Better Scores, Worse Grounding: Hidden Regressions after Fine-Tuning in Dialogue Fact Verification

Hyunkyung Park and Arkaitz Zubiaga
Queen Mary University of London
{hyunkyung.park, a.zubiaga}@qmul.ac.uk

Abstract

In dialogue fact verification (DFV), responses often depend on prior turns for correct interpretation, yet systems are still judged mainly by aggregate benchmark scores. We study a *hidden grounding regression*: aggregate Macro-F1 improves after fine-tuning while previously correct, context-dependent pronoun cases become newly wrong and show stronger premise-side sensitivity than cases that remain correct. On three referent-annotated audit sets constructed from DialFact and FaithDial, we audit six encoder-only verifiers before and after matched source-specific fine-tuning through prediction-transition analysis and the Premise-Preference Score (PPS), a control-adjusted masking diagnostic. Fine-tuning improves Macro-F1 across the six-model/three-evaluation-set panel, yet newly regressed cases show stronger premise-side sensitivity than stable-correct cases under PPS in 17 of 18 evaluated comparisons, showing that aggregate gains can conceal regressions on a controlled dialogue-grounding audit.

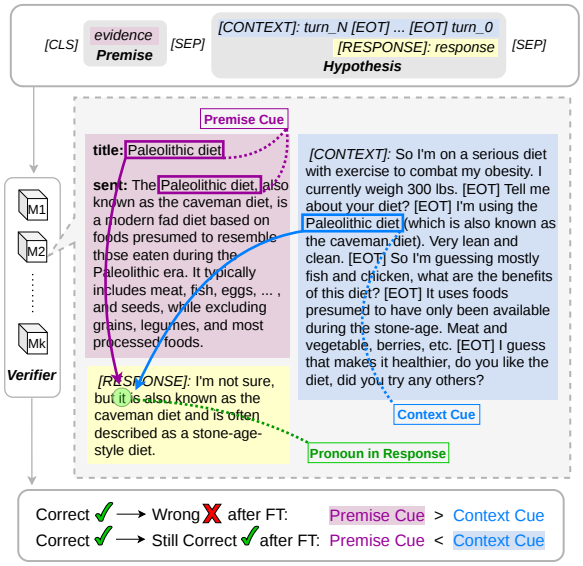

Figure 1: Negative-flip example: after fine-tuning, a previously correct prediction becomes wrong when the response pronoun requires dialogue-context grounding but the paired evidence provides a competing cue.

1 Introduction

Fact verification (FV) is usually framed as deciding whether a claim is supported, refuted, or unverifiable given evidence. In many settings, especially premise–hypothesis verification, the claim is largely interpretable in isolation (Thorne et al., 2018). Dialogue fact verification (DFV) differs because the verifier keeps paired evidence on the premise side but must interpret a response jointly with prior dialogue context on the hypothesis side, as responses often contain pronouns, ellipsis, and deictic expressions whose meaning is fixed only by earlier turns (Gupta et al., 2022; Chamoun et al., 2023). This paper examines whether fine-tuning can improve aggregate verification performance while worsening behavior on a controlled audit of dialogue-conditioned grounding.

We use *hidden regressions* as shorthand for *hidden grounding regressions*: aggregate Macro-F1 improves after fine-tuning, but examples that were previously correct and require dialogue-contextual reference resolution become newly wrong, and those newly regressed cases show relatively stronger premise-side sensitivity than cases that remain correct after fine-tuning. The regression is hidden because it is obscured by aggregate benchmark scores. This coexistence is possible because Macro-F1 is computed over the full evaluation set, so gains on other cases can outweigh regressions on a narrower referent-sensitive subset that was previously correct. It is dialogue-grounding-specific because the target response meaning is fixed by earlier turns, while the paired evidence may contain plausible but competing referential cues. We study this failure mode on a deliberately narrow, response-side pronoun slice rather than claiming to cover dialogue grounding in full generality.

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 720–737
August 2–5, 2026.

A common response is to rewrite the response into a standalone claim before verification (Chamoun et al., 2023; Wanner et al., 2025). Rewriting can be useful, but it can also change the claim to be verified and obscure the grounding dependence that makes DFV difficult. We therefore keep the original joint input intact and examine whether fine-tuning improves the controlled dialogue-grounding behavior measured here or instead raises aggregate accuracy while increasing relative sensitivity to premise-side referent cues.

Figure 1 shows this configuration with a concrete negative-flip case. The verifier sees paired evidence as the premise and the dialogue context plus response as the hypothesis, so the response becomes the claim to verify only after its pronoun is grounded in prior context. Here, *it* refers to the diet discussed in the dialogue, but the paired evidence contains a lexically similar diet cue; a verifier can therefore remain locally compatible with the response surface by leaning more on premise-side cues than on the context-conditioned interpretation. The audited failure is not simply a wrong post-FT label: it is a transition from a pre-FT gold prediction to a post-FT non-gold prediction, together with stronger post-FT premise-side sensitivity under the masking diagnostic introduced below. We use the figure only to illustrate this cue contrast; the aggregate analysis below measures the pattern over controlled audit sets.

We study this question on a conservative referent-sensitive slice of DFV in which response-side pronouns can be reliably annotated in dialogue context and have surface-matched referent cues on the premise side. We use DialFact as the in-domain DFV source and FaithDial as out-of-domain corroboration from knowledge-grounded dialogue, allowing us to test whether the same audit pattern recurs outside the original DFV source. From these sources, we construct the referent-annotated audit files used in the study: DF-Coref-Valid and DF-Coref-Test from DialFact, and two FaithDial-derived sets, FD-Random and FD-Topic. We evaluate six encoder-only verifiers before and after fine-tuning under a shared premise–hypothesis verification interface.

Our audit has two components. First, we partition examples by pre/post-fine-tuning correctness transitions, focusing on the contrast between *negative flips* (NF; correct before fine-tuning, wrong after fine-tuning) and *positive no flips* (PNF; correct both before and after fine-tuning). Second, we introduce a control-adjusted masking diagnostic, the Premise-Preference Score (PPS), which measures whether the model's gold-label support depends more on premise-side or context-side referent cues within the same example.

The main result is a hidden grounding regression under a matched source-specific fine-tuning setup. Across the six-model/three-evaluation-set panel, fine-tuning improves Macro-F1 throughout, yet NF cases show stronger premise-side sensitivity under PPS than PNF cases in 17 of the 18 pairs overall and in all six model comparisons on DF-Coref-Test. On this audited slice, aggregate benchmark gains can therefore coexist with worse behavior on a controlled dialogue-grounding audit.

Our contributions are threefold. First, we introduce a transition-conditioned diagnostic framework for DFV that combines pre/post-fine-tuning prediction transitions with PPS, a control-adjusted masking diagnostic of premise- versus context-side referent sensitivity, while keeping the original joint verifier input intact. Second, we construct and analyze referent-annotated audit files for this study: DF-Coref-Valid/Test from DialFact and FD-Random/FD-Topic from FaithDial. Finally, we provide empirical evidence that aggregate Macro-F1 improvements after matched source-specific fine-tuning can coexist with hidden grounding regressions: newly regressed cases show stronger premise-side sensitivity under PPS than stable-correct cases in 17 of 18 model–evaluation-set comparisons. We also report the PPS-stratification ablation as an exploratory analysis.

2 Related Work

2.1 Dialogue Claims Require Context

Fact verification (FV) typically assumes claims are largely interpretable in isolation (Thorne et al., 2018). In dialogue fact verification (DFV), by contrast, the hypothesis-side text is not just the response, but the response interpreted jointly with prior dialogue context (Gupta et al., 2022; Chamoun et al., 2023). DFV is particularly challenging because dialogue claims are often colloquial, underspecified, and rich in referring expressions, making them harder to interpret, retrieve evidence for, and verify than stand-alone formal claims (Gupta et al., 2022; Chamoun et al., 2023).

