79.82	77.88
	Qwen3-235B (7-shot) [†]	Dec-only	235B	1*	78.80	77.49	78.14	89.51	<u>81.92</u>	<u>85.55</u>	<u>81.85</u>
Fine-tuned	DeBERTa-v3 (MNLI) [‡]	Enc-only	435M	73	66.03	34.48	45.30	52.15	69.92	59.74	52.52
	BART-large (MNLI) [‡]	Enc-Dec	406M	73	69.84	34.15	45.87	40.47	68.32	50.83	48.35
	DeBERTa-v3 (ETB)	Enc-only	435M	73	71.51	<u>73.73</u>	72.60	74.03	76.64	75.31	73.96
	BART-large (ETB)	Enc-Dec	406M	73	74.27	<u>70.40</u>	72.28	71.75	77.60	74.56	73.42
	$\mathcal{M}_{pair}$ (ETB)	Enc-only	395M	73	74.70	70.07	72.31	74.11	66.40	70.04	71.18
	$\mathcal{M}_{con}$ (ETRACER)	Enc-only	395M	1	84.10	68.63	75.58	90.59	87.84	89.20	82.39

Table 3: Main grounding results on the ETRACER-BENCH test set. **Infer.**: forward passes per response (pair-wise requires ≈ 73 on average; context-aware requires 1; 1* denotes a single prompt for LLMs). ETB: fine-tuned on ETRACER-BENCH train split. S./C.: support/contradict. **Bold**: best, underline: second best. [†]: API-based; GPT-4o parameter count is not publicly disclosed. [‡]: off-the-shelf, no task-specific training. 95% bootstrap confidence intervals are reported in Table 15 (Appendix C.1).

only +3.27 over $\mathcal{M}_{pair}$, with overlapping intervals (75.58, [73.03, 77.98] vs. 72.31, [69.90, 74.74]), consistent with support being well-served by local cues while contradiction requires comparing co-occurring sentences of similar surface form (§7.5). Robustness across contradiction types follows the same pattern: ETRACER’s implicit-vs-explicit gap is the smallest among all methods ($\Delta = -9.77$, Table 5; §7.2).

2) Competitive performance at two orders of magnitude fewer parameters. ETRACER matches Qwen3-235B 7-shot on Citation F1 (82.39, [80.78, 83.89] vs. 81.85, [80.31, 83.31]; overlapping CIs, statistically on par) in a single forward pass without few-shot demonstrations, at roughly 1/600 the parameter count. This indicates that polarity-aware grounding is amenable to specialized small architectures rather than requiring general-purpose LLM scale.

3) Targeted supervision dominates prompt scaling. Fine-tuning on ETRACER-BENCH lifts MNLI checkpoints by +21.44 (DeBERTa-v3) and +25.07 (BART-large) Citation F1, dwarfing the gains from few-shot demonstrations over zero-shot (+5.82 for GPT-4o, +3.97 for Qwen3-235B). ETRACER (395M, fine-tuned) remains statistically on par with Qwen3-235B 7-shot (82.39 vs. 81.85 Citation F1, overlapping CIs) despite a $\sim 600\times$ parameter gap, indicating that for polarity-aware grounding, task-specific supervision is at least as parameter-efficient as in-context scaling.

4) The context-aware advantage persists across training seeds. In addition to the original seed 42, we train ETRACER and the matched $\mathcal{M}_{pair}$ baseline with seeds 52 and 62, following the same

Model	S.F1	C.F1	Cit.F1
ETRACER	75.47 $\pm$ 0.10	88.71 $\pm$ 1.10	82.09 $\pm$ 0.58
$\mathcal{M}_{pair}$	72.54 $\pm$ 0.61	71.03 $\pm$ 1.07	71.79 $\pm$ 0.81
Δ	+2.93 $\pm$ 0.63	+17.69 $\pm$ 1.33	+10.31 $\pm$ 0.79

Table 4: Mean $\pm$ standard deviation across training seeds {42, 52, 62}. Each Δ is computed between ETRACER and $\mathcal{M}_{pair}$ for the same seed.

Method	C.F1	C _e .F1	C _i .F1	Δ
GPT-4o (7-shot)	74.50	89.50	61.54	−27.96
Qwen3-235B (7-shot)	<u>85.55</u>	<u>91.31</u>	<u>81.48</u>	−9.83
DeBERTa-v3 (MNLI)	59.74	67.26	53.30	−13.96
BART-large (MNLI)	50.83	57.36	45.16	−12.20
DeBERTa-v3 (ETB)	75.31	83.71	69.35	−14.36
BART-large (ETB)	74.56	83.83	68.14	−15.69
$\mathcal{M}_{pair}$ (ETB)	70.04	82.82	59.91	−22.91
$\mathcal{M}_{con}$ (ETRACER)	89.20	95.18	85.41	−9.77

Table 5: Contradict F1 by contradiction type. C_e/C_i: explicit/implicit; Δ : C_i−C_e (smaller = more robust). **Bold**/underline: best/second best.

data-processing, optimization, and model-selection protocol. As shown in Table 4, ETRACER yields a mean Citation F1 gain of +10.31 over $\mathcal{M}_{pair}$, driven primarily by Contradict F1 (+17.69) rather than Support F1 (+2.93). The Citation F1 gain is positive for each of the three seeds, indicating that the context-aware advantage is not merely an artifact of a particular random seed.

7.2 Contradiction Detection by Type

Table 5 stratifies Contradict F1 by explicit and implicit contradictions (§4). All methods handle explicit contradictions reasonably, where surface cues such as negation and antonyms suffice. Implicit contradictions widen the gap: fine-tuned

pair-wise methods drop 14.4 . . . 22.9 points (Δ column), whereas ETRACER drops only 9.77 and retains 85.41 C_i F1, the smallest gap among all methods. Even Qwen3-235B 7-shot, the strongest LLM baseline ($\Delta = -9.83$), trails ETRACER on both types (+3.87 C_e , +3.93 C_i). This robustness is the more diagnostic signal, as it cannot stem from training-set style transfer: the train/dev/test contradictions were constructed by three disjoint annotators (§4). It is corroborated by two independent analyses, attention masking (§7.5) and natural-vs-constructed generalization (§7.3), confirming that context-aware scoring is the architectural ingredient closing the C_e/C_i gap.

7.3 Generalization to Naturally Occurring Contradictions

Only 24 naturally occurring contradiction pairs appear in the test split, limiting the precision of real-world robustness estimates. We therefore conduct two analyses to assess whether the observed advantage extends beyond constructed contradictions.

Generalization by contradiction structure. The 24 natural test pairs comprise 12 self-contained contradictions and 12 cross-sentence contradictions requiring additional context. As shown in Table 6, ETRACER correctly identifies 18/24 pairs, compared with 1/24 for $\mathcal{M}_{pair}$; an exact paired McNemar test confirms a significant overall difference ($p=1.53\times 10^{-5}$). The larger cross-sentence gap (10/12 vs. 0/12) is consistent with context-aware evidence aggregation, although the subgroup results remain descriptive given their small size.

Model	All	Self	Cross-sent.
ETRACER	18/24	8/12	10/12
$\mathcal{M}_{pair}$	1/24	1/12	0/12

Table 6: Accuracy on natural contradictions. Self: self-contained; Cross-sent.: cross-sentence.

Training setting	Natural accuracy
Full model	18/24
Constructed-only	18/24

Table 7: Natural-test accuracy by training supervision.

Transfer without natural-contradiction supervision. We train an additional ETRACER model while masking all natural contradiction pairs from the training loss and excluding them from the development-set model-selection metric, retaining

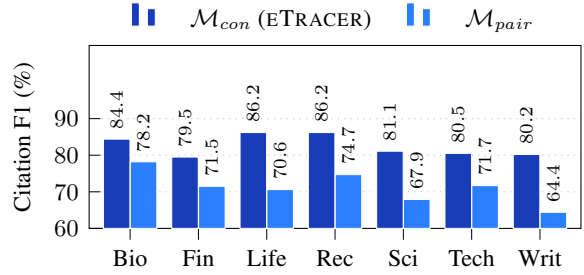


Figure 2: Per-domain Citation F1 for ETRACER and the controlled pair-wise baseline $\mathcal{M}_{pair}$. Full breakdown in Appendix C.2.

all constructed contradictions and other supervision. As shown in Table 7, removing natural-contradiction supervision leaves aggregate natural-test accuracy unchanged at 18/24. This provides preliminary evidence that supervision from constructed contradictions transfers to natural conflicts, although the small test set and single training run preclude a stronger conclusion.

7.4 Domain-wise Analysis

Figure 2 reports per-domain Citation F1 on the ETRACER-BENCH test split. ETRACER outperforms $\mathcal{M}_{pair}$ in all seven domains, with Citation F1 ranging from 79.5 (FiQA) to 86.2 (Lifestyle/Recreation). Gains are largest in Writing (+15.8), Lifestyle (+15.6), and Science (+13.2), where contradictions hinge on specific claims, mechanisms, or factual details that reward cross-sentence comparison, and smallest in BioASQ (+6.2), where $\mathcal{M}_{pair}$ already attains its strongest pair-wise Citation F1 (78.2) and leaves less headroom. The asymmetry is sharper on contradiction alone: ETRACER exceeds 85 Contradict F1 in every domain, peaking at 95.3 (Recreation) and 93.5 (BioASQ), whereas $\mathcal{M}_{pair}$ falls as low as 63.4 (Writing) and 63.7 (Science), yielding per-domain contradiction gaps up to +25.7 (Lifestyle) and +24.1 (Science) that underscore the value of context-aware scoring in analytically intensive domains.

7.5 Ablation Study

Table 8 isolates the contribution of context-aware scoring by selectively masking cross-sentence attention at test time. All ablations share the same trained ETRACER model, identical input sequences, and identical decoding procedure; only the attention mask differs, so any performance change reflects the masked information flow alone.

Masking context-side cross-attention. When each context sentence can see the full response

System	S.F1	C.F1	Cit.F1	Δ
ETRACER (full)	75.58	89.20	82.39	—
Mask context cross-attn.	57.47	75.58	66.53	−15.86
Mask response cross-attn.	73.78	85.22	79.50	−2.89
Mask both	60.04	67.56	63.80	−18.59

Table 8: Ablation on the ETRACER-BENCH test set. All rows share the same trained ETRACER model with selective attention masking at test time. Δ : change in Citation F1 vs. full system.

but not any other context sentence, Support F1 drops by 18.11 and Contradict F1 by 13.62 (Cit.F1: −15.86). The decline on Support reveals what we term *joint evidence aggregation*: multiple context sentences each supply partial information that only jointly constitutes sufficient evidence, so disabling cross-context attention forces the model into pair-wise reasoning and erases this signal. The drop on Contradict confirms a complementary role: polarity discrimination relies on comparing co-occurring context sentences, since a supporting and a contradicting sentence often share surface form and differ only in polarity cues; without seeing both, the model loses the contrastive signal needed to tell them apart. This is the most impactful single-axis ablation, establishing context-side cross-attention as the dominant source of ETRACER’s advantage over pair-wise scoring.

Masking response-side cross-attention. When each response sentence can see all context sentences but not any other response sentence, Support F1 drops by 1.80 and Contradict F1 by 3.98 (Cit.F1: −2.89). The modest decline suggests that response-side cross-attention contributes secondarily, likely by resolving coreference across response sentences, for instance linking a pronoun in r_2 back to an entity introduced in r_1 , rather than driving polarity classification directly.

Masking both. Isolating context and response sentences simultaneously drops Cit.F1 by 18.59, only marginally beyond the context-side ablation alone (−15.86). Since the response-side ablation costs −2.89 Cit.F1 in isolation and contributes −2.73 on top of the context-side ablation, the two signals operate near-additively: their effects accumulate without meaningful overlap, with the context side clearly dominant ($\sim 85\%$ of the joint degradation).

Context shuffling. To verify that ETRACER exploits cross-sentence *structure* rather than mere co-occurrence, we shuffle the order of context sentences at inference time. $\mathcal{M}_{pair}$ is invariant to

Condition	S.F1	C.F1	Cit.F1
$\mathcal{M}_{con}$ (original)	75.58	89.20	82.39
$\mathcal{M}_{con}$ (shuffled)	70.42 $\pm$ 0.12	70.35 $\pm$ 1.60	70.39 $\pm$ 0.85
Δ	−5.16	−18.85	−12.00
$\mathcal{M}_{pair}$ (invariant)	72.31	70.04	71.18

Table 9: Context shuffling on the test split. Shuffled results: mean $\pm$ std over 3 seeds $\in \{42, 52, 62\}$. $\mathcal{M}_{pair}$ serves as a null control, invariant to shuffling by construction. Shuffled $\mathcal{M}_{con}$ converges to $\mathcal{M}_{pair}$ ’s native level across all pair-level metrics.

this perturbation by construction and serves as a null control. Shuffling drops Contradict F1 by 18.85 points and Support F1 by only 5.16 (3 seeds $\in \{42, 52, 62\}$, std ≤ 1.60 ; Appendix C.5). The asymmetry is diagnostic: contradiction detection hinges on relative positions among co-occurring sentences, whereas support detection relies more on local sentence-pair semantics. Strikingly, shuffled $\mathcal{M}_{con}$ converges almost exactly to $\mathcal{M}_{pair}$ ’s native level (S.F1: 70.42 vs. 72.31; C.F1: 70.35 vs. 70.04; Cit.F1: 70.39 vs. 71.18): stripping cross-sentence structure leaves $\mathcal{M}_{con}$ at essentially the same operating point as a model structurally incapable of seeing it. The context-aware gain (Table 3) is thus *entirely* mediated by cross-sentence structure, not by incidental factors such as parameter count or effective training data.