One common response to this problem is to rewrite or decontextualize the response into a more self-contained claim before verification (Chamoun

et al., 2023; Deng et al., 2024; Wanner et al., 2025; Gunjal and Durrett, 2024); related work on claim extraction likewise shows that poor reformulations can distort downstream fact-checking (Metropolitansky and Larson, 2025). Earlier dialogue fact-checking work showed that making claims more self-contained can help retrieval while harming downstream verification when the rewrite drifts semantically, and BiConGate makes this tension explicit as a rewrite-induced IR–FV trade-off (Chamoun et al., 2023; Park and Zubiaga, 2026).

DialFact identifies both coreference and retrieval ambiguity as central challenges in DFV (Gupta et al., 2022), and related fact-checking work shows that some cases are genuinely ambiguous enough to require explicit disambiguation rather than a single unqualified verdict (Glockner et al., 2024; Staliunaite and Vlachos, 2025). In this paper, we keep the joint verifier input fixed and focus on a narrower pronoun-grounding slice in which the gold referent is recoverable from dialogue context and a matching surface cue is available in the paired source evidence. This lets us ask whether fine-tuning changes relative reliance on premise-side versus context-side referent cues, especially when dialogue context fixes the intended referent but the paired evidence still offers a competing cue (Chen et al., 2025).

2.2 Correctness vs. Source Use

Prior work on grounded dialogue shows that correct predictions do not necessarily mean that the model used context and source evidence appropriately. FaithDial evaluates faithfulness to provided knowledge, while the BEGIN benchmark shows that attribution metrics can latch onto spurious correlations and become less reliable as sources get longer. More broadly, attribution work distinguishes factuality from source-groundedness (Dziri et al., 2022a,b; Rashkin et al., 2023).

CiteEval further argues that NLI-style support from cited passages is only an incomplete proxy for source use, because citation quality also depends on the full retrieval set and the broader question–response context (Xu et al., 2025). In dialogue settings, context attribution must handle follow-up questions grounded in previous turns, often involving coreference and other dialogue-specific phenomena, and fine-tuned attribution models improve evidence recall in these settings (Radevski et al., 2025).

In dialogue and other context-dependent settings, strong end-task performance can therefore mask weak use of conversation history or source evidence. Recent robustness work on fact verification similarly shows the need to evaluate behavior under perturbations and subpopulations rather than relying only on aggregate performance (Mamta and Cocarascu, 2025). Related context-conditioned generation work makes a complementary do-no-harm distinction, separating context-induced errors on baseline-correct items from gains on genuinely helpful contexts (Tao and Agrawal, 2026). Models also remain vulnerable to distractor mentions and other superficially plausible but misleading cues in referential reasoning (Ghaddar et al., 2024; Tighidet et al., 2024; Gautam et al., 2024; Manikantan et al., 2025).

2.3 Fine-Tuning May Shift Cue Sensitivity

Fine-tuning can improve benchmark performance without preserving the same cue-use behavior. In DFV, dialogue-specific adaptation improves in-domain performance while reducing transfer to standard fact-checking (Chamoun et al., 2023). More broadly, supervised fine-tuning has been linked to forgetting, non-monotonic changes in knowledge, and shifts in reasoning behavior (Li et al., 2024; Ye et al., 2025; Lobo et al., 2025). Related dialogue work also studies behavioral fine-tuning to improve faithfulness to provided knowledge (Razumovskaia et al., 2024). Ye et al. show that as many as 90% of parameter updates during SFT may not contribute to knowledge enhancement, and that restoring many of those updates can recover or even improve closed-book QA performance (Ye et al., 2025). Zheng et al. further report a knowledge-preference bias induced by SFT data properties: long-context SFT promotes contextual knowledge, whereas short-context SFT favors parametric knowledge (Zheng et al., 2025). Recent evidence of context-parametric inversion during instruction tuning further shows that context reliance can decline even as standard benchmark performance continues to improve (Goyal et al., 2025). Together, these findings suggest that we should compare pre- and post-fine-tuning behavior rather than looking only at post-fine-tuning accuracy. Our specific question is whether fine-tuning changes a verifier's relative sensitivity to premise-side versus context-side cues under a fixed verifier input.

3 Methodology

We study DFV cases in which the target response is not self-contained, i.e., cases in which context from the dialogue is necessary for understanding the claim for the purposes of verification. This section specifies the shared verifier input, describes the referent-sensitive setting that motivates the audit, and defines the transition-conditioned audit signals used in the experiments.

3.1 Problem Setup

Across datasets, we represent each example as $x = (C, R, E, y)$, where C is the dialogue context, R is the target response whose claim content is interpreted in context, E is the paired evidence or knowledge snippet, and y is the gold label. This shared interface lets us analyze both DialFact and FaithDial under the same verifier setting while preserving each dataset's native label space. When we use verifier-side language, the E field is the *premise side*, while the text judged against it is the *hypothesis side*. In our joint-input DFV setting, that hypothesis-side text is the concatenation of dialogue context C and response R, not the response alone.

A verifier can therefore rely more on premise-side cues than on the context-conditioned interpretation while still matching surface cues in the response. The framework below does not directly observe referent substitution or identify the full causal contribution of premise use. Instead, it measures a control-adjusted behavioral signature of relative premise-side sensitivity: how much the verifier's gold-label margin changes under premise-side masking versus context-side masking after matched perturbations. We instantiate the audit on pronoun-resolved cases because they permit high-precision referent annotation and matched masking controls across datasets. Our empirical claims are limited to this conservative referent-sensitive slice of dialogue verification.

3.2 Audit Signals

Prediction transitions. For an example $x = (C, R, E, y)$, let $\hat{y}^{\text{pre}}(x)$ and $\hat{y}^{\text{post}}(x)$ denote the model predictions before and after fine-tuning. We assign each example to one of four mutually exclusive transition groups:

- **Negative No Flip (NNF)**: $\hat{y}^{\text{pre}}(x) \neq y \ \wedge \ \hat{y}^{\text{post}}(x) \neq y$;
- **Positive Flip (PF)**: $\hat{y}^{\text{pre}}(x) \neq y \ \wedge \ \hat{y}^{\text{post}}(x) = y$;
- **Negative Flip (NF)**: $\hat{y}^{\text{pre}}(x) = y \ \wedge \ \hat{y}^{\text{post}}(x) \neq y$;
- **Positive No Flip (PNF)**: $\hat{y}^{\text{pre}}(x) = y \ \wedge \ \hat{y}^{\text{post}}(x) = y$.

Here, *positive* and *negative* are defined according to post-fine-tuning correctness, and *flip* indicates whether correctness changes from pre-FT to post-FT. We use these groups as an observational partition of outcomes rather than a causal decomposition of error sources.

Our main interest lies in comparing NF versus PNF: newly regressed cases versus cases that remain correct after fine-tuning. Both groups are correct before fine-tuning and differ only in whether that earlier success is preserved after fine-tuning. This removes one obvious confound—pre-FT errors—while not ruling out others, such as residual ambiguity, low confidence margins, or inherent example difficulty. § 5.2 therefore interprets the NF-versus-PNF contrast as a controlled observational comparison and evaluates its recurrence across model families and datasets rather than as a causal estimate of fine-tuning.

Gold-label margin. We quantify verifier support for the gold decision using the gold-label margin. For a model state θ and input x, we define the gold-label margin as

$$m(x) = \log p_\theta(y \mid x) - \max_{y' \neq y} \log p_\theta(y' \mid x) \quad (1)$$

where y is the gold label and y' ranges over all labels other than y. Using the gold label in Eq. 1, rather than each state's predicted label, keeps the audit target fixed when pre-FT and post-FT predictions differ. This is especially important for PF and NF cases: the issue is the redistribution of support for the correct decision when fine-tuning changes the outcome, not the confidence with which each state supports its predicted label. PPS diagnoses relative support redistribution around the gold decision; it is neither a coreference score nor a causal attribution of entity use.

Target spans and diagnosable examples. For each example, let $S_{\text{ctx}}(x)$ be the set of gold referent mentions annotated in `referent_info.mentions`. Let $S_{\text{pre}}(x)$ be the set of case-insensitive exact surface matches of those referent strings in evidence titles and contents on the premise side. We form x^{ctx} by masking all mentions in $S_{\text{ctx}}(x)$ and x^{pre} by masking all mentions in $S_{\text{pre}}(x)$. If the same referent cue appears more than once on one side, we mask all matched mentions on that side.

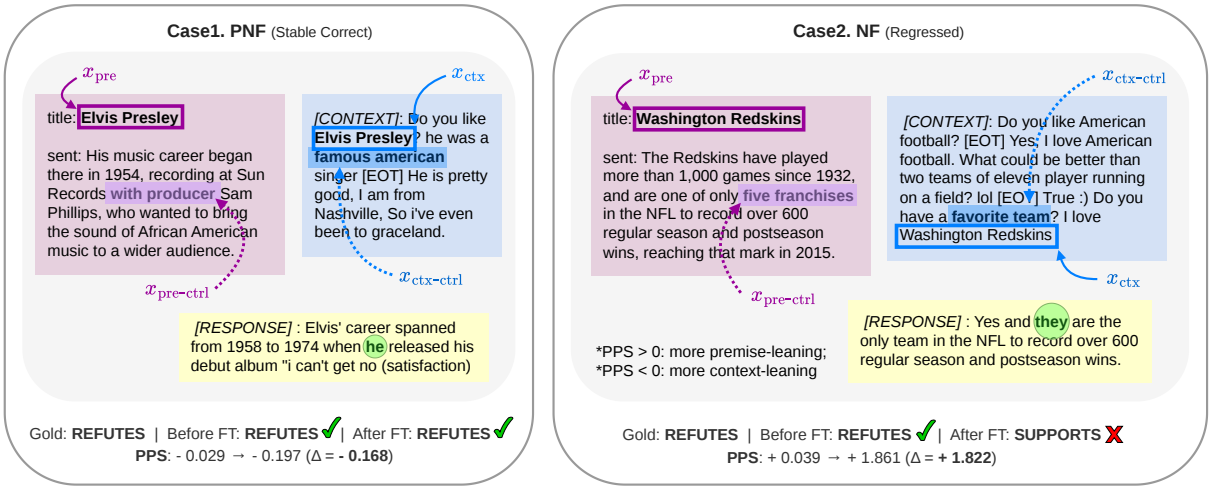

Figure 2: PNF and NF examples for the surface-matched PPS diagnostic. Highlighted spans show premise/context referent masks and same-side non-referential controls; the response pronoun is marked in green.

Operationally, each target or control span is replaced before model encoding with a word-count-preserving sequence of the checkpoint tokenizer's `mask_token`, falling back to `[MASK]` only if no tokenizer mask token is defined. Thus a two-word span is replaced by two mask-token instances, and the resulting string is then encoded with the model's native tokenizer.

PPS is computed only on *diagnosable* examples. The annotated referent must be visible in context, at least one premise-side surface match must be available, and all masked variants must remain valid model inputs for the verifier.

For the main tables, this restriction prefers auditability over coverage and keeps the premise-side target definition conservative: only case-insensitive exact surface matches of the annotated referent are masked on the premise side. Under this exact-match operationalization, PPS is a diagnostic; it is not an alias- or paraphrase-invariant estimate of premise use.

Matched controls. To control simple masking burden, we also construct same-side matched control variants $x^{\text{ctx-ctrl}}$ and $x^{\text{pre-ctrl}}$. These controls are not intended to equalize semantic importance or discourse function. Instead, each control span is selected algorithmically from the same side as the target so that the number of masked spans is matched, word count is preserved when possible, span-length mismatch is minimized, and overlap with the gold referent spans or their premise-side surface matches is avoided. When multiple control candidates are available, the search prefers the same context turn or premise field as the target, then candidates with the target word count, then smaller character-length mismatch, and finally earlier occurrence in the serialized text.

The adjustment therefore controls only simple span-count, span-length, and placeholder effects. Figure 2 illustrates this operationalization with one PNF case and one NF case; these labels denote correctness transitions before and after fine-tuning rather than PPS categories. In the PNF case, fine-tuning preserves the correct prediction, whereas in the NF case it changes a previously correct prediction into an error while producing a large positive PPS, illustrating the transition-conditioned contrast used in the audit.

Eligibility, target spans, and matched-control spans are determined once from the raw example and then reused for both model states. Pre/post PPS differences therefore come from model behavior under a fixed valid perturbation pattern, not from state-dependent changes in what counts as diagnosable. Unless otherwise stated, Macro-F1 is computed on the full post-adjudication evaluation file. Transition shares, transition-wise PPS means, NF–PNF gaps, and intervention denominators are computed on the fixed joint-diagnosable subset, defined as examples that are diagnosable in both the pre-FT and post-FT effect tables. If either transition group is empty, the corresponding PPS mean and gap are undefined and are not counted as a positive or negative directional gap.

Premise-Preference Score (PPS). Let $E_{\text{ctx}} = m(x) - m(x^{\text{ctx}})$ and $E_{\text{pre}} = m(x) - m(x^{\text{pre}})$ be the margin drops caused by masking context-side and premise-side referent cues. Let $E_{\text{ctx}}^{\text{ctrl}}$ and $E_{\text{pre}}^{\text{ctrl}}$

be the corresponding control drops. We define

$$\mathrm{PPS} = \left(E_{\mathrm{pre}} - E_{\mathrm{pre}}^{\mathrm{ctrl}}\right) - \left(E_{\mathrm{ctx}} - E_{\mathrm{ctx}}^{\mathrm{ctrl}}\right) \quad (2)$$

Positive PPS indicates greater premise-side sensitivity under the chosen target-set definition and matched-control operationalization. Negative PPS indicates stronger context-side sensitivity. We interpret PPS comparatively, both within an example as a relative premise-versus-context contrast and across transition groups through aggregated group-level comparisons, especially the post-FT NF–PNF gap reported in § 5.2.

3.3 Evaluation Setting

Evaluation sets. Because PPS requires referent annotations and matched masking controls, the main PPS tables do not use the original DialFact splits directly. We use *DialFact* and *FaithDial* only for the original source datasets and their native splits. For the audited files used in this study, we use split-specific names: DF-Coref-Valid (derived from the original DialFact validation split), DF-Coref-Test (derived from the original DialFact test split), FD-Random, and FD-Topic. When the split distinction is irrelevant, we use DF-Coref for the in-domain audited family as a whole.

Native label spaces. Because the two datasets use different annotation schemes, direct comparison at the label level is not meaningful. DialFact follows its original three-way scheme—Supports, Refutes, and Not Enough Information (NEI)—(Gupta et al., 2022), whereas FaithDial is evaluated here in the binary non-hallucination/hallucination label space defined in § 4. We therefore report Macro-F1 within each dataset's native label space and use the shared audit signals above for cross-dataset comparison.

Fine-tuning protocol. Fine-tuning is source-specific rather than performed on the audit files. As detailed in § 5.1, DialFact-derived audits use a rebuilt DialFact training pool excluding DF-Coref-Valid, whereas FaithDial-derived audits use the native FaithDial training file and binary label space.

4 Evaluation Set Construction

We build one in-domain audited family, DF-Coref, and two out-of-domain corroboration sets, FD-Random and FD-Topic, for the PPS analysis. DF-Coref contains two split-specific files: DF-Coref-Valid, derived from the original DialFact validation split, and DF-Coref-Test, derived from the original DialFact test split. DF-Coref-Test supports the main audit in § 5.2; DF-Coref-Valid is used only for PPS-tertile estimation and intervention selection in § 5.3.