8 User Study

Automatic metrics measure label agreement; their value rests on whether they translate to better human verification. To isolate the value of the grounding paradigm from model accuracy, all evidence highlights in the user study are drawn from gold annotations rather than from ETRACER predictions. This prevents prediction errors from confounding the comparison and allows us to evaluate the support-only and polarity-aware paradigms under correct evidence grounding. We test this with a pre-registered², within-subjects experiment ($n=30$, recruited via Prolific). Full protocol is described in Appendix D.

Design. We compare three conditions: **(D)** *document-level*, the response beside its full source without citation; **(S)** *support-only*, where clicking a response sentence highlights its supporting context sentences in green; and **(E)** *polarity-aware*, additionally highlighting contradicting sentences in red. Highlights in both S and E are drawn from

²Pre-reg.: <https://aspredicted.org/vv3v2k.pdf>.

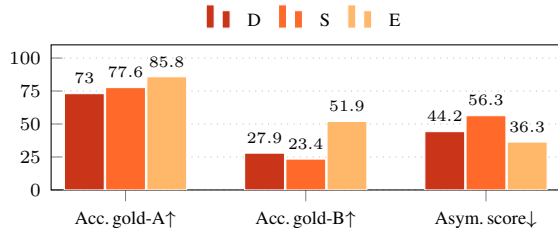


Figure 3: User-study results ($n=30$, within-subjects). S ’s collapsed gold-B accuracy paired with elevated gold-A accuracy shows signs of a false-completeness bias, which E eliminates while raising both accuracies.

gold annotations rather than from ETRACER predictions, isolating each paradigm from model quality and preventing prediction errors from obscuring the comparison. The 15 tasks are rotated across conditions via a 6-sequence Williams Latin square, yielding 10 participants per (group, condition) cell.

Task and Measures. Each task pairs a RAG response with its full cited context. Participants evaluate the response sentences against the source evidence rather than assessing ETRACER predictions. They click each response sentence, view the condition-appropriate evidence, and record a binary citation decision: (A) *include as-is* or (B) *do not include*. Gold-A sentences have only supporting evidence; gold-B sentences also contain contradicting evidence, where citing as-is would ignore the conflict. The 15 tasks comprise 35 sentences (15 gold-A, 20 gold-B), yielding 1,050 decisions across 30 participants and 3 conditions. We report *decision accuracy* by gold label and the *within-response asymmetry score* (gold-A minus gold-B accuracy): scores near 0 indicate symmetric judgment; large positive scores indicate accepting support at face value while missing conflict. The pre-registered primary contrast is S vs. E on gold-B accuracy.

Findings. Two patterns emerge (Figure 3; full statistics in Appendix D.4). (F1) *Polarity-aware grounding eliminates the bias and raises contradiction accuracy*: confirming our pre-registered primary hypothesis, E more than doubles gold-B accuracy over S (51.9% vs. 23.4%, +28.5 pp, $d=0.82$, $p=.0003$) and lowers the asymmetry score to 36.3; gold-A accuracy is also higher under E (85.8%). The $S \rightarrow E$ gap meets our pre-registered threshold ($d \geq 0.8$) and survives Holm-Bonferroni correction (Holm-corrected $p=.0029$). (F2) *Support-only grounding shows signs of a false-completeness bias*: S yields numerically lower gold-B accuracy than D (23.4% vs. 27.9%), though the effect is small ($d=0.16$, $p=0.135$) and

our sample is underpowered to confirm it; support-only highlights may signal sufficiency, leading participants to overlook unhighlighted conflicts, and the asymmetry score under S (56.3) exceeds that under D (44.2).

9 Conclusion

We presented ETRACER, a post-hoc framework for polarity-aware evidence grounding that jointly localizes sentence-to-sentence evidence and classifies its polarity in a single forward pass. Its context-aware signed scorer exploits a transformer encoder’s self-attention, reading the full $n \times m$ signed evidence map off the boundary-token interactions the encoder already computes. This single-pass design surpasses a matched pair-wise baseline by +19.16 Contradict F1 while matching strong LLM baselines at roughly 1/600 their scale. A pre-registered user study shows that polarity-aware grounding significantly improves contradiction identification (Cohen’s $d=0.82$), and further suggests a *false-completeness bias* under which support-only grounding fails to improve over the no-citation baseline. We release ETRACER-BENCH, with continued maintenance, to support future work on grounding that is both fine-grained and polarity-aware.

Limitations

Context length. ETRACER is built on ModernBERT-large, whose native 8,192-token window accommodates every ETRACER-BENCH instance without truncation. Since individual retrieved documents rarely exceed this budget, the limit is reached only when the *number* of retrieved documents grows large. For such cases, ETRACER admits a natural document-level sliding-window extension: partition the documents into disjoint groups that each fit the window, concatenate every group with the full response as an independent input, and merge the resulting signed maps along the context axis into a complete $n \times m$ map. Partitioning at document boundaries preserves each document’s semantic integrity and keeps context-aware scoring intact within each window, at the cost of cross-window dependencies. We validate this extension in Appendix C.8: at $T=2,048$ only 3.6% of test instances require multiple windows, with pooled Citation F1 unchanged from the native window (within run-to-run noise); even at the aggressive $T=1,024$ (25.0% triggered), pooled Ci-

tation F1 stays within 0.8 points of the native result. A systematic study of long-context strategies (e.g., hierarchical scoring or retrieval-augmented chunking) is left to future work.

Naturally occurring contradictions. ETRACER-BENCH contains 94 naturally occurring sentence-to-sentence contradicting pairs, split 59/11/24 across train/dev/test. The test subset suffices for a statistically separable comparison among methods (§7.2, Appendix C.6), but it is modest in size and skews toward single-sentence conflicts; the multi-document and cross-sentence cases that most directly probe context-aware reasoning are underrepresented—though the architectural advantage we report rests on the controlled M_{pair} comparison (§7.5), independent of this subset’s size. Two directions follow: scaling natural-contradiction annotation by systematically retrieving genuinely conflicting sources in domains such as evidence-based medicine, with deliberate emphasis on cross-sentence and multi-document conflicts; and extending the schema to pragmatic contradiction types such as deprecation, supersession, and evidence-strength divergence.

User study scale. Our within-subjects study ($n=30$) is powered to detect large effects (Cohen’s $d \geq 0.8$) at $\alpha=0.05$ after Holm–Bonferroni correction, matching the primary $S \rightarrow E$ contrast in §8. Smaller effects, such as the D vs. S difference suggestive of a false-completeness bias ($d=0.16$, $p=0.135$) or domain- and task-type-specific differences between conditions, would require substantially larger samples to confirm and are left to follow-up work.

Training variability. The main result tables report the run with seed 42 for consistency with the original evaluation. We additionally train ETRACER and $\mathcal{M}_{\text{pair}}$ with seeds 52 and 62; the context-aware gain remains positive across all three seeds (Table 4). Although this analysis shows that the main result is not specific to seed 42, three runs provide only a limited estimate of training variability.

Ethics Statement

Annotation. Our work re-annotates LFRQA, a publicly released corpus distributed under the Apache License 2.0. The BIOASQ subset is derived from publicly available biomedical literature rather than patient records and contains no personally identifiable information. Dataset annotation was conducted by an in-house team compensated

at the authors’ institution’s standard hourly rate for student research assistants (§A.2).

User study. The user study recruited participants through Prolific. Participants provided informed consent, could withdraw at any time without penalty, and received £10.00 for a median completion time of approximately 38 minutes (£15.85 per hour). We collected only task responses and completion times; no names, contact details, or other directly identifying information were collected. Study records were pseudonymized using Prolific identifiers, with recruitment and payment handled by the platform. Under our institution’s policy, this minimal-risk study involving judgments of publicly derived text did not require formal ethics or IRB review.

Acknowledgments

This work was supported by WisPerMed (AI for Personalized Medicine), funded by the German Research Foundation (DFG) under Research Training Group 2535. Generative AI tools were used for language editing, proofreading, and checklist preparation. All scientific claims, analyses, citations, and final text were reviewed and verified by the authors.

References

- Bohao Chu, Hendrik Damm, Tabea M. G. Pakull, Sameh Frihat, Georg Lodde, Elisabeth Livingstone, Christoph M. Friedrich, and Norbert Fuhr. 2026a. [MAPLE: Medical Aspect-Based Summarization with Phrase-Level Evidence](#). *arXiv preprint arXiv:2601.03418*.
- Bohao Chu, Hendrik Damm, Qianli Wang, Hui Wang, Shuning Zhang, Christoph M. Friedrich, and Norbert Fuhr. 2026b. [Locally Sound, Globally Insufficient: The Local-Global Gap in Multi-Hop Reasoning](#). *arXiv preprint arXiv:2609.32496*.
- Bohao Chu, Meijie Li, Sameh Frihat, Chengyu Gu, Georg Lodde, Elisabeth Livingstone, and Norbert Fuhr. 2025. [TracSum: A new benchmark for aspect-based summarization with sentence-level traceability in medical domain](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 844–864, Suzhou, China. Association for Computational Linguistics.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. [Multi-news: A large-scale](#)

- multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. **Enabling large language models to generate text with citations**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.
- Anisha Gunjal and Greg Durrett. 2024. **Molecular facts: Desiderata for decontextualization in LLM fact verification**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3751–3768, Miami, Florida, USA. Association for Computational Linguistics.
- Deepak Gupta, Dina Demner-Fushman, William Hersh, Steven Bedrick, and Kirk Roberts. 2024. **Overview of trec 2024 biomedical generative retrieval (biogen) track**. *arXiv preprint arXiv:2411.18069*.
- Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jinyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. **RAG-QA arena: Evaluating domain robustness for long-form retrieval augmented question answering**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4354–4374, Miami, Florida, USA. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. **Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing**. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*.
- Or Honovich, Roei Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. **TRUE: Re-evaluating factual consistency evaluation**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. **Gpt-4o system card**. *arXiv preprint arXiv:2410.21276*.
- Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. 2023. **Hagrid: A human-llm collaborative dataset for generative information-seeking with attribution**. *arXiv preprint arXiv:2307.16883*.
- Philipp Koehn. 2004. **Statistical significance tests for machine translation evaluation**. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. **SummaC: Re-visiting NLI-based models for inconsistency detection in summarization**. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. **BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Ilya Loshchilov and Frank Hutter. 2019. **Decoupled weight decay regularization**. In *International Conference on Learning Representations*.
- Microsoft. 2026. Microsoft copilot. <https://copilot.microsoft.com>. Accessed: 2026-04-20.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. **FActScore: Fine-grained atomic evaluation of factual precision in long form text generation**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.
- Perplexity AI. 2026. Perplexity ai. <https://www.perplexity.ai>. Accessed: 2026-04-20.
- David L. Sackett, William M. C. Rosenberg, J. A. Muir Gray, R. Brian Haynes, and W. Scott Richardson. 1996. **Evidence based medicine: what it is and what it isn't**. *BMJ*, 312(7023):71–72.
- Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. **Attribute first, then generate: Locally-attributable grounded text generation**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3344, Bangkok, Thailand. Association for Computational Linguistics.
- 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, Miami, Florida, USA. Association for Computational Linguistics.
- Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. **A closer look at claim decomposition**. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
- Edwin B Wilson. 1927. [Probable inference, the law of succession, and statistical inference](#). *Journal of the American Statistical Association*, 22(158):209–212.
- Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelim Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. 2024. [DocLens: Multi-aspect fine-grained evaluation for medical text generation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–679, Bangkok, Thailand. Association for Computational Linguistics.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. [Qwen3 technical report](#). *arXiv preprint arXiv:2505.09388*.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating factual consistency with a unified alignment function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.
- Jiajie Zhang, Yushi Bai, Xin Lv, Wanjuan Gu, Danqing Liu, Minhao Zou, Shulin Cao, Lei Hou, Yuxiao Dong, Ling Feng, and Juanzi Li. 2025. [LongCite: Enabling LLMs to generate fine-grained citations in long-context QA](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5098–5122, Vienna, Austria. Association for Computational Linguistics.