Audit set	#	Label dist.	Avg. turns	Ref. men./turns
DF-Coref-Valid	467	176 / 163 / 128	4.41	1.82 / 1.70
DF-Coref-Test	315	115 / 102 / 98	3.56	1.59 / 1.50
FD-Random	405	111 / 294	4.47	1.00 / 1.00
FD-Topic	483	114 / 369	4.33	1.00 / 1.00

Table 1: Summary statistics for the post-adjudication audit files used in the PPS analysis.

Reviewed file	Review n	Correct	Wrong	Ambig.
DF-Coref-Valid	94	85	7	2
DF-Coref-Test	63	59	2	2
FD-Random	83	70	13	0
FD-Topic	108	105	3	0

Table 2: Independent second-review outcome counts for the reviewed stratified samples. DF-Coref rows summarize the valid-coref and test-coref audit reviews; correct counts include annotations with an alternative valid span under the selection rule.

FD-Random and FD-Topic are supplementary out-of-domain checks rather than co-equal primary benchmarks. Table 1 reports the post-adjudication audit-file sizes, label distributions, average dialogue lengths, and referent-annotation coverage used in this study. Label distributions are ordered as Supports/Refutes/NEI for DF-Coref and non-hallucination/hallucination for FD-Random and FD-Topic; Ref. men./turns denotes the mean number of annotated referent mentions and referent-bearing turns per example. Table 2 summarizes the independent second-review outcomes.

4.1 DF-Coref

Selection rule. We derive DF-Coref-Valid and DF-Coref-Test from the original DialFact validation and test splits by restricting attention to single-pronoun cases in which the response-side reference is unambiguously recoverable from the prior dialogue context (Gupta et al., 2022). This restriction supports reliable antecedent annotation and matched masking for the audit. DF-Coref should therefore be understood as a targeted pronoun-grounding audit set rather than a general coreference benchmark.

A *clear textual antecedent* is one uniquely recoverable from prior context without relying on response-internal content or outside world knowledge. Multiple mentions of the same entity are

allowed, but multiple plausible antecedents are excluded. For *they*, we require only that the intended antecedent be recoverable from context. We exclude pleonastic or non-referential uses of *it*, cases with no antecedent in prior dialogue context, and cases where the antecedent appears only in the response.

Audit annotation. Each retained example is manually annotated with three minimal additions: the response-side pronoun span, all non-overlapping context mentions of the same discourse referent, and the numeric form of the original three-way label. Annotating all visible context mentions is necessary because PPS later masks all context-side referent cues. Example audit records are provided in Appendix Tables A8–A9.

Adjudication. Each DF-Coref file first received an initial annotation pass, and a senior researcher who did not produce those annotations independently reviewed a fixed stratified sample, stratified by `data_type`, `response_label`, and `type_label`, as a second-review quality check. Each sampled item was labeled as correct, wrong, or ambiguous. Cases with an alternative valid span were counted as correct unless the original annotation violated the selection rule or identified the wrong discourse entity.

Only wrong and ambiguous cases triggered adjudication. This second pass followed a fixed decision rule rather than open-ended reannotation, because open-span antecedent annotation can admit multiple acceptable surface mentions of the same discourse referent.

4.2 FD-Random and FD-Topic

Role and label mapping. We use FaithDial as external corroboration rather than as a second main benchmark. FD-Random asks whether the audited pattern recurs in the seen/random condition, whereas FD-Topic asks whether it survives the harder unseen/topic condition inherited from Wizard of Wikipedia (Dziri et al., 2022a; Dinan et al., 2019).

To align FaithDial with the verifier setting in § 3, we map dialogue history to C, the audited response text to R, the gold knowledge snippet to E, and a binary label y. We derive y from `BEGIN`: an instance is labeled `hallucination` if `BEGIN` contains `Hallucination`, and `non-hallucination` otherwise.

Construction and adjudication. For the audited response text, we use `original_response` when available and otherwise fall back to `response`. We then apply the same response-side pronoun criterion and antecedent annotation guideline as in DF-Coref.

Each FaithDial file was checked with the same second-review setup: the senior coauthor independently reviewed a stratified sample of each file, again stratified by `data_type`, `response_label`, and `type_label`. Each sampled item was labeled as correct or wrong, with alternative valid spans counted as correct unless they violated the selection rule or identified the wrong discourse entity. In FD-Random, the 13 reviewed errors in Table 2 were resolved by antecedent edits or removals, depending on whether a clear context antecedent remained after applying the selection rule. In FD-Topic, the three reviewed errors were resolved under the same existing guideline and did not require any change to the annotation policy. All reported FD-Random and FD-Topic experiments use the consolidated post-adjudication files.

Across the reviewed samples reported in Table 2, 319/348 items (91.7%) were judged correct under the selection rule used in this study.

5 Experiments and Analysis

5.1 Experimental Setup

All verifiers receive the same joint verification input: paired evidence on the premise side and the context-conditioned response on the hypothesis side, using all dialogue turns and the first five evidence items. Each model keeps its native tokenizer and fixed input cap. The six-model panel should therefore be read as a controlled audit under stable model-native settings, not as an equal-context-budget ranking.

We restrict the comparison to encoder-only verifiers because PPS and the intervention-selection analysis rely on comparable class-margin outputs under masked or transformed inputs. Detailed checkpoint specifications and tokenizer-overflow counts are reported in Appendix Tables A3–A4; overflow is near zero for all models except DeBERTa-v3-B.

For the cross-set audit in § 5.2, fine-tuning is source-specific: DF-Coref results use a rebuilt DialFact fine-tuning pool that excludes DF-Coref-Valid, whereas FD-Random and FD-Topic use the FaithDial training file and the native binary

Model	Post-FT Macro-F1			Post-FT PPS gap		
	DF-Coref Test	FD-Random	FD-Topic	DF-Coref Test	FD-Random	FD-Topic
NLI-pretrained						
ModernBERT-L	86.6 (+25.7)	81.9 (+54.9)	85.4 (+59.5)	+0.81 [0.13, 1.55]	0.00 [-1.88, 1.90]	+0.42 [-0.92, 1.80]
DeBERTa-B-long	69.0 (+1.3)	73.1 (+51.1)	72.4 (+51.8)	+0.56 [0.05, 1.05]	+0.63 [-0.36, 1.63]	+0.38 [-0.87, 1.83]
DeBERTa-v3-L	84.1 (+20.4)	77.9 (+47.5)	82.7 (+52.5)	+0.64 [-0.22, 1.44]	+0.88 [-0.44, 2.22]	+1.81 [-0.33, 3.77]
Plain base						
ModernBERT-B	78.7 (+59.7)	79.9 (+31.0)	79.5 (+31.9)	+0.64 [0.29, 1.01]	+0.32 [-1.08, 2.16]	+0.41 [-0.58, 1.62]
DeBERTa-v3-B	57.8 (+42.0)	80.7 (+59.2)	79.4 (+60.3)	+0.26 [0.02, 0.52]	+0.08 [-0.38, 0.56]	+0.57 [0.08, 1.10]
RoBERTa-B	75.9 (+60.1)	80.5 (+38.4)	79.3 (+36.0)	+0.72 [0.17, 1.41]	+0.73 [-0.73, 2.21]	+0.65 [0.16, 1.22]

Table 3: Cross-set audit results on DF-Coref-Test, FD-Random, and FD-Topic. Macro-F1 cells report full-file post-FT Macro-F1 with ΔMacro-F1 (post-FT − pre-FT) in parentheses, and PPS-gap cells report $\mathrm{PPS}_{\mathrm{NF}} - \mathrm{PPS}_{\mathrm{PNF}}$ on the joint-diagnosable subset with 95% bootstrap confidence intervals in brackets. The transition counts and shares underlying these PPS denominators are shown in Figure A1 and Appendix Table A5. Positive gaps indicate that negative flips (NF; correct→wrong) show stronger premise-side sensitivity than positive no flips (PNF; correct→correct) under PPS.