Appendix

A	Dataset Construction Details	14
A.1	Base Corpus	14
A.2	Annotation Team	14
A.3	Biomedical Domain Support	14
A.4	Natural Annotation Stage	15
A.5	Dataset Split	15
A.6	Adversarial Construction Stage	15
A.7	Dataset Statistics	17
B	Implementation Details	18
B.1	Shared Data Pipeline	18
B.2	Input Construction	18
B.3	Optimization	18
B.4	Document-Order Data Augmentation	19
B.5	Reproducibility	19
B.6	Off-the-Shelf Baselines	20
B.7	Inference Cost	20
C	Supplementary Experiments	20
C.1	Main Results with Confidence Intervals	20
C.2	Per-Domain Grounding Results	20
C.3	Class-Weighting for Context-Aware	21
C.4	Class-Weighting for Pair-wise Baselines	22
C.5	Context Shuffling: Per-Seed Results	23
C.6	Analysis on Naturally Occurring Contradictions	23
C.7	Document-Order Augmentation Ablation	24
C.8	Context-Window Scaling via Sliding Windows	24
D	User Study Details	25
D.1	Protocol and Participants	25
D.2	Task Materials	26
D.3	Measures	26
D.4	Full Statistical Results	27
E	Instructions and Demonstration	29
E.1	Prompt for Zero-Shot	29
E.2	Prompt for Seven-Shot	30

A Dataset Construction Details

ETRACER-BENCH is constructed in two annotation stages on top of a publicly released long-form RAG QA corpus. In the first **natural annotation** stage (§A.4), for each response sentence, annotators identify the supporting and contradicting context sentences; all remaining context sentences are automatically labeled Irrelevant. Because naturally occurring contradictions are rare (only 94 contradicting pairs vs. 4,565 supporting pairs), a second **adversarial construction** stage (§A.6) augments the corpus with human-written contradictions at two difficulty levels (explicit and implicit).

A.1 Base Corpus

The base corpus is LFRQA (Han et al., 2024), a long-form retrieval-augmented QA corpus of human-written answers spanning seven domains: BioASQ (biomedical), FiQA (finance), and five LoTTE domains (Lifestyle, Recreation, Science, Technology, Writing). We adopt LFRQA because (i) its answers are human-authored rather than LLM-generated, providing a high-quality substrate for sentence-level re-annotation; (ii) each answer is paired with curated source documents, enabling grounded evidence annotation; and (iii) its domain coverage spans both general-purpose and specialized content. We randomly sample 100 instances per domain, yielding 700 instances allocated 70/10/20 across train/dev/test (§A.5).

Licensing. LFRQA is released under Apache License 2.0, which permits derivative works with preserved attribution. We release ETRACER-BENCH under the same license, retain the original LFRQA attribution, and document modifications in the accompanying NOTICE file. Upstream retrieval corpora (LoTTE, FiQA, BioASQ) retain their original licenses; we redistribute only the document subsets referenced by LFRQA answers, consistent with its distribution terms.

Maintenance plan. ETRACER-BENCH will be hosted on a public repository (e.g., HuggingFace and GitHub) with version-controlled releases. We commit to long-term maintenance, including responding to erratum reports, incorporating community feedback, and releasing updates as additional annotations or domains become available.

Task redefinition. LFRQA’s native task is response-level pairwise preference; ETRACER-BENCH redefines the evaluation unit at sentence-

to-sentence granularity, adding a fresh annotation layer over the original data: sentence segmentation, per-sentence polarity labels (§A.4), and adversarial contradiction documents on a subset of instances (§A.6). No original content is modified or removed.

A.2 Annotation Team

Annotations are produced by an in-house team of five NLP graduate students from the authors’ institution, all fluent in English (the language of both the corpus and the guidelines). Annotators are compensated at the institution’s standard hourly rate for student research assistants, and no personally identifiable information is collected or released. Stage-specific procedures, role assignments, and consent are detailed in §A.4 and §A.6.

Cross-stage avoidance. The same five annotators participate in both stages. Beyond role disjointness within each stage, we enforce response-level cross-stage avoidance: no constructor produces contradictions for a response they previously labeled or adjudicated. Role rotation in the natural stage ensures that at least two of the five annotators remain unfamiliar with each response and thus eligible as constructors, making this feasible across all splits. This prevents familiarity gained during labeling from leaking into the style of constructed contradictions.

A.3 Biomedical Domain Support

For BIOASQ, we engage external collaborators from two institutions through cross-departmental agreements, serving both annotation stages.

Consultation during annotation. Two clinical experts from the Department of Dermatology at University Hospital Essen³ serve as domain advisors. Annotators and constructors encountering ambiguous biomedical content first consult authoritative sources (biomedical databases, clinical guidelines) or a large language model; unresolved cases are escalated to the experts for final adjudication.

Post-hoc verification. One medical student from the Annotation Lab Essen⁴ independently reviews a random 30% sample (30/100 instances) of the BIOASQ annotations along three dimensions: (i) completeness of the identified evidence, (ii) correctness of the polarity labels, and (iii) validity of the constructed adversarial contradictions. All 30 instances pass on all three dimensions. Verification

³<https://hautklinik.uk-essen.de/>

⁴<https://annotationlab.ikim.nrw/>

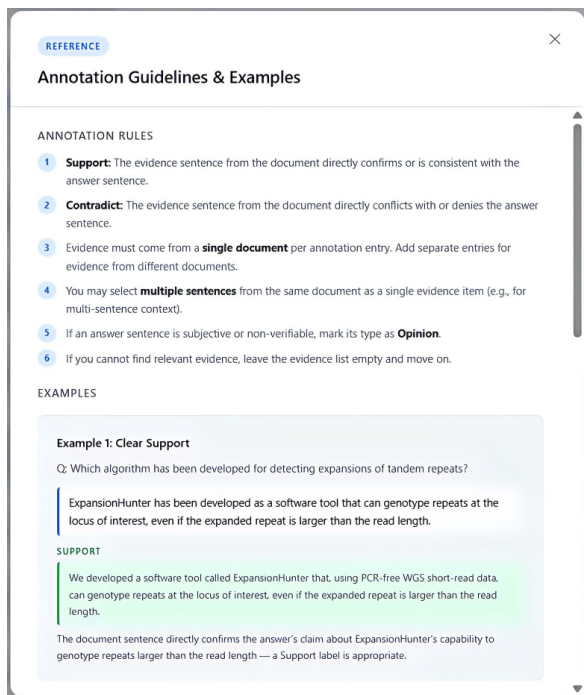


Figure 4: Annotation guidelines shown in the interface, covering label definitions, boundary cases, and worked examples for each domain.

is double-blind, with the verifier shown neither the original annotators’ identities nor their labels.

A.4 Natural Annotation Stage

Procedure and role assignment. For each instance, two annotators independently label the context sentences, and a third resolves disagreements, yielding three distinct team members per instance. The team receives written guidelines and completes a 10-sample pilot round with iterative feedback before beginning production annotation.

Annotation interface and consent. We develop a custom web-based interface for this stage. Before starting, each annotator views a consent page describing Apache License 2.0 release terms and the right to withdraw, and must enter an identifier and tick an acknowledgment checkbox to proceed. In the main view (Figure 5), the question and answer sentences appear in the left panel and source documents with per-sentence indexing on the right. For each answer sentence, the annotator selects candidate context sentences and tags them as supporting (green) or contradicting (red) via single-click buttons; unselected sentences default to irrelevant. Guidelines (Figure 4) cover label definitions, boundary cases, and per-domain worked examples, and are accessible at any time without leaving the current item.

Split	Items	Per-domain	Purpose
Train	490	70	Scorer training
Dev	70	10	Model selection, tuning, few-shot
Test	140	20	Final evaluation

Table 10: ETRACER-BENCH split sizes (70/10/20 per-domain ratio). **Items:** number of response-level instances, distinct from the response *sentences* counted in Table 2. Splits are stratified by domain; each response and its documents appear in exactly one split.

Throughput. Annotating a single instance takes 5–8 minutes on average, depending on context length, covering reading the response, inspecting candidate context sentences, and assigning labels for every response sentence; adjudication is excluded.

Inter-annotator agreement. Cohen’s κ (Cohen, 1960) is computed on the raw independent annotations *prior to* adjudication, providing an unbiased reliability estimate. We obtain $\kappa=0.87$ on evidence identification (any support or contradict label vs. none) and $\kappa=0.92$ on polarity classification (support vs. contradict, on sentences labeled as evidence by both annotators), both indicating strong agreement. The 7.3% of disagreements are resolved by a third adjudicator, whose decisions constitute the final gold labels.

A.5 Dataset Split

Table 10 reports the exact per-split allocation. Responses are split at the response level, stratified by domain; each response and its source documents therefore appear in exactly one split, preventing context leakage.

Treatment of natural contradictions. The 94 naturally occurring contradicting pairs follow the response-level random split, distributing as 59/11/24 across train/dev/test; given their small count, we do not stratify further on this axis. Natural and adversarial contradictions share the same label schema (polarity -1) and are treated identically in training and evaluation: the model receives no indicator distinguishing the two sources, and metrics in §6.2 pool them together. Adversarial contradictions (§A.6) are constructed *after* the split, with split-specific constructors assigned under the constraints in §A.2.

A.6 Adversarial Construction Stage

This stage produces 1,205 explicit and 2,041 implicit contradicting pairs following the protocol below.

Procedure and role assignment. Three disjoint

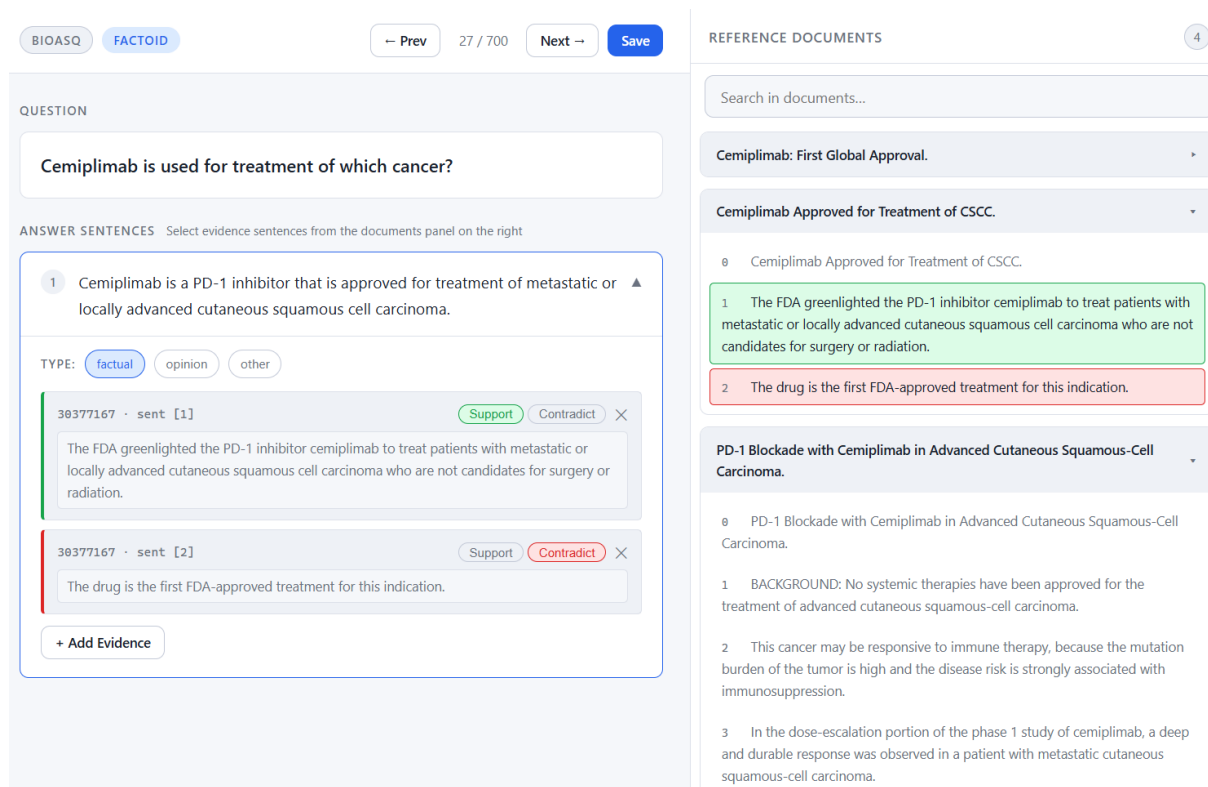


Figure 5: Main annotation view of the ETRACER-BENCH annotation interface. The left panel displays the question and answer sentences, with supporting (green) and contradicting (red) evidence attached via single-click buttons; the right panel lists the reference documents, with selected evidence sentences highlighted in the corresponding color.

constructors from the team (§A.2) produce contradictions for the train, dev, and test splits respectively, and a fourth member verifies all constructed instances. This stage does not use the web interface: response sentences are distributed as plain-text task sheets, and written contradictions are collected and verified offline. Constructors receive written instructions on task scope and difficulty levels, and provide written consent at the same time.