FaithDial label space. Within each source-specific fine-tuning pool, we form a stratified 90/10 train–checkpoint-selection split with the reported seed; this checkpoint-selection split is separate from all audited evaluation files. We train for at most three epochs with the Hugging Face `Trainer`, evaluate after each epoch, and select the post-FT checkpoint by Macro-F1 on the checkpoint-selection split. For DialFact, early stopping is disabled and the selected after-FT checkpoint is chosen by a post-hoc sweep over the saved fine-tuned epoch checkpoints; for FaithDial, `load_best_model_at_end` is enabled with early-stopping patience 2.

Optimization uses AdamW-style training with weight decay 0.01 and warmup ratio 0.06. DialFact runs use learning rate 3×10^{-5}; FaithDial runs use 2×10^{-5} for NLI-pretrained checkpoints and 3×10^{-5} for plain-base checkpoints. Micro-batch sizes are selected before training from fixed OOM-safe candidate lists on the training data, with the resulting training plan logged in the run metadata; these choices are not tuned on the audit or evaluation files.

For NLI-adapted checkpoints, pre-FT evaluation uses raw zero-shot NLI; for plain-base checkpoints, it uses a random-head diagnostic from a freshly initialized classifier head. Main tables use the fixed seed-7 run set, and Appendix Table A6 reports the additional-seed 42 rerun; other exploratory seeds are not used in the reported tables.

5.2 Main Audit Results: Cross-Set Hidden Grounding Regression

Table 3 reports the main audit result on DF-Coref-Test, FD-Random, and FD-Topic. Under this matched fine-tuning setup, fine-tuning improves Macro-F1 throughout the six-model/three-evaluation-set panel, while the post-FT NF–PNF PPS gap is positive in 17 of 18 pairs and essentially zero in the remaining one. These results show the central pattern: aggregate Macro-F1 improves, while NF cases show stronger premise-side sensitivity than PNF cases on the dialogue-grounding audit. A validated seed-42 rerun over the same 18 model–evaluation-set pairs shows the same directional pattern (Appendix Table A6), supporting the robustness of the main audit finding.

The clearest evidence comes from DF-Coref-Test, the in-domain evaluation file most tightly matched to the audit design. It is explicitly constructed for response-side pronoun grounding, and its annotation policy matches the conservative PPS operationalization used in the main text. We therefore place the main evidential weight on this in-domain held-out file, using FD-Random and FD-Topic as supplementary out-of-domain checks.

The qualitative pattern behind this gap is the one illustrated in Figure 1: the response pronoun has a context-fixed referent, while the paired evidence contains a lexically plausible competing cue. In negative flips, the post-FT model changes from a previously correct decision to a wrong label in this setting, and a higher PPS indicates stronger relative sensitivity to the premise-side cue under PPS.

We use the example only to make the group-level

	Held-out NF recovery		Held-out PNF harm	
Subset	Single best	Stratified	Single best	Stratified
High	10/84	8/84	33/241	32/241
Mid	3/38	3/38	32/256	11/256
Low	4/41	6/41	17/207	8/207
Overall	17/163	17/163	82/704	51/704

Table 4: Held-out NF recovery and PNF harm on DF-Coref-Test for *Single best* and *Stratified* selection over the fixed intervention family defined in § 5.3.

NF–PNF comparison interpretable; the evidence for the claim remains the aggregate transition-conditioned analysis. The NLI-adapted checkpoints show the same directional pattern in almost all settings, and the plain-base checkpoints reproduce it consistently as well. Among the two out-of-domain conditions, FD-Topic provides the more consistent signal, whereas FD-Random is weaker and includes the near-zero exception.

Because some confidence intervals still cross zero, we do not present the observed NF–PNF separation as a formal causal estimate of fine-tuning or as a universal property of fine-tuned verifiers. Instead, we interpret it as a repeated descriptive pattern on a controlled referent-sensitive slice: fine-tuning can improve aggregate verification accuracy while still making newly regressed cases relatively more premise-side sensitive under PPS than stable-correct cases.

5.3 Exploratory Ablation: PPS-Stratified Intervention Selection

We include a small auxiliary ablation over the same five oracle, post-hoc input transformations for every model: no change, response-side gold-referent clarification (*pronoun=referent*), response-side pronoun replacement with the canonical gold referent, context-side appending of the same clarification, and the combined response-replacement/context-clarification variant. These transformations use referent annotations only at analysis time, are applied to the inputs of the already fine-tuned verifier without further parameter updates, and are not proposed as a repair method.

For each model, we compute High/Mid/Low PPS tertile thresholds on DF-Coref-Valid and transfer the selected intervention choices unchanged to DF-Coref-Test. We compare two selection settings over this fixed family: *Single best*, which chooses one intervention per model on DF-Coref-Valid, and *Stratified*, which chooses interventions separately within the three DF-Coref-Valid PPS tertiles. Table 4 reports the held-out NF recovery and PNF harm for these two settings; denominators aggregate transition-conditioned NF and PNF cases over the six model runs on the DF-Coref-Test joint-diagnosable subset, which contains 305 examples per model before transition conditioning.

Stratified matches Single best on NF recovery (17/163) but reduces PNF harm from 82/704 to 51/704, with this reduction concentrated in the Mid and Low tertiles rather than the High tertile. In this narrow setting, PPS provides only a limited stratification signal for reducing collateral harm; it does not support a monotonic intervention rule or a broadly transferable repair strategy.

Because selection on DF-Coref-Valid optimizes NF recovery rather than PNF harm, we treat this as a DF-Coref-Test transfer ablation rather than as a tuned utility objective. The effect remains limited and model-dependent; when avoiding PNF harm is the priority, the no-intervention baseline is still safest. Model-wise breakdowns and oracle upper bounds are given in Appendix Table A7.

6 Conclusions

This paper asked whether better benchmark performance after fine-tuning in DFV necessarily reflects better dialogue-conditioned grounding. Across six models and three evaluation sets, Macro-F1 improved, yet the post-FT NF–PNF PPS gap was positive in 17 of 18 comparisons. This supports a hidden-grounding-regression pattern: aggregate gains can mask newly regressed cases with stronger premise-side sensitivity under PPS, especially on DF-Coref-Test.

The finding matters for dialogue research because DFV labels can depend on cross-turn reference resolution, not only on claim–evidence compatibility. In practice, DFV model selection should not rely on benchmark gains alone; fine-tuned verifiers should also be checked on controlled referent-sensitive audits. The auxiliary PPS-stratified ablation reduced collateral PNF harm without improving NF recovery, suggesting limited diagnostic value rather than a repair method. Future work should test referent-aware supervision, speaker/entity-chain cues, and context-use regularization so that aggregate gains are not selected at the expense of cross-turn grounding.

7 Limitations

Our claims concern a conservative, pronoun-resolved, surface-recoverable slice of dialogue fact verification rather than dialogue grounding in full generality. PPS is computed only when the intended referent is visible in the dialogue context, a matching cue is available on the premise side, and the required masked/control variants remain well-formed. The paper therefore establishes the existence and recurrence of hidden grounding regressions on this audited slice, but does not estimate their prevalence over full benchmark distributions or over other forms of referring expression.

The audited dialogues are also short relative to long multi-session dialogue settings. As shown in Table 1, the audited files average only a few turns per example, so they should not be treated as a proxy for long-context DFV. Longer conversations may introduce more competing referents, topic shifts, and accumulated ambiguity; the prevalence of hidden grounding regressions and the behavior of PPS may therefore differ in long-dialogue settings.

A second limitation concerns PPS itself. Under our operationalization, PPS is a within-example, control-adjusted diagnostic of relative premise-versus-context sensitivity. It should not be interpreted as a full causal account of entity use, a general attribution method, or a standalone coreference metric.