Selection criterion. We construct contradictions only for response sentences containing *falsifiable factual claims*, such as quantities, causal relations, or temporal and comparative judgments. Sentences expressing subjective preferences or definitional statements (e.g., “A hash function maps inputs to fixed-size outputs”) are skipped, since inverting them yields a different opinion or definition rather than a contradiction.

Construction procedure. For each selected response sentence, the assigned constructor writes a contradicting paragraph of 1–4 sentences at one of two difficulty levels (assigned at random):

- **Explicit** contradictions use surface-level alterations: negation, antonym substitution, or factual replacement (e.g., “increases” → “decreases”; “30%” → “70%”).

- **Implicit** contradictions introduce logically incompatible statements that preserve the topic but require multi-step inference (e.g., “Drug X reduces mortality” is contradicted by “Patients receiving Drug X in the 2023 trial showed no survival benefit over placebo at 5-year follow-up”).

Each constructed paragraph is inserted into a *new* document appended to the source context, rather than edited into an existing one. This preserves the original documents as-is and simulates realistic RAG scenarios where retrieval returns conflicting sources from different origins. The entire context is then shuffled with a fixed seed (seed=42) to eliminate positional bias and prevent models from exploiting the append order. Contradicting sentences are annotated with polarity -1 for the target response sentence.

Throughput. Constructing a contradicting paragraph takes 4–8 minutes per eligible response sentence, covering reading the sentence, drafting the paragraph, and labeling the difficulty level. This exceeds the natural annotation cost.

Verification. The verifier inspects each constructed instance and judges (i) whether the paragraph genuinely contradicts the target response sentence and

Domain	Total	Pre-augmentation				Post-augmentation			
		π_F	π_C	π_A	π_U	π_F	π_C	π_A	π_U
BioASQ	163	159 (97.5%)	1 (0.6%)	0 (0.0%)	3 (1.8%)	3 (1.8%)	4 (2.5%)	156 (95.7%)	0 (0.0%)
FiQA	285	276 (96.8%)	0 (0.0%)	9 (3.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	285 (100.0%)	0 (0.0%)
Lifestyle	315	305 (96.8%)	0 (0.0%)	10 (3.2%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	315 (100.0%)	0 (0.0%)
Recreation	214	206 (96.3%)	0 (0.0%)	8 (3.7%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	214 (100.0%)	0 (0.0%)
Science	233	225 (96.6%)	1 (0.4%)	7 (3.0%)	0 (0.0%)	4 (1.7%)	1 (0.4%)	228 (97.9%)	0 (0.0%)
Technology	266	260 (97.7%)	0 (0.0%)	6 (2.3%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	266 (100.0%)	0 (0.0%)
Writing	260	253 (97.3%)	0 (0.0%)	6 (2.3%)	1 (0.4%)	6 (2.3%)	1 (0.4%)	253 (97.3%)	0 (0.0%)
Total	1,736	1,684 (97.0%)	2 (0.1%)	46 (2.6%)	4 (0.2%)	13 (0.7%)	6 (0.3%)	1,717 (98.9%)	0 (0.0%)

Table 11: Response-level faithfulness profile of ETRACER-BENCH by domain, under two views. The *pre-augmentation* view reflects the corpus labeled by our natural annotation stage (§A.4), before any adversarial modification. The *post-augmentation* view reflects the corpus after injecting adversarial contradiction documents (§A.6). Each cell reports the class count with its corpus share π in parentheses; $\pi_F/\pi_C/\pi_A/\pi_U$ correspond to FAITHFUL/CONTRADICTED/AMBIGUOUS/UNGROUNDED (§A.7). The shift from FAITHFUL-dominated (pre) to AMBIGUOUS-dominated (post) reflects preserving original supporting documents rather than overwriting them, simulating retrieval returning conflicting sources.

(ii) whether the difficulty level matches the intended assignment. First-pass agreement with the constructor’s intent is 95.2% for explicit and 91.4% for implicit contradictions; given the high acceptance rate, we report percent agreement rather than chance-adjusted measures. Flagged instances are returned for revision until both dimensions are confirmed, so all released instances pass final verification. The final corpus distribution by domain is reported in Table 2.

A.7 Dataset Statistics

Table 12 summarizes length and label-distribution statistics of ETRACER-BENCH, complementing the per-domain pair counts in Table 2. The corpus comprises 51,206 labeled response–context sentence pairs across 700 responses, of which 7,905 carry an evidential (support or contradict) label and the rest are irrelevant.

Length distributions. Responses contain a median of 2 sentences (range 1–9), and source contexts after adversarial augmentation a median of 17 (range 2–171). When the full response and context are concatenated into a single input sequence (§5.1), every instance fits within ModernBERT-large’s 8,192-token window (median 548 tokens, maximum 5,621), so no truncation is applied.

Label distribution. Across all $n \times m$ response–context sentence pairs, supporting pairs constitute 8.9%, contradicting pairs (natural and adversarial combined) 6.5%, and irrelevant pairs 84.6% (percentages may not sum to 100 due to rounding).

Response-level faithfulness profile. The signed evidence map (§3.1) categorizes each response sentence by the polarity of its evidence into

Statistic	Median Mean		Range
<i>Per response</i>			
Response sentences	2	2.5	1–9
Context sentences	17	24.0	2–171
Supporting evidence	5	6.5	0–69
Contradicting evidence	4	4.8	0–17
<i>Label shares (over 51,206 response–context sentence pairs)</i>			
Support			8.9%
Contradict			6.5%
Irrelevant			84.6%
<i>Input sequence length (tokens, ModernBERT-large tokenizer)</i>			
Tokens per instance	548	729	72–5,621

Table 12: Length and label-distribution statistics of ETRACER-BENCH (700 responses across seven domains; 51,206 labeled pairs).

four classes, with corpus shares π_F (FAITHFUL, only supporting), π_C (CONTRADICTED, only contradicting), π_A (AMBIGUOUS, both), and π_U (UNGROUNDED, neither). Table 11 reports the distribution under two views. The *pre-augmentation* view, after the natural annotation stage (§A.4) but before contradiction injection, is overwhelmingly FAITHFUL (97.0%), with a small AMBIGUOUS tail (2.6%, 46 sentences) where naturally co-occurring supporting and contradicting evidence coexists, and only 2 CONTRADICTED and 4 UNGROUNDED sentences in total, reflecting the high factual quality of the underlying human-written answers. The 94 naturally occurring contradicting pairs in Table 2 are hosted by these 48 sentences (2 CONTRADICTED + 46 AMBIGUOUS), since a single response sentence can be contradicted by multiple context sentences. The *post-augmentation* view, after injecting adversarial contradiction documents (§A.6), shifts almost entirely to AMBIGUOUS (98.9%): each adversarial paragraph is appended as a new document without removing the original supporting evidence, so most response sentences end up with evidence

Baseline	Checkpoint / Snapshot	Params	Access
DeBERTa-v3 (MNLI)	MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli	435M	HF
BART-large (MNLI)	facebook/bart-large-mnli	406M	HF
DeBERTa-v3 (ETB)	microsoft/deberta-v3-large (fine-tuned)	435M	HF
BART-large (ETB)	facebook/bart-large (fine-tuned)	406M	HF
$\mathcal{M}_{con}$ (ETRACER)	answerdotai/ModernBERT-large (fine-tuned)	395M	HF
$\mathcal{M}_{pair}$	answerdotai/ModernBERT-large (fine-tuned)	395M	HF
GPT-4o	gpt-4o-2024-08-06	—	OpenAI API
Qwen3-235B	qwen/qwen3-235b-a22b-2507	235B	OpenRouter API

Table 13: Baseline model sources. HF: Hugging Face checkpoints. Parameter counts include embeddings. Fine-tuned variants use the listed checkpoint as initialization and are trained on the ETRACER-BENCH (ETB) train split.

of both polarities. This bimodal structure is intentional but does not trivialize the task: our primary metrics (Support F1 and Contradict F1, §6.2) operate at the sentence-to-sentence level, so a model cannot win by predicting AMBIGUOUS. It must still decide, for each context sentence, whether it supports or contradicts a given response sentence. The per-domain breakdown in Table 11 confirms this pattern holds uniformly across all seven domains.

B Implementation Details

This appendix details the training and inference configuration of the models in §7. The four fine-tuned models are our context-aware $\mathcal{M}_{con}$ (ETRACER) and three pair-wise baselines trained on identical data: $\mathcal{M}_{pair}$ (ModernBERT-large), DeBERTa-v3-large, and BART-large. We additionally evaluate two off-the-shelf MNLI checkpoints and two LLM baselines.

B.1 Shared Data Pipeline

All four fine-tuned models are trained from the same canonical form derived from the ETRACER-BENCH train split, differing only in how each response–context instance is presented. A single preprocessing pass converts every annotation item into the canonical tuple $(\mathcal{R}, S, \mathbf{Y})$: the response $\mathcal{R}=[r_1, \dots, r_n]$, the context $S=[s_1, \dots, s_m]$ obtained by concatenating sentences across all cited documents in retrieval order, and the polarity matrix $\mathbf{Y} \in \{+1, -1, 0\}^{n \times m}$. The context-aware model consumes each instance whole, while pair-wise baselines decompose it into $n \times m$ independent examples $\{(r_i, s_j, \mathbf{Y}_{ij})\}$.

B.2 Input Construction

Table 14 summarizes the input format, maximum sequence length, and scoring head for each fine-tuned model. The three pair-wise baselines cover three distinct input conventions: ModernBERT-style with explicit boundary tokens, standard NLI

encoding, and BART’s native pair format, ensuring our comparison does not conflate context-awareness with encoder idiosyncrasy.

Context-aware $\mathcal{M}_{con}$. The full response and context are packed into one sequence with [R_SEP] and [S_SEP] marking sentence boundaries (§5.1). If the sequence exceeds 8,192 tokens, it is truncated at the last complete [S_SEP] so that no partial context sentence is scored; in practice, every instance fits within the window without truncation. A single forward pass yields hidden states at all boundary positions, from which the $n \times m$ pair scores are gathered in one shot.

Pair-wise $\mathcal{M}_{pair}$. Shares ETRACER’s backbone, special tokens, and classification head, but encodes one response sentence and one context sentence per forward pass. This is the controlled ablation of §7.1: any gap versus $\mathcal{M}_{con}$ is attributable solely to the loss of cross-sentence attention, since tokenizer, vocabulary, head, and training signal are otherwise identical.

DeBERTa-v3-large. Uses the standard NLI encoding and predicts from the [CLS] hidden state, the conventional format for fine-tuning DeBERTa-v3 on sentence-pair classification.

BART-large. Uses BART’s native sequence-pair format with classification on the decoder hidden state at the final </s>, as prescribed by BartForSequenceClassification, avoiding distribution shift from non-native encodings.

B.3 Optimization

All four fine-tuned models share the same optimization protocol, differing only in per-device batch size to accommodate sequence-length differences.

Optimizer and schedule. Models are optimized with AdamW (Loshchilov and Hutter, 2019) at peak learning rate 2×10^{-5} , linear warmup over the first 3% of steps followed by linear decay to zero, weight decay 0.01, gradient clipping at ℓ_2 -norm 1.0, dropout 0.15 on the classification head,

Model	Input sequence (per forward pass)	Max. len.	Head input
<i>Context-aware (single pass per response)</i>			
$\mathcal{M}_{con}$ (ModernBERT-large)	[CLS] r_1 [R_SEP] ... [R_SEP] [SEP] s_1 [S_SEP] ... [S_SEP]	8,192	$[h_i^{[R_SEP]}, h_j^{[S_SEP]}]$
<i>Pair-wise (one pass per (r_i, s_j) pair)</i>			
$\mathcal{M}_{pair}$ (ModernBERT-large)	[CLS] r_i [R_SEP] [SEP] s_j [S_SEP]	512	$[h^{[R_SEP]}, h^{[S_SEP]}]$
DeBERTa-v3-large	[CLS] r_i [SEP] s_j [SEP]	512	$h^{[CLS]}$
BART-large	<S> r_i </S> </S> s_j </S>	1,024	$h_{last}^{</S>}$

Table 14: Input construction for the four fine-tuned models. For $\mathcal{M}_{con}$, hidden states at every [R_SEP] and [S_SEP] form $n \times m$ pair representations in a single forward pass; pair-wise models encode each (r_i, s_j) independently. All four use a three-way head over {support, contradict, irrelevant}, trained from scratch rather than inherited from any MNLI checkpoint.

and bf16 mixed precision. Training runs up to 35 epochs with early stopping (patience 5) and final checkpoints selected by dev Citation F1.

Batch configuration. For $\mathcal{M}_{con}$, each instance is a full response–context pair jointly supervising up to $n \times m$ polarity predictions; we use a per-device batch size of 2 with 8-step gradient accumulation, yielding an effective batch size of 16 instances. Gradient checkpointing is enabled to fit the 8,192-token sequences within GPU memory. The three pair-wise models ($\mathcal{M}_{pair}$, DeBERTa-v3-large, BART-large) decompose each instance into $n \times m$ independent examples; we use a per-device batch size of 16 with the same 8-step accumulation, yielding an effective batch size of 128 pairs.

Class weights. All four models share the same class-weighted cross-entropy objective (§5.2) with weights $[w_{sup}, w_{irr}, w_{con}] = [1, 10, 1]$, selected on the dev split for $\mathcal{M}_{con}$ (§C.3) and applied uniformly to the pair-wise baselines; this never costs any baseline more than 0.8 dev Citation F1 relative to its own optimum (§C.4). Holding the loss fixed ensures that comparisons in Table 3 reflect input construction rather than class-imbalance treatment.

Hardware and training time. All four models are trained on a single NVIDIA A6000 GPU (48 GB). Wall-clock training time ranges from 1.9 hours ($\mathcal{M}_{pair}$) to 11.4 hours ($\mathcal{M}_{con}$), with BART-large at 4.2 and DeBERTa-v3-large at 9.4. $\mathcal{M}_{con}$ ’s longer time reflects its augmented training corpus (5,386 vs. 490 instances; Appendix B.4) rather than a per-step cost difference. Early stopping (patience 5) reliably triggers well before the 35-epoch cap.

B.4 Document-Order Data Augmentation

Procedure. We expand the 490-response train split to 5,386 instances ($\sim 11 \times$) via two complementary augmentations: *document-order shuffling* and *document-type filtering*. Order shuffling generates eight permutations per response

with seeds {42, 52, 62, 72, 82, 92, 102, 112}, each preserving the response $\mathcal{R}$, retrieved documents, within-document sentence order, and polarity matrix $\mathbf{Y}$, differing only in document concatenation order; this yields 3,920 instances (490×8). Type filtering derives two additional views by partitioning documents by origin: a *natural-only* view retaining only the original LFRQA-retrieved documents (preserving naturally occurring evidence; 490 instances), and a *synthetic-only* view retaining only the adversarially constructed contradictions from §A.6 (486 instances; four train responses lack adversarial contradictions and are dropped). Together with the 490 original instances, the corpus totals $490 + 3,920 + 490 + 486 = 5,386$. Labels are re-indexed in each view so every (r_i, s_j) pair retains its original annotation relative to its filtered context. All permutations and views are precomputed offline as a static training corpus, used for all $\mathcal{M}_{con}$ results in the main paper.

Why pair-wise baselines cannot benefit. A pair-wise model consumes one (r_i, s_j) tuple per forward pass and is therefore invariant to the order of all other context sentences. Document-order shuffling produces multiple training instances containing the *identical* set of (r_i, s_j) pairs as the original, offering no new learning signal. Type-filter views likewise only restrict the pair set without exposing cross-pair structure. Empirically, training pair-wise baselines on the augmented corpus triggers early stopping several epochs earlier with no gain in dev Citation F1. We therefore report them on the non-augmented corpus, which matches the inputs they actually consume.

B.5 Reproducibility

All reported numbers come from fully deterministic training. We fix random seeds for Python’s random, NumPy, and PyTorch (CPU and all CUDA devices) to 42, enable cuDNN deterministic mode and PyTorch’s strict deterministic algorithm mode, disable

memory-efficient SDP attention kernels, set the cuBLAS workspace configuration to :4096:8 prior to PyTorch initialization, and seed the DataLoader with an explicit torch.Generator. Under this configuration, repeated runs with identical hyperparameters yield bit-identical per-step losses and final checkpoints on matching hardware (NVIDIA A6000, CUDA 12.8, PyTorch 2.9).

Seed sensitivity. The main result tables report the seed-42 run for all models. To assess training variability in the controlled architectural comparison, we additionally train $\mathcal{M}_{con}$ (ETRACER) and $\mathcal{M}_{pair}$ with seeds 52 and 62 using the same data-processing, optimization, and model-selection protocol. Across seeds {42, 52, 62}, the paired mean gains are +2.93 Support F1, +17.69 Contradict F1, and +10.31 Citation F1, and all three gains remain positive for every seed (Table 4). The context-aware advantage is therefore not specific to the original seed-42 run. We do not repeat the multi-seed analysis for the remaining baselines, and three seeds provide only a limited estimate of training variability. The paired bootstrap CIs in Table 15 separately quantify evaluation-sampling uncertainty for the main seed-42 results and should not be interpreted as estimates of training-seed variance.

B.6 Off-the-Shelf Baselines

MNLI baselines. DeBERTa-v3-large-MNLI and BART-large-MNLI are used as released on Hugging Face without further training. Their pretrained 3-way MNLI heads output {entailment, contradiction, neutral}, which we map to {support, contradict, irrelevant}. The mapping is imperfect: MNLI’s *neutral* is a coarser superset of our *irrelevant*, conflating topically-related-but-uninferable pairs with genuinely off-topic ones. This is the closest available correspondence for NLI-pretrained models, and the fine-tuned variants of the same backbones in Table 3 isolate the label-mismatch effect from the backbone effect.

LLM baselines. GPT-4o is accessed via the OpenAI API and Qwen3-235B-Instruct via OpenRouter, both with temperature 0 and max output length 4,096 tokens. Each model receives the full response and context in a single prompt and outputs structured sentence-level polarity labels; all test instances fit within both context windows. Few-shot prompts use 7 demonstrations, one per domain, drawn from the dev split; templates appear in Ap-

pendix E.

B.7 Inference Cost

We characterize inference cost by the number of forward passes per response, a hardware- and batching-agnostic measure that directly reflects the architectural difference between the two paradigms. For a response with n sentences and context with m sentences, $\mathcal{M}_{con}$ performs 1 forward pass to produce the full $n \times m$ evidence map, whereas each pair-wise baseline requires $n \times m$ passes. Across the test split, pair-wise scoring averages 73.1 passes per response ($\frac{1}{N} \sum_i n_i m_i$ over N test instances) versus 1 for $\mathcal{M}_{con}$, and this gap grows linearly with both response and context length. The per-instance mean exceeds the product of marginal means ($\bar{n} \cdot \bar{m} \approx 60$) because response and context length are positively correlated. We do not report wall-clock latency, as it depends on batching, hardware, and (for LLM baselines) external API conditions orthogonal to the architecture being compared.

C Supplementary Experiments

C.1 Main Results with Confidence Intervals

Table 15 reports the full main-results table with 95% paired bootstrap confidence intervals for every model and metric, complementing the point estimates in Table 3. Intervals are computed by resampling response sentences with 10,000 bootstrap replicates (§6.2), paired across models so that each replicate draws the same set of response sentences for all systems. We additionally report explicit and implicit Contradict F1 ($C.F1_e/C.F1_i$) with their own intervals. Three comparisons are separable at the CI level. (i) ETRACER’s Contradict F1 interval [87.31, 91.16] is disjoint from $\mathcal{M}_{pair}$ ’s [66.91, 72.96], confirming the contradiction-detection gain is not an artifact of evaluation sampling. (ii) On Support F1, by contrast, ETRACER and $\mathcal{M}_{pair}$ overlap ([73.03, 77.98] vs. [69.90, 74.74]), consistent with support being well-served by local cues. (iii) ETRACER and Qwen3-235B 7-shot have overlapping Citation F1 intervals ([80.78, 83.89] vs. [80.31, 83.31]), so the two are statistically on par despite the $\sim 600 \times$ parameter gap.

C.2 Per-Domain Grounding Results

Table 16 reports per-domain Support, Contradict, and Citation F1 for $\mathcal{M}_{con}$ (ETRACER) and the con-

Category	Method	S.P	S.R	S.F1	C.P	C.R	C.F1	Cit.F1	C.F1 _e	C.F1 _i
Prompt	GPT-4o (0-shot) [†]	88.60 [86.23,90.92]	68.96 [65.84,72.00]	77.56 [75.21,79.78]	<u>91.02</u> [87.89,93.84]	48.64 [44.98,52.51]	63.40 [59.93,66.80]	70.48 [68.34,72.61]	86.15 [82.50,89.21]	42.25 [36.67,47.36]
	GPT-4o (7-shot) [†]	<u>85.39</u> [82.85,88.00]	71.95 [68.99,74.97]	<u>78.10</u> [75.97,80.26]	87.31 [84.13,90.18]	<u>64.96</u> [61.27,68.85]	<u>74.50</u> [71.47,77.38]	<u>76.30</u> [74.42,78.07]	89.50 [86.55,92.20]	61.54 [56.89,66.10]
	Qwen3-235B (0-shot) [†]	79.18 [76.56,81.72]	72.95 [70.14,75.93]	75.94 [73.71,78.08]	92.42 [89.93,94.54]	70.24 [66.88,73.95]	79.82 [77.22,82.42]	77.88 [76.18,79.56]	91.10 [88.16,93.53]	70.99 [66.88,75.09]
	Qwen3-235B (7-shot) [†]	78.80 [76.06,81.84]	77.49 [74.74,80.17]	78.14 [76.12,80.24]	89.51 [86.96,91.94]	<u>81.92</u> [78.86,84.88]	<u>85.55</u> [83.26,87.58]	<u>81.85</u> [80.31,83.31]	<u>91.31</u> [88.65,93.79]	<u>81.48</u> [78.22,84.66]
	DeBERTa-v3 (MNLI) [‡]	66.03 [61.62,70.48]	34.48 [31.24,37.74]	45.30 [41.97,48.66]	52.15 [48.56,55.75]	69.92 [66.40,73.61]	59.74 [56.54,62.72]	52.52 [50.20,54.76]	67.26 [63.11,71.03]	53.30 [48.95,57.60]
Fine-tuned	BART-large (MNLI) [‡]	69.84 [65.51,74.38]	34.15 [30.92,37.54]	45.87 [42.51,49.35]	40.47 [37.44,43.46]	68.32 [64.63,72.21]	50.83 [47.83,53.72]	48.35 [46.05,50.54]	57.36 [53.41,61.69]	45.16 [40.68,49.22]
	DeBERTa-v3 (ETB)	71.51 [68.54,74.37]	<u>73.73</u> [70.53,76.71]	72.60 [70.19,74.81]	74.03 [70.35,77.61]	76.64 [73.12,80.00]	75.31 [72.28,78.00]	73.96 [72.04,75.78]	83.71 [80.23,87.07]	69.35 [65.13,72.97]
	BART-large (ETB)	74.27 [71.36,77.26]	70.40 [67.48,73.38]	72.28 [69.88,74.69]	71.75 [68.09,75.07]	77.60 [74.19,80.53]	74.56 [71.54,77.08]	73.42 [71.56,75.18]	83.83 [80.40,87.15]	68.14 [63.98,71.89]
	$\mathcal{M}_{pair}$ (ETB)	74.70 [71.92,77.68]	70.07 [66.89,73.10]	72.31 [69.90,74.74]	74.11 [70.25,77.80]	66.40 [62.93,70.02]	70.04 [66.91,72.96]	71.18 [69.18,73.15]	82.82 [79.27,86.34]	59.91 [55.29,64.17]
	$\mathcal{M}_{con}$ (ETRACER)	84.10 [81.08,86.68]	68.63 [65.60,71.59]	75.58 [73.03,77.98]	90.59 [88.19,92.89]	87.84 [85.34,90.59]	89.20 [87.31,91.16]	82.39 [80.78,83.89]	95.18 [93.17,97.10]	85.41 [82.48,88.14]

Table 15: Main grounding results with 95% paired bootstrap confidence intervals (10,000 replicates over response sentences, §6.2). Point estimates match Table 3. Gray brackets give the lower/upper interval bounds. S/C.: support/contradict; C.F1_e/C.F1_i: explicit/implicit Contradict F1. **Bold**: best, underline: second best (on point estimates). The Contradict F1 intervals of ETRACER and $\mathcal{M}_{pair}$ are disjoint, while their Support F1 intervals overlap, isolating context-aware scoring as the source of the contradiction-detection gain. ETRACER and Qwen3-235B 7-shot have overlapping Citation F1 intervals.