A third limitation concerns experimental scope. We study encoder-only discriminative verifiers because PPS relies on comparable class margins under masked variants, and the intervention-selection analysis likewise compares class-score changes under post-hoc textual input transformations.

Prompted or decoder-only LLM verifiers would require a different margin or log-probability extraction protocol, and prompting, instruction tuning, RLHF, and generation-time behavior may change how context is used. We therefore do not claim that the same prevalence or mitigation behavior transfers to generative LLM settings, and we place the main evidential weight on within-model pre/post contrasts rather than on fine-grained cross-model ranking.

Finally, the intervention-selection analysis is intentionally narrow. It is evaluated only on DF-Coref-Test and within a small, interpretable intervention family. PPS-Stratified Selection shows localized operational value in this setup, but should not be treated as a broadly transferable repair strategy.

Data availability. The audited examples are derived from third-party dialogue datasets and include an annotation layer constructed for this study. This version reports the annotation protocol, aggregate statistics, second-review summaries, and representative audit records, but does not release the full audit files. The audit files are used here as an internal evaluation resource for the reported experiments, not as a released benchmark.

References

Eric Chamoun, Marzieh Saeidi, and Andreas Vlachos. 2023. Automated fact-checking in dialogue: Are specialized models needed? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16009–16020, Singapore. Association for Computational Linguistics.

Sihao Chen, Chaitanya Malaviya, Alex Fabrikant, Hagai Taitelbaum, Tal Schuster, Senaka Buthpitiya, and Dan Roth. 2025. On reference (in-)determinacy in natural language inference. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8081–8093, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhenyun Deng, Michael Schlichtkrull, and Andreas Vlachos. 2024. Document-level claim extraction and decontextualisation for fact-checking. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11943–11954, Bangkok, Thailand. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. *Preprint*, arXiv:1811.01241.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M. Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Preprint*, arXiv:2204.10757.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022b. Evaluating attribution in dialogue systems: The BEGIN benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Vagrant Gautam, Eileen Bingert, Dawei Zhu, Anne Lauscher, and Dietrich Klakow. 2024. Robust pronoun fidelity with English LLMs: Are they reasoning, repeating, or just biased? *Transactions of the Association for Computational Linguistics*, 12:1755–1779.

Abbas Ghaddar, David Alfonso-Hermelo, Philippe Langlais, Mehdi Rezagholizadeh, Boxing Chen, and Prasanna Parthasarathi. 2024. Charp: Conversation history awareness probing for knowledge-grounded dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1534–1551.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Sachin Goyal, Christina Baek, J. Zico Kolter, and Aditi Raghunathan. 2025. Context-parametric inversion: Why instruction finetuning can worsen context reliance. *Preprint*, arXiv:2410.10796.

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.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. DialFact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801, Dublin, Ireland. Association for Computational Linguistics.

Hongyu Li, Liang Ding, Meng Fang, and Dacheng Tao. 2024. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4297–4308, Miami, Florida, USA. Association for Computational Linguistics.

Elita Lobo, Chirag Agarwal, and Himabindu Lakkaraju. 2025. On the impact of fine-tuning on chain-of-thought reasoning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11679–11698, Albuquerque, New Mexico. Association for Computational Linguistics.

Mamta and Oana Cocarascu. 2025. FactEval: Evaluating the robustness of fact verification systems in the era of large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10647–10660, Albuquerque, New Mexico. Association for Computational Linguistics.

Kawshik Manikantan, Makarand Tapaswi, Vineet Gandhi, and Shubham Toshniwal. 2025. IdentifyMe: A challenging long-context mention resolution benchmark for LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 768–777, Albuquerque, New Mexico. Association for Computational Linguistics.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6996–7045, Vienna, Austria. Association for Computational Linguistics.

Hyunkyung Park and Arkaitz Zubiaga. 2026. Bicongate: Consistency-gated de-colloquialisation for dialogue fact-checking. In *Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER)*, pages 59–73. Association for Computational Linguistics.

Gorjan Radevski, Kiril Gashteovski, Shahbaz Syed, Christopher Malon, Sebastien Nicolas, Chia-Chien Hung, Timo Sztyler, Verena Heußer, Wiem Ben Rim, Masafumi Enomoto, Kunihiro Takeoka, Masafumi Oyamada, Goran Glavaš, and Carolin Lawrence. 2025. On synthesizing data for context attribution in question answering. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16929–16950, Vienna, Austria. Association for Computational Linguistics.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Evgeniia Razumovskaia, Ivan Vulić, Pavle Marković, Tomasz Cichy, Qian Zheng, Tsung-Hsien Wen, and Paweł Budzianowski. 2024. Dial BeInfo for faithfulness: Improving factuality of information-seeking dialogue via behavioural fine-tuning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17139–17152, Miami, Florida, USA. Association for Computational Linguistics.

Ieva Staliunaite and Andreas Vlachos. 2025. Dis2Dis: Explaining ambiguity in fact-checking. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 246–267, Albuquerque, New Mexico. Association for Computational Linguistics.

Yufei Tao and Ameeta Agrawal. 2026. No-worse context-aware decoding: Preventing neutral regression in context-conditioned generation. *Preprint*, arXiv:2604.16686.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

Papers), pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Zineddine Tighidet, Jiali Mei, Benjamin Piwowarski, and Patrick Gallinari. 2024. Probing language models on their knowledge source. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 604–614, Miami, Florida, US. Association for Computational Linguistics.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2025. DnDScore: Decontextualization and decomposition for factuality verification in long-form text generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23609–23626, Suzhou, China. Association for Computational Linguistics.

Yumo Xu, Peng Qi, Jifan Chen, Kunlun Liu, Rujun Han, Lan Liu, Bonan Min, Vittorio Castelli, Arshit Gupta, and Zhiguo Wang. 2025. CiteEval: Principle-driven citation evaluation for source attribution. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32759–32778, Vienna, Austria. Association for Computational Linguistics.

Junjie Ye, Yuming Yang, Yang Nan, Shuo Li, Qi Zhang, Tao Gui, Xuanjing Huang, Peng Wang, Zhongchao Shi, and Jianping Fan. 2025. Analyzing the effects of supervised fine-tuning on model knowledge from token and parameter levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 471–513, Suzhou, China. Association for Computational Linguistics.

Yingming Zheng, Hanqi Li, Kai Yu, and Lu Chen. 2025. When long helps short: How context length in supervised fine-tuning affects behavior of large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10293–10308, Suzhou, China. Association for Computational Linguistics.

A Appendix

A.1 Audit dataset summaries

Split	Type	#	S	R	NEI	Ev.Item	Turns(C)	Ref.Men.	Ref.Turns
Original full splits									
Valid	factual	8,691	3,342	3,363	1,986	1.13	4.58	-	-
	personal	1,745	0	0	1,745	1.01	4.36	-	-
	total	10,436	3,342	3,363	3,731	1.11	4.54	-	-
Test	factual	10,420	3,939	3,935	2,546	1.32	4.35	-	-
	personal	1,389	0	0	1,389	1.10	3.80	-	-
	total	11,809	3,939	3,935	3,935	1.29	4.28	-	-
DF-Coref audit files									
Coref-Valid	factual	409	176	163	70	1.10	4.37	1.76	1.66
	personal	58	0	0	58	1.00	4.67	2.26	1.98
	total	467	176	163	128	1.09	4.41	1.82	1.70
Coref-Test	factual	287	115	102	70	1.09	3.56	1.57	1.49
	personal	28	0	0	28	1.00	3.54	1.71	1.57
	total	315	115	102	98	1.09	3.56	1.59	1.50

Table A1: Original DialFact splits and the corresponding post-adjudication DF-Coref files. In the audited block, Coref-Valid corresponds to DF-Coref-Valid (derived from the original DialFact validation split) and Coref-Test corresponds to DF-Coref-Test (derived from the original DialFact test split). Ref.Men. and Ref.Turns denote annotated referent mentions and referent-bearing turns per example.