Metric	BioASQ		FiQA		Life.		Rec.		Sci.		Tech.		Writ.	
	$\mathcal{M}_{con}$	$\mathcal{M}_{pair}$	$\mathcal{M}_{con}$	$\mathcal{M}_{pair}$	$\mathcal{M}_{con}$	$\mathcal{M}_{pair}$	$\mathcal{M}_{con}$	$\mathcal{M}_{pair}$	$\mathcal{M}_{con}$	$\mathcal{M}_{pair}$	$\mathcal{M}_{con}$	$\mathcal{M}_{pair}$	$\mathcal{M}_{con}$	$\mathcal{M}_{pair}$
Support F1	75.3	72.4	72.9	73.5	79.6	73.9	77.0	75.5	74.3	72.1	75.1	72.0	75.1	65.4
Contradict F1	93.5	84.0	86.2	69.5	92.9	67.2	95.3	73.9	87.8	63.7	86.0	71.4	85.4	63.4
Citation F1	84.4	78.2	79.5	71.5	86.2	70.6	86.2	74.7	81.1	67.9	80.5	71.7	80.2	64.4

Table 16: Per-domain Support, Contradict, and Citation F1 on the ETRACER-BENCH test split. **Bold**: higher of the two per cell. Domain abbreviations follow Figure 2.

trolled pair-wise baseline $\mathcal{M}_{pair}$, complementing Figure 2. Three patterns emerge consistent with our main findings.

Contradict F1 is uniformly strong for ETRACER. ETRACER achieves Contradict F1 of at least 85 in every domain, peaking at 95.3 in Recreation and 93.5 in BioASQ. $\mathcal{M}_{pair}$ varies more widely (63.4 in Writing to 84.0 in BioASQ), with the per-domain gap reaching +25.7 in Lifestyle and +24.1 in Science. Context-aware scoring delivers its largest contradiction-detection gains in domains where polarity depends on cross-sentence comparison of mechanisms, quantities, or causal chains.

Support F1 varies by domain structure. ETRACER’s Support F1 ranges from 72.9 (FiQA) to 79.6 (Lifestyle), with FiQA the only domain where pair-wise scoring is marginally stronger (73.5, −0.6). Locally self-contained financial claims are often verifiable from a single context sentence, where pair-wise scoring is already competitive. In the remaining six domains, $\mathcal{M}_{con}$ leads $\mathcal{M}_{pair}$ by +1.5 to +9.7, with the largest margin

in Writing where supporting evidence frequently spans multiple sentences, echoing the ablation finding (§7.5) that context-aware scoring helps support detection most when evidence must be aggregated across sentences.

Citation F1 gains are concentrated on contradiction. In every domain, the Contradict F1 gap between $\mathcal{M}_{con}$ and $\mathcal{M}_{pair}$ substantially exceeds the Support F1 gap. Averaged across the seven domains, $\mathcal{M}_{con}$ exceeds $\mathcal{M}_{pair}$ by +19.1 on Contradict F1 but only +3.5 on Support F1 (cf. the macro-averaged +19.16/+3.27 in Table 3), a five-to-one ratio. ETRACER’s Citation F1 advantage is thus driven by contradiction detection across the board, not by any single domain.

C.3 Class-Weighting for Context-Aware

We ablate class weights for $\mathcal{M}_{con}$ across 10 configurations on the dev split (Table 17), spanning aggressive majority-class down-weighting (sklearn’s $w_c \propto 1/f_c$ heuristic), flatter and uniform weights, and progressive majority-class up-weighting. The

configuration was selected by grid search on dev Cit.F1; we report the full sweep here for transparency and to characterize the loss surface.

(i) Down-weighting the majority class is harmful.

The sklearn balanced heuristic (row a) compresses w_{irr} to 0.40 and drops Cit.F1 by 3.5 points below uniform (row d, 78.4), driven by a collapse in Support Precision: under-weighting the irrelevant class biases the model toward predicting support or contradict on ambiguous pairs, inflating false positives. The Sqrt variant (row b, 78.6) sits closer to uniform, confirming the monotone trend.

(ii) Up-weighting the majority class within a window yields the best result.

Cit.F1 peaks at 79.6 with $[1, 10, 1]$ (row i), which we adopt. The trajectory across the irr-boost spectrum is non-monotonic, with two near-tied peaks (rows f and i) bracketing a modest dip at row h; Support F1 follows a similar shape while Contradict F1 stays within a narrow ± 1 band throughout. Up-weighting therefore primarily improves Support without sacrificing Contradict, but excessive up-weighting (beyond row i) yields no further gains.

(iii) Why the counterintuitive direction works.

The asymmetry, that majority up-weighting helps while down-weighting hurts, is opposite to the conventional remedy for imbalanced classification, where the minority class is up-weighted. We initially expected the reverse and let the dev sweep dictate the choice. The empirical pattern is interpretable post-hoc: with an 84% irrelevant prior, the dominant gradient direction at training initialization already pushes toward correct irrelevant predictions, so additional minority up-weighting risks pushing the model into false-positive territory on the much larger irrelevant population. Under context-aware scoring, the bottleneck is reliable *non-evidence rejection*, identifying which context sentences are *not* evidence for a given response sentence, since precise evidence localization depends on first eliminating distractors. Up-weighting w_{irr} sharpens this rejection boundary at minimal cost to minority recall, because the minority gradient signal is already strong on the small number of true evidence pairs. We do not claim this generalizes to all class-imbalanced tasks; the result is specific to a localization-style objective on a long-tailed three-way label space.

(iv) Contradict F1 is architecture-bound, not weight-bound.

Across all 10 configurations, Con-

Config	Weights	S.F1	C.F1	Cit.F1
(a) sklearn	[4.92, 0.40, 3.72]	71.4	78.4	74.9
(b) Sqrt	[1.85, 0.48, 1.62]	72.9	84.2	78.6
(c) Flatter	[1.5, 1.2, 1.5]	73.8	83.5	78.7
(d) Uniform	[1.0, 1.0, 1.0]	72.2	84.6	78.4
(e) Irr-boost-2	[1.0, 2.0, 1.0]	72.7	84.6	78.6
(f) Irr-boost-4	[1.0, 4.0, 1.0]	74.0	85.2	79.5
(g) Irr-boost-6	[1.0, 6.0, 1.0]	74.3	84.1	79.2
(h) Irr-boost-8	[1.0, 8.0, 1.0]	70.7	83.7	77.0
(i) Irr-boost-10 (Adopt)	[1.0, 10.0, 1.0]	74.7	84.5	79.6
(j) Irr-boost-12	[1.0, 12.0, 1.0]	73.8	83.8	78.8

Table 17: Class-weighting ablation for $\mathcal{M}_{\text{con}}$ on the ETRACER-BENCH dev split. Weights are ordered $[w_{\text{sup}}, w_{\text{irr}}, w_{\text{con}}]$. Configurations span aggressive majority-class down-weighting (a–b), flatter and uniform baselines (c–d), and progressive irrelevant-class up-weighting (e–j). **Bold**: adopted configuration, which achieves the highest Cit.F1 alongside the best S.F1.

Model	Weights	S.F1	C.F1	Cit.F1	Δ
DeBERTa-v3	[1, 1, 1]	77.5	70.6	74.0	—
	[1, 10, 1] (adopted)	76.9	74.9	75.9	+1.9
$\mathcal{M}_{\text{pair}}$	[1, 1, 1]	73.9	66.2	70.1	—
	[1, 10, 1] (adopted)	71.3	67.4	69.3	−0.8
BART-large	[1, 1, 1]	72.0	65.9	69.0	—
	[1, 10, 1] (adopted)	72.6	66.0	69.3	+0.3

Table 18: Class-weighting comparison for the three pair-wise baselines on the ETRACER-BENCH dev split. Weights are ordered $[w_{\text{sup}}, w_{\text{irr}}, w_{\text{con}}]$. Δ : change in Cit.F1 relative to uniform weighting. Main-paper results in Table 3 use the adopted $[1, 10, 1]$ configuration.

tradict F1 stays within a 6.8-point band (78.4–85.2), with 9 of 10 above 83.0. Contradiction detection is therefore primarily enabled by context-aware scoring rather than by class reweighting, consistent with the main-results finding (§7.1) that architecture drives the Contradict F1 gap over the pair-wise baseline.

C.4 Class-Weighting for Pair-wise Baselines

To ensure a fair comparison in Table 3, the three pair-wise baselines ($\mathcal{M}_{\text{pair}}$, DeBERTa-v3-large, BART-large) are fine-tuned with the same class-weighting adopted for $\mathcal{M}_{\text{con}}$, $[1, 10, 1]$ (§C.3). To verify that this choice does not systematically disadvantage them relative to uniform weighting, we additionally fine-tune each baseline under $[1, 1, 1]$, keeping all other settings unchanged, and compare dev Cit.F1 (Table 18). DeBERTa-v3-large gains +1.9 Cit.F1 from majority up-weighting, mirroring the trend observed for $\mathcal{M}_{\text{con}}$. BART-large also narrowly prefers the adopted configuration (+0.3), suggesting it is largely insensitive to class weighting. $\mathcal{M}_{\text{pair}}$ is the only baseline where uniform weighting is mildly preferred (−0.8), though still

within 0.8 points of its own optimum. Across all three baselines, the adopted $[1, 10, 1]$ configuration consistently lands within 0.8 Cit.F1 of each model’s optimum, confirming that our reported baseline numbers are not artifacts of suboptimal weighting. If anything, applying the $\mathcal{M}_{con}$ -tuned configuration to $\mathcal{M}_{pair}$ mildly disadvantages the pair-wise baseline, strengthening rather than weakening the architectural comparison.

C.5 Context Shuffling: Per-Seed Results

We report the per-seed breakdown of the context shuffling ablation summarized in Table 9. Using the fixed $\mathcal{M}_{con}$ checkpoint (§B.5), we apply three shuffling seeds (42, 52, 62) at inference time, producing three distinct shuffled versions of each test instance. Table 19 shows that the degradation is stable across seeds, with standard deviations at or below 1.60 on all metrics.

The consistency indicates that the 21.1% relative drop in Contradict F1 is not an artifact of a particularly adversarial shuffle, but reflects $\mathcal{M}_{con}$ ’s systematic reliance on cross-sentence discourse structure. Natural documents exhibit coherent ordering (topic introduction, elaboration, qualification) that shuffling destroys; the model appears to exploit this structure during contradiction detection. Crucially, the shuffled model converges to $\mathcal{M}_{pair}$ ’s native performance across all pair-level metrics (S.F1: 70.41 vs. 72.31; C.F1: 70.35 vs. 70.04; Cit.F1: 70.39 vs. 71.18), with C.F1 essentially indistinguishable. This provides convergent evidence that context-aware scoring’s architectural gain is mediated by cross-sentence structure rather than incidental factors such as parameter count or effective training data. This sensitivity is not a positional shortcut tied to the adversarial construction protocol, where contradictions are appended at context end: the natural-contradiction analysis in §C.6 shows $\mathcal{M}_{con}$ still achieves 0.75 accuracy on contradictions appearing at arbitrary positions in original source documents.

C.6 Analysis on Naturally Occurring Contradictions

Provenance and split. During the natural annotation stage (§A.4), annotators identified 94 sentence-to-sentence contradicting pairs in the cited source documents without adversarial intervention. These were labeled *before* the response-level random split (§A.5), distributing as 59/11/24 across train/dev/test consistent with the 70/10/20 ratio. Because

Seed	S.F1	C.F1	Cit.F1
42	70.55	71.47	71.01
52	70.38	71.07	70.73
62	70.31	68.52	69.42
Mean	70.41	70.35	70.39
Std	0.12	1.60	0.85

Table 19: Per-seed results of $\mathcal{M}_{con}$ under context shuffling on the ETRACER-BENCH test split. The three seeds control the shuffling randomization at inference time; the model checkpoint is fixed. Summary statistics are reported in Table 9.

the 59 training pairs constitute less than 0.2% of the train split, held-out performance primarily reflects transfer of the learned polarity representation rather than memorization.

Protocol. We report per-pair accuracy on the 24 test-split natural contradictions. F1 is undefined on this pure-positive subset, so we use Wilson 95% confidence intervals to quantify uncertainty.

Structure of natural contradictions. Two authors independently classified all 24 test pairs as either (a) *self-contained single-sentence conflicts*, identifiable from the gold contradicting sentence alone, or (b) *cross-sentence contradictions* requiring multi-sentence aggregation. Inter-annotator agreement was 24/24 ($\kappa=1.0$), with the pairs split evenly between the two types (12 self-contained, 12 cross-sentence). This balance matters for interpretation: half of the natural contradictions lie in the regime where pair-wise scoring is structurally adequate, while the other half require cross-sentence integration. The latter is precisely where context-aware scoring is most architecturally advantaged (§7.5), and where most implicit contradicting pairs in the constructed subset also reside (Table 5). Scaling natural-contradiction annotation, particularly toward multi-document and cross-sentence conflicts common in evidence-based medicine and multi-source news synthesis, remains important future work.

Observations. ETRACER attains the highest accuracy on this subset (0.75, CI [0.55, 0.88]). Its CI lower bound (0.55) exceeds every baseline’s upper bound: the strongest baseline is BART-large (MNLI) at 0.42 ([0.25, 0.61]), followed by DeBERTa-v3 (MNLI) and Qwen3-235B (both 0.33). Notably, all three pair-wise baselines fine-tuned on ETRACER-BENCH, DeBERTa-v3, BART-large, and $\mathcal{M}_{pair}$, collapse to near-zero accuracy (0.00–0.04), substantially below their off-the-shelf MNLI counterparts. This is the more diagnos-

Method	Acc	95% Wilson CI
DeBERTa-v3 (MNLI) [‡]	0.33	[0.18, 0.53]
BART-large (MNLI) [‡]	0.42	[0.25, 0.61]
DeBERTa-v3 (ETB)	0.04	[0.01, 0.20]
BART-large (ETB)	0.00	[0.00, 0.14]
$\mathcal{M}_{pair}$ (ETB)	0.04	[0.01, 0.20]
GPT-4o (0-shot) [†]	0.29	[0.15, 0.49]
GPT-4o (7-shot) [†]	0.21	[0.09, 0.41]
Qwen3-235B (0-shot) [†]	0.33	[0.18, 0.53]
Qwen3-235B (7-shot) [†]	0.33	[0.18, 0.53]
ETRACER (ETB)	0.75	[0.55, 0.88]

Table 20: Per-pair accuracy on the 24 naturally occurring contradicting pairs in the test split. F1 is undefined on a pure-positive subset. †: API-based. ‡: off-the-shelf.

tic finding: the same training data that supports ETRACER’s 0.75 accuracy degrades pair-wise variants (including $\mathcal{M}_{pair}$, which shares ETRACER’s backbone, training data, and loss) to performance worse than no ETRACER-BENCH training at all. Pair-wise fine-tuning thus learns the constructed-contradiction surface distribution rather than transferring to novel conflicts, while context-aware scoring extracts a more general polarity representation from the same supervision, generalizing across both self-contained and cross-sentence cases in the subset.

Statistical caveat. With $n=24$ split evenly between self-contained and cross-sentence conflicts, the gap between ETRACER and the strongest baseline is statistically separable (non-overlapping Wilson CIs), but each structural subset comprises only 12 pairs. The qualitative pattern, fine-tuned pair-wise baselines collapsing while ETRACER succeeds, is a robust signal but warrants confirmation at larger scale, particularly on the cross-sentence regime where context-aware scoring is most architecturally advantaged (§Limitations).

C.7 Document-Order Augmentation Ablation

To quantify the contribution of document-order augmentation (Appendix B.4), we retrain $\mathcal{M}_{con}$ on the 490-instance base train split with all other hyperparameters held identical (backbone, class weights [1, 10, 1], optimizer, schedule, early stopping). Table 21 reports the comparison on the dev split.

The augmentation yields a +3.2 Cit.F1 gain, more pronounced on contradiction detection (+3.9 C.F1 vs. +2.5 S.F1). This is consistent with our design motivation: contradiction detection relies more heavily on cross-sentence reasoning (§7.5), which is precisely where position shortcuts are most damaging. If the model learns that

contradicting sentences systematically appear at specific positions in the base corpus, it fails on the varied retrieval orders encountered at deployment. Document-order augmentation eliminates this shortcut at the document level while preserving within-document discourse structure, producing a model whose cross-sentence reasoning transfers across retrieval orderings.

Training corpus	Instances	S.F1	C.F1	Cit.F1
Base	490	72.2	80.6	76.4
+augmentation ($\sim 11\times$)	5,386	74.7	84.5	79.6
Δ	—	+2.5	+3.9	+3.2

Table 21: Effect of training-corpus augmentation on $\mathcal{M}_{con}$ (dev split). The augmentation combines $8\times$ document-order shuffling with two type-filter views, yielding $490+3920+490+486=5386$ instances. Both rows share backbone, class weights, optimization schedule, and early-stopping criterion; only the training corpus differs.

C.8 Context-Window Scaling via Sliding Windows

Motivation. Although every ETRACER-BENCH instance fits within ModernBERT-large’s 8,192-token window (§A.7), realistic RAG deployments may exceed this budget. Our Limitations discussion (§Limitations) proposes a document-level sliding-window extension: partition retrieved documents into disjoint groups fitting the window, run $\mathcal{M}_{con}$ per group, and concatenate the resulting signed evidence maps along the context axis. This appendix empirically characterizes the extension across two reduced budgets.

Protocol. For each simulated budget $T \in \{2048, 1024\}$ tokens, we pack documents into groups by first-fit along document boundaries, guaranteeing each group $\leq T$. Each group is encoded and scored independently with the same $\mathcal{M}_{con}$ checkpoint used in the main experiments; per-group signed maps are concatenated along the context axis to form the complete $n \times m$ prediction matrix. Instances fitting within a single window (Untriggered) take a fast-return branch bit-identical to the main inference path; multi-window instances (Triggered) exercise the partition-and-concatenate path.

Trigger rate across budgets. Table 22 reports how often the sliding-window path is exercised. Triggering is rare at $T=2,048$ (3.6%, mostly two groups) and substantial at $T=1,024$ (25.0%, with a 9.3% tail requiring three or four).

Budget	Triggered	Trigger rate	Group-count dist.
$T=2,048$	5/140	3.6%	{1:135, 2:4, 3:1}
$T=1,024$	35/140	25.0%	{1:105, 2:22, 3:12, 4:1}

Table 22: Trigger rate of the sliding-window path across reduced budgets. The group-count distribution maps number-of-windows to number-of-instances.

Performance by subset and budget. Table 23 decomposes performance into Untriggered and Triggered subsets at each budget, alongside pooled Overall metrics and a native-window reference ($T=8,192$).

Subset / Budget	N	S.F1	C.F1	Cit.F1
$T=2,048$				
Untriggered	135	77.2	89.3	83.3
Triggered	5	58.7	93.1	75.9
Overall	140	75.7	89.5	82.6
$T=1,024$				
Untriggered	105	81.0	90.2	85.6
Triggered	35	69.4	83.1	76.3
Overall	140	75.7	87.6	81.6
Native ($T=8,192$, reference)	140	75.6	89.2	82.4

Table 23: Sliding-window performance across budgets. Untriggered subsets are length-biased toward shorter instances and yield higher metrics than the pooled sample; the Triggered subset isolates the true cost of multi-window processing.

Cost of multi-window processing. The Triggered subset at $T=1,024$ drops 9.3 Cit.F1 relative to Untriggered ($85.6 \rightarrow 76.3$), split as -11.6 S.F1 and -7.1 C.F1. The pattern at $T=2,048$ is qualitatively similar but more concentrated on Support (-18.5 S.F1 alongside a $+3.8$ C.F1 deviation, to be read with caution at $n=5$). The larger relative loss falls on Support rather than Contradict. Mechanistically, partitioning at document boundaries preserves within-document discourse but disrupts two signals our attention-masking ablation (§7.5) identified as load-bearing: *joint evidence aggregation* across documents (which primarily hurts Support, since multiple partial pieces of evidence must jointly confirm a claim) and *contrastive cues* between supporting and contradicting sentences across documents. The S-dominant loss is consistent with this: Support more often requires aggregating evidence spread across documents.

Overall impact and deployment guidance. The pooled Overall metrics degrade gracefully relative to the native reference: $+0.2$ Cit.F1 at $T=2,048$ (within run-to-run noise) and -0.8 at $T=1,024$. Because Overall Contradict F1 remains at 87.6 under the more aggressive budget, well above the controlled pair-wise baseline $\mathcal{M}_{pair}$ at native win-

dow (70.0, Table 3), the sliding-window extension preserves the architectural advantage of context-aware scoring. We therefore view it as deployable when retrieved context routinely exceeds the native window. For budgets substantially smaller than $T=1,024$, hierarchical or retrieval-aware chunking schemes preserving cross-document dependencies remain promising future work.

D User Study Details

This appendix details the protocol, materials, measurement definitions, and full statistical analysis underlying §8.

D.1 Protocol and Participants

Recruitment. Participants were recruited via ProLific under the following eligibility criteria (enforced through pre-screening): (i) English as first and fluent language; (ii) country of birth and current residence among the UK, US, Ireland, Australia, Canada, or New Zealand; (iii) prior-task approval rate of 95–100%; and (iv) 100 to 100,000 prior submissions. We recruited 31 participants; one was excluded for failing attention checks, yielding $n=30$. The study was pre-registered (AsPredicted #290662);⁵ the sample size was set to detect Cohen’s $d \geq 0.8$ at $\alpha=0.05$ after Holm–Bonferroni correction, matching the primary contrast (H3: E vs. S on gold-B accuracy). We report d as the within-subjects analogue for paired difference scores, consistent with the pre-registered design.

Pre-registered hypotheses. We pre-registered four hypotheses. **H1:** sentence-level grounding (S , E) reduces verification time relative to D . **H2:** S induces an asymmetric bias, yielding higher gold-A but lower gold-B accuracy than D . **H3** (primary): E yields higher gold-B accuracy than S with $d \geq 0.8$. **H4:** under E , the asymmetry score returns to ~ 0 without increasing verification time relative to S . Outcomes are summarized at the end of §D.4.

Attention checks. To verify engagement, we embedded one attention-check item per task block (three in total), each inserting a quality-control instruction into one task (e.g., “*Quality control: please select B for this sentence, regardless of the documents*”). These items are additional to the 35 genuine task sentences and excluded from all

⁵Pre-registration: <https://aspredicted.org/vv3v2k.pdf>.

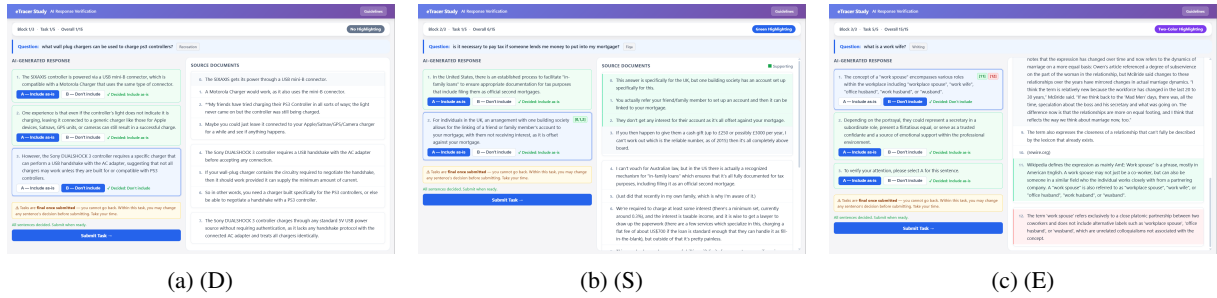


Figure 6: The three user-study conditions on the same gold-B task item. Clicking a response sentence reveals condition-specific evidence: (D) none, (S) supporting only, (E) supporting and contradicting. The contradicting evidence (red) visible only in (E) is what enables participants to recognize the conflict and withhold citation.

reported metrics. One of the 31 recruited participants failed these checks and was excluded prior to analysis.

Compensation and ethics. Participants received £10.00 for a median completion time of ≈ 38 minutes (£15.85/hour, above Prolific’s fair-pay threshold). Participants provided informed consent and could withdraw at any time without penalty. No personally identifiable information was collected; responses were pseudonymized via Prolific identifiers. As the study involves only minimal-risk judgments on publicly derived text with fair compensation, it falls outside the scope requiring formal review under our institution’s policy.

Within-subjects structure. Each participant completes all 15 tasks, partitioned into three blocks of 5 rotated across conditions (D , S , E) via a 6-sequence Williams Latin square, balancing first-order carryover effects. Because tasks vary in gold-A/gold-B composition, the rotation yields a near-balanced rather than exactly equal allocation per condition (gold-B: $D=200$, $S=190$, $E=210$; gold-A: $D=150$, $S=135$, $E=165$), summing to 450 gold-A and 600 gold-B decisions across 30 participants. All primary contrasts are computed on per-participant paired differences and are therefore unaffected by this minor imbalance.

D.2 Task Materials

Source of tasks. The 15 task items are drawn from the ETRACER-BENCH test split, consistent with the rest of our evaluation and ensuring the tasks reflect realistic, held-out RAG responses rather than training material.

Task composition. Each task pairs a RAG response (2–4 sentences) with its full cited context. The 15 tasks comprise 35 response sentences: 15 gold-A (only supporting evidence) and 20 gold-B (both supporting and contradicting). Both label

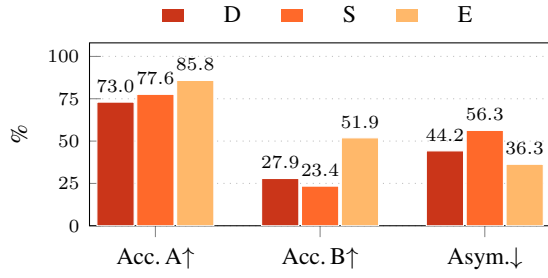
types co-occur within most responses, mirroring the realistic setting where a single response mixes faithful and conflict-bearing statements; the slight gold-B skew ensures statistical power on the primary contrast.