File	#	NonHall	Hall	KnowTok	Turns(H)	orig_empty	OrigUsed	RespUsed	Ref.Men.	Ref.Turns
Train	10,000	2,643	7,357	22.56	4.55	1,489	8,511	1,489	-	-
FD-Topic	483	114	369	22.74	4.33	48	435	48	1.00	1.00
FD-Random	405	111	294	22.60	4.47	88	317	88	1.00	1.00

Table A2: Processed FaithDial files. Train is the original FaithDial training split shown only for reference; the audited evaluation files used in the paper are FD-Topic and FD-Random. `orig_empty`, `OrigUsed`, and `RespUsed` indicate missing `original_response` and the source of the audited claim text.

A.2 Model protocol and tokenizer-overflow audit

Model	Checkpoint	Pre-FT	FT setup	Max len / cap
NLI-pretrained				
ModernBERT-L	`modernbert-large-nli`	zero-shot	native head / IDs	2048 / 1024
DeBERTa-B-long	`deberta-base-long-nli`	zero-shot	native head / IDs	1280 / 1024
DeBERTa-v3-L	`deberta-v3-large-mnli-fever-anli-ling-wanli`	zero-shot	native head / IDs	512 / 512
Plain base				
ModernBERT-B	`modernbert-base`	rand. head	new head / DialFact IDs	8192 / 1024
DeBERTa-v3-B	`deberta-v3-base`	rand. head	new head / DialFact IDs	512 / 512
RoBERTa-B	`roberta-base`	rand. head	new head / DialFact IDs	512 / 512

Table A3: Checkpoints, pre-FT semantics, DialFact FT label interface, and sequence-length limits. Checkpoint entries are the exact model keys/local checkpoint directory names resolved by the released configuration; tokenizers are loaded from the same checkpoints, with no additional lowercasing or tokenizer substitution. "Max len / cap" reports the model-side maximum input length shown for reference and the sequence cap actually used for fine-tuning and evaluation. *NLI-pretrained* models are evaluated before FT with their native zero-shot NLI heads, and DialFact FT preserves the native 3-way label order by mapping DialFact gold labels to native IDs. *Plain base* models are evaluated before FT with a randomly initialized head, and DialFact FT uses a new 3-way head with DialFact label IDs. FaithDial uses the same checkpoints and sequence caps, but fine-tunes in the native binary FaithDial label space.

Model	DialFact				FaithDial	
	Train pool (n=9958)	DialFact Test (n=11809)	DF-Coref Valid (n=467)	DF-Coref Test (n=315)	FD-Random (n=405)	FD-Topic (n=483)
NLI-pretrained						
ModernBERT-L	0	0	0	0	0	0
DeBERTa-B-long	0	0	0	0	0	0
DeBERTa-v3-L	0	11 (0.09%)	0	2 (0.63%)	0	0
Plain base						
ModernBERT-B	0	0	0	0	0	0
DeBERTa-v3-B	719 (7.22%)	1042 (8.82%)	24 (5.14%)	14 (4.44%)	29 (7.16%)	23 (4.76%)
RoBERTa-B	4 (0.04%)	9 (0.08%)	0	2 (0.63%)	0	0

Table A4: Counts and percentages of examples whose serialized input exceeds each model's chosen cap under its native tokenizer. Train pool denotes the final DialFact fine-tuning pool used in the reported runs; it excludes DF-Coref-Valid and reflects the same preprocessing filters used before training.

A.3 Supplementary result breakdowns

Model	Evaluation set	Macro-F1		Transition share (%)				Post-FT PPS				
		Post	Δ	NNF (W→W)	PF (W→C)	NF (C→W)	PNF (C→C)	NNF	PF	NF	PNF	NF–PNF gap
NLI-pretrained												
ModernBERT-L	DF-Coref-Test	86.6	+25.7	8.2	30.2	5.2	56.4	-0.18	-0.27	+0.34	-0.47	+0.81
	FD-Random	81.9	+54.9	4.3	66.4	7.8	21.5	+0.66	-0.07	-0.26	-0.26	0.00
	FD-Topic	85.4	+59.5	6.5	63.7	5.1	24.7	+2.28	+0.14	+0.03	-0.39	+0.42
DeBERTa-B-long	DF-Coref-Test	69.0	+1.3	17.4	15.7	12.5	54.4	-0.34	-1.22	-0.02	-0.57	+0.56
	FD-Random	73.1	+51.1	12.1	61.3	8.2	18.4	-0.41	+0.03	-0.34	-0.97	+0.63
	FD-Topic	72.4	+51.8	11.2	62.8	9.8	16.3	+1.03	-0.05	-0.52	-0.90	+0.38
DeBERTa-v3-L	DF-Coref-Test	84.1	+20.4	7.9	27.5	7.9	56.7	-0.44	-1.85	+0.07	-0.57	+0.64
	FD-Random	77.9	+47.5	2.7	65.6	9.8	21.9	+0.58	+0.04	-0.04	-0.92	+0.88
	FD-Topic	82.7	+52.5	4.7	64.2	8.4	22.8	+3.64	+0.05	+0.34	-1.46	+1.81
Plain base												
ModernBERT-B	DF-Coref-Test	78.7	+59.7	14.4	54.8	6.6	24.3	+0.04	-0.40	+0.24	-0.40	+0.64
	FD-Random	79.9	+31.0	8.6	24.0	3.9	63.4	-0.13	-0.50	+0.27	-0.05	+0.32
	FD-Topic	79.5	+31.9	8.4	20.9	6.5	64.2	-0.14	-0.12	+0.47	+0.06	+0.41
DeBERTa-v3-B	DF-Coref-Test	57.8	+42.0	27.9	42.0	12.1	18.0	-0.05	-0.20	+0.35	+0.09	+0.26
	FD-Random	80.7	+59.2	5.5	67.6	8.2	18.8	+1.17	+0.05	-0.41	-0.48	+0.08
	FD-Topic	79.4	+60.3	6.5	68.8	7.9	16.7	+0.93	+0.01	-0.08	-0.65	+0.57
RoBERTa-B	DF-Coref-Test	75.9	+60.1	15.1	54.8	9.2	21.0	+0.10	-0.73	+0.73	+0.01	+0.72
	FD-Random	80.5	+38.4	9.8	17.2	3.1	69.9	+0.35	-0.17	+0.68	-0.05	+0.73
	FD-Topic	79.3	+36.0	8.8	15.8	8.4	67.0	+0.18	-0.15	+0.57	-0.08	+0.65

Table A5: Full cross-benchmark results across the three referent-annotated evaluation sets. Macro-F1 is computed on the full post-adjudication evaluation file; transition shares and transition-wise post-FT PPS are computed on the fixed joint-diagnosable subset used for PPS. NF–PNF gap denotes $\mathrm{PPS}_{\mathrm{NF}} - \mathrm{PPS}_{\mathrm{PNF}}$, and transition definitions follow Section 3. Positive NF–PNF gaps appear in 17 of 18 pairs, and the remaining estimate is 0.00.