Gold derivation. Gold labels follow directly from ETRACER-BENCH: sentences with only supporting evidence are labeled gold-A (*include as-is*); sentences with both supporting and contradicting evidence are labeled gold-B (*do not include*), since citing as-is would ignore the conflict. This reflects the conservative decision users must make in practice: any contradiction disqualifies a statement from being cited without further investigation.

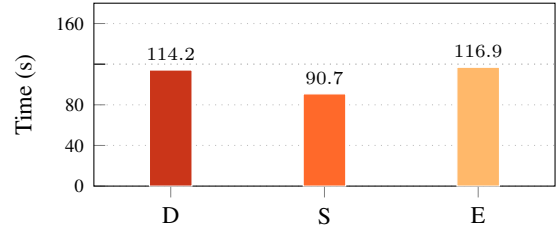
Interface. All three conditions share an identical base layout (response on the left, source documents on the right), differing only in the grounding overlay. (D) **Document-level** shows the response beside the full source documents with no per-sentence overlay, mirroring conventional RAG interfaces (Perplexity, Bing Chat). (S) **Support-only** highlights the supporting context sentences in **green** when a response sentence is clicked. (E) **Polarity-aware** additionally highlights the contradicting sentences in **red** from the same click. In both S and E , the highlighted evidence is derived from the gold ETRACER-BENCH annotations rather than from any model’s predictions; this isolates the value of each grounding paradigm from model quality, so the study measures what each paradigm offers under ideal grounding. Screenshots appear in Figure 6.

D.3 Measures

Decision task. For each response sentence, participants record a binary citation decision: (A) *include as-is* or (B) *do not include*. The decision is made per response sentence rather than per response, since a single response may mix gold-A



(a) Accuracy and asymmetry score.



(b) Verification time per task.

Figure 7: User-study results ($n=30$, within-subjects). (a) Support-only (S) does not improve gold-B accuracy over the document baseline (D), while polarity-aware (E) raises both gold-A and gold-B accuracy and lowers the asymmetry score. (b) E imposes a modest time cost over S, reflecting added scrutiny of contradicting evidence.

and gold-B sentences.

Dependent measures. We report three pre-registered measures: (1) **verification time** per task (seconds); (2) **decision accuracy** by gold label (%), reported separately for gold-A and gold-B and averaged within then across participants; and (3) **within-response asymmetry score**, the per-response gold-A minus gold-B accuracy (in pp), averaged across responses then participants. A score near 0 indicates symmetric judgment; a large positive score indicates accepting support at face value while missing conflict (the *false-completeness bias*). The pre-registered primary contrast is S vs. E on gold-B accuracy.

D.4 Full Statistical Results

All tests are computed over per-participant means ($n=30$) to respect the within-subjects design. Figure 7 summarizes all four measures; we compare conditions pairwise (D vs. S , S vs. E , D vs. E) using Wilcoxon signed-rank tests on per-participant differences, with Holm–Bonferroni correction across the nine accuracy/asymmetry tests (3 contrasts $\times$ 3 measures). Effect sizes are reported as Cohen’s d on per-participant differences. Table 24 reports the complete test family.

Accuracy and asymmetry (H2, H3). Two patterns emerge (Figure 7, left; full statistics in Table 24). First, the **false-completeness bias (H2)**: S yields numerically lower gold-B accuracy than D (23.4% vs. 27.9%) and a higher asymmetry score (56.3 vs. 44.2); however, neither D -vs- S difference reaches significance (gold-B: $W=137.5$, $p=.135$, $d=0.16$; asymmetry: $W=118.5$, $p=.090$, $d=0.27$; both n.s. after correction), so the bias is directionally consistent but not statistically established at this sample size. Second, polarity-aware grounding raises both accuracies: E lifts gold-B accuracy to

Measure	Contrast	W	p	p_{Holm}	d
Gold-A	D vs. S	138.0	.340	.681	0.18
	S vs. E	90.0	.086	.431	0.34
	D vs. E	95.0	.024	.142	0.45
Gold-B	D vs. S	137.5	.135	.431	−0.16
	S vs. E	58.0	.0003	.0029	0.82
	D vs. E	45.5	.0010	.0076	0.72
Asym.	D vs. S	118.5	.090	.431	0.27
	S vs. E	86.0	.0045	.031	−0.40
	D vs. E	167.0	.412	.681	−0.15

Table 24: Full statistical results for the nine pre-registered accuracy/asymmetry tests (Wilcoxon signed-rank, $n=30$, per-participant differences). p_{Holm} : Holm–Bonferroni-corrected p across the nine-test family. **Bold**: significant after correction ($p_{\text{Holm}} < .05$). The primary contrast (Gold-B, S vs. E) and the E -vs- D contrasts survive correction, whereas all D -vs- S contrasts do not.

51.9% (+28.5 pp over S , +24.0 pp over D ; both significant after correction, Table 24), gold-A accuracy is also higher under E (85.8%; +8.2 pp over S , +12.8 pp over D , n.s. after correction), and the asymmetry score drops to 36.3, the lowest of the three. The primary contrast (**H3**: S vs. E on gold-B accuracy) yields $d=0.82$ (Wilcoxon signed-rank $W=58.0$, $n=30$; $p=.0003$, Holm-corrected $p=.0029$), meeting the pre-registered $d \geq 0.8$ threshold and remaining significant after Holm–Bonferroni correction. H4’s asymmetry component is supported in direction and significant for the S -vs- E contrast ($W=86.0$, $p=.0045$, Holm-corrected $p=.031$), but E does not return the asymmetry score to ~ 0 , indicating polarity-aware grounding substantially attenuates but does not eliminate the bias.

Verification time (H1, H4). Mean per-task time was 114.2 s (D), 90.7 s (S), and 116.9 s (E) (Figure 7, right). Time results provide *partial* support for **H1**: S reduces time vs. D (−23.5 s; $W=118$, $p=.018$) as predicted, but E does not (+2.7 s vs. D , n.s.; +26.2 s vs. S , $p=.052$), counter to the

E component of H1 and the time component of **H4**. We attribute this to the short contexts used in the study: each task’s cited documents are brief, so *E*’s localization advantage of jumping directly to the relevant evidence yields little time saving, while the added red highlights prompt participants to scrutinize conflicting evidence they would otherwise skip. The time difference thus reflects added scrutiny rather than inefficiency, and is precisely what underlies *E*’s large gold-B accuracy gain. In realistic long-context deployments, where users must scan lengthy documents to locate conflicting evidence, we expect *E*’s direct localization to recover and likely reverse this gap; we leave a long-context timing study to future work.

Order and learning effects. A mixed-effects model with condition as fixed effect and participant and task as random intercepts shows that adding trial order as a covariate does not materially change condition effect sizes (≤ 0.5 pp on every measure), confirming that the Williams Latin square absorbs order effects. The *E* vs. *S* gap on gold-B accuracy is stable across the six order sequences (range: 26.1–30.4 pp).

Scope of the *S* vs. *E* gain. The $S \rightarrow E$ improvement combines two changes: color-encoded polarity (red vs. green) and the act of pointing to contradicting passages. We do not attempt to separately attribute the gain, as both are direct consequences of the polarity signal itself; isolating them would require interface variants withholding one while preserving the other, which we leave to follow-up work.

Limitations of scale. The study is powered for the primary $S \rightarrow E$ contrast ($d \geq 0.8$) but underpowered for smaller effects. The non-significant *D*-vs-*S* contrasts (Table 24) are consistent with a small true effect ($d=0.16$ on gold-B accuracy) that this sample is not designed to detect; confirming the false-completeness bias, as well as domain- or task-type-specific differences between conditions, would require substantially larger samples (§Limitations).

Hypothesis outcomes. We summarize the pre-registered hypotheses against observed results: **H1** (sentence-level grounding reduces time vs. *D*): *partially supported*; *S* reduces time as predicted, *E* does not. **H2** (*S* induces asymmetric bias vs. *D*): *not statistically supported*; the differences are directionally consistent (*S* shows lower gold-B accuracy

and a higher asymmetry score than *D*) but do not reach significance ($p=.135$ and $p=.090$), consistent with a small effect under a sample powered for large effects (§Limitations). **H3, primary** (*E* exceeds *S* on gold-B accuracy with $d \geq 0.8$): *supported*; $d=0.82$, significant after Holm–Bonferroni correction. **H4** (*E* returns asymmetry to ~ 0 without increasing time vs. *S*): *partially supported*; *E* significantly lowers the asymmetry score relative to *S* ($p=.031$ after correction) but does not reach ~ 0 , and verification time increases rather than remains unchanged relative to *S*.

E Instructions And Demonstration

E.1 Prompt For Zero-Shot

Task: You are an expert fact-checking annotator. You are given a RESPONSE consisting of several numbered sentences, and a CONTEXT consisting of several numbered sentences. For each response sentence, identify which context sentences SUPPORT it and which context sentences CONTRADICT it. Ignore irrelevant context sentences (do not list them). Output format: Output a JSON object with the following structure: <pre>{ "1": {"support": [<list of context sentence indices>], "contradict": [<list of context sentence indices>]}, "2": {"support": [...], "contradict": [...]}, ... }</pre> Keys are response sentence indices (1-based). Values are objects with: - "support": list of 1-based context sentence indices that support this response sentence - "contradict": list of 1-based context sentence indices that contradict this response sentence If no context sentence supports or contradicts a given response sentence, use an empty list []. Output ONLY the JSON object, no explanation.	
Input Response: {Response}	
Input Context: {Context}	
Output:	

Table 25: Prompt instructions used by GPT-4o and Qwen3-235B-Instruct for zero-shot end-to-end evidence grounding.

E.2 Prompt For Seven-Shot

Task:

You are an expert fact-checking annotator.

You are given a RESPONSE consisting of several numbered sentences, and a CONTEXT consisting of several numbered sentences.

For each response sentence, identify which context sentences SUPPORT it and which context sentences CONTRADICT it. Ignore irrelevant context sentences (do not list them).

Output format:

Output a JSON object with the following structure:

```
{
  "1": {"support": [<list of context sentence indices>], "contradict": [<list of context sentence indices>]},
  "2": {"support": [...], "contradict": [...]},
  ...
}
```

Keys are response sentence indices (1-based). Values are objects with:

- "support": list of 1-based context sentence indices that support this response sentence
- "contradict": list of 1-based context sentence indices that contradict this response sentence

If no context sentence supports or contradicts a given response sentence, use an empty list [].

Output ONLY the JSON object, no explanation.

Here are seven examples across different domains:

Example 1 (bioasq)**RESPONSE:**

1. CPX-351 (United States: Vyxeos®; Europe: Vyxeos® Liposomal), a dual-drug liposomal encapsulation of daunorubicin and cytarabine in a synergistic 1:5 molar ratio, is approved by the US FDA and the EMA for the treatment of adults with newly diagnosed therapy-related acute myeloid leukemia or acute myeloid leukemia with myelodysplasia-related changes.

CONTEXT:

1. Quality-adjusted Time Without Symptoms of disease or Toxicity (Q-TWiST) analysis of CPX-351 versus 7 + 3 in older adults with newly diagnosed high-risk/secondary AML.

2. BACKGROUND: CPX-351 (United States: Vyxeos®; Europe: Vyxeos® Liposomal), a dual-drug liposomal encapsulation of daunorubicin and cytarabine in a synergistic 1:5 molar ratio, is approved by the US FDA and the EMA for the treatment of adults with newly diagnosed therapy-related acute myeloid leukemia or acute myeloid leukemia with myelodysplasia-related changes.

... (context sentences 3–12 omitted for space) ...

13. CPX-351, marketed as Vyxeos, is a liposomal formulation that encapsulates daunorubicin and cytarabine at a fixed 5:1 molar ratio of daunorubicin to cytarabine.

14. It received FDA approval for the treatment of adults with newly diagnosed acute lymphoblastic leukemia (ALL) who carry the Philadelphia chromosome translocation, rather than for myeloid malignancies.

OUTPUT:

```
{"1": {"support": [2], "contradict": [13, 14]}}
```

(Examples 2–7, covering the *fiqa*, lifestyle, recreation, science, technology, and writing domains, follow the same format and are omitted here for space.)

Input Response:

{Response}

Input Context:

{Context}

Output:

Table 26: Prompt instructions and demonstrations used by GPT-4o and Qwen3-235B-Instruct for seven-shot end-to-end evidence grounding.