Model	Evaluation set	Macro-F1		Transition share (%)				Post-FT PPS				
		Post	Δ	NNF (W→W)	PF (W→C)	NF (C→W)	PNF (C→C)	NNF	PF	NF	PNF	NF–PNF gap
NLI-pretrained												
ModernBERT-L	DF-Coref-Test	87.2	+27.0	7.9	30.8	4.9	56.4	-0.57	-0.60	+0.45	-0.38	+0.83
	FD-Random	83.9	+56.8	5.1	65.6	5.1	24.2	-1.11	+0.27	-0.03	-0.17	+0.14
	FD-Topic	84.5	+58.6	7.0	63.3	6.5	23.3	+2.13	+0.26	+0.22	-0.92	+1.14
DeBERTa-B-long	DF-Coref-Test	70.3	+2.6	15.7	17.4	12.8	54.1	-0.17	-1.31	+0.14	-0.49	+0.63
	FD-Random	63.5	+41.5	26.2	47.3	6.2	20.3	-0.15	-0.06	+0.01	-0.65	+0.66
	FD-Topic	65.4	+44.8	25.1	48.8	6.5	19.5	+0.30	-0.16	+0.22	-0.48	+0.70
DeBERTa-v3-L	DF-Coref-Test	82.8	+19.2	8.9	26.6	8.5	56.1	-0.23	-2.17	+0.56	-0.73	+1.29
	FD-Random	78.6	+48.3	5.9	62.5	8.6	23.0	+1.87	+0.16	+0.35	-0.90	+1.25
	FD-Topic	82.2	+52.0	6.5	62.3	6.0	25.1	+2.59	+0.04	+0.90	-1.11	+2.01
Plain base												
ModernBERT-B	DF-Coref-Test	74.2	+57.9	18.0	49.2	8.2	24.6	+0.05	-0.25	+0.14	-0.15	+0.29
	FD-Random	84.0	+42.1	7.0	20.7	2.7	69.5	+0.32	-0.59	+0.72	-0.07	+0.79
	FD-Topic	81.3	+38.0	7.9	16.7	4.7	70.7	+0.01	-0.69	+1.01	-0.03	+1.05
DeBERTa-v3-B	DF-Coref-Test	61.0	+44.7	17.0	50.2	20.7	12.1	-0.01	-0.09	+0.01	-0.13	+0.14
	FD-Random	75.7	+33.7	9.0	18.0	7.0	66.0	-0.12	-0.14	+0.07	+0.04	+0.03
	FD-Topic	80.7	+37.4	6.5	18.1	4.7	70.7	-0.10	-0.16	+0.25	+0.00	+0.25
RoBERTa-B	DF-Coref-Test	77.1	+59.3	16.4	46.6	6.6	30.5	+0.50	-0.03	-0.51	-1.14	+0.64
	FD-Random	82.2	+60.6	4.3	68.8	8.6	18.4	-0.20	-0.02	+0.23	+0.12	+0.11
	FD-Topic	81.8	+62.7	6.0	69.3	6.0	18.6	+0.53	-0.05	+0.36	-0.01	+0.37

Table A6: Validated seed-42 cross-benchmark rerun results under the same denominator rules as Appendix Table A5. NNF/PF post-FT PPS values are recovered from the archived per-run transition summaries, and NF–PNF gap denotes $\mathrm{PPS}_{\mathrm{NF}} - \mathrm{PPS}_{\mathrm{PNF}}$ with transition definitions as in Section 3. The NF–PNF gap remains positive in all 18 model–evaluation-set pairs.

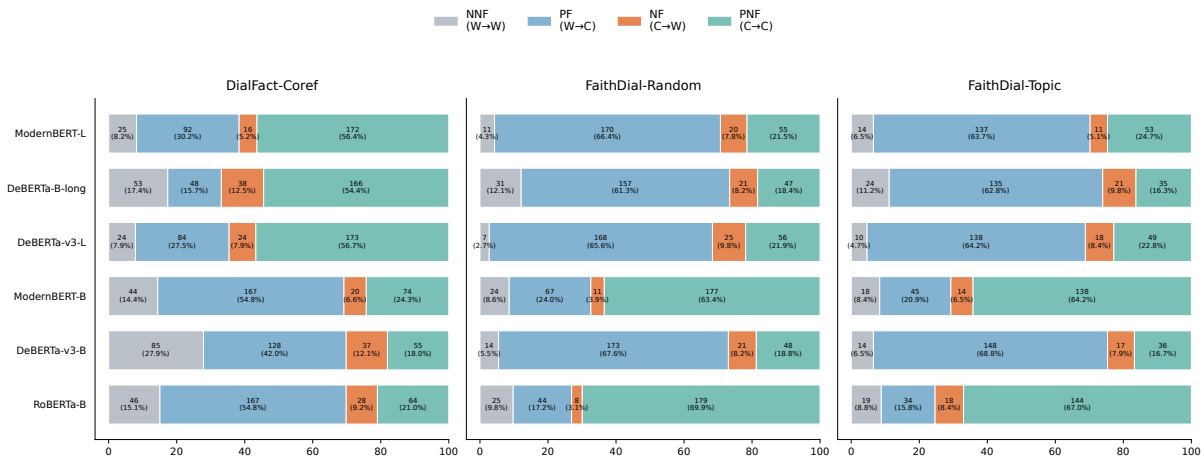

Figure A1: Before→after transition shares across models and referent-annotated evaluation sets. Bars sum to 100% and are partitioned into NNF (W→W), PF (W→C), NF (C→W), and PNF (C→C); labels report counts and within-slice percentages.

Model	Subset	Held-out NF recovery				Held-out PNF harm		
		n	Single best	Stratified	Oracle	n	Single best	Stratified
NLI-pretrained								
ModernBERT-L	Overall	16	1/16	1/16	3/16	172	11/172	11/172
	High PPS	8	0/8	0/8	1/8	47	7/47	7/47
	Mid PPS	1	0/1	0/1	0/1	68	0/68	0/68
	Low PPS	7	1/7	1/7	2/7	57	4/57	4/57
DeBERTa-B-long	Overall	38	3/38	4/38	6/38	166	65/166	29/166
	High PPS	16	2/16	2/16	4/16	58	23/58	23/58
	Mid PPS	14	0/14	1/14	1/14	60	30/60	3/60
	Low PPS	8	1/8	1/8	1/8	48	12/48	3/48
DeBERTa-v3-L	Overall	24	4/24	3/24	5/24	173	3/173	9/173
	High PPS	12	1/12	1/12	2/12	68	1/68	1/68
	Mid PPS	6	1/6	0/6	1/6	59	1/59	7/59
	Low PPS	6	2/6	2/6	2/6	46	1/46	1/46
Plain base								
ModernBERT-B	Overall	20	5/20	7/20	10/20	74	0/74	0/74
	High PPS	11	3/11	3/11	5/11	19	0/19	0/19
	Mid PPS	5	2/5	2/5	3/5	20	0/20	0/20
	Low PPS	4	0/4	2/4	2/4	35	0/35	0/35
DeBERTa-v3-B	Overall	37	2/37	2/37	3/37	55	1/55	1/55
	High PPS	22	2/22	2/22	2/22	26	1/26	1/26
	Mid PPS	5	0/5	0/5	0/5	17	0/17	0/17
	Low PPS	10	0/10	0/10	1/10	12	0/12	0/12
RoBERTa-B	Overall	28	2/28	0/28	2/28	64	2/64	1/64
	High PPS	15	2/15	0/15	2/15	23	1/23	0/23
	Mid PPS	7	0/7	0/7	0/7	32	1/32	1/32
	Low PPS	6	0/6	0/6	0/6	9	0/9	0/9
Overall	Overall	163	17/163	17/163	29/163	704	82/704	51/704
	High PPS	84	10/84	8/84	16/84	241	33/241	32/241
	Mid PPS	38	3/38	3/38	5/38	256	32/256	11/256
	Low PPS	41	4/41	6/41	8/41	207	17/207	8/207

Table A7: Tertile-wise transfer results for PPS-Stratified Selection on DF-Coref-Test, reporting NF recovery and PNF harm for the Single best, Stratified, and Oracle settings. High/Mid/Low PPS tertiles are defined per model from PPS tertiles over all diagnosable examples in DF-Coref-Valid. Rows condition on transition type; the DF-Coref-Test PPS base contains 305 joint-diagnosable examples per model before transition conditioning. The clean baseline is omitted because it is 0/n in every subset by construction. Oracle reports the per-example NF recovery attainable by choosing the best candidate in hindsight within the same fixed family; it is an upper bound and is not used as a deployable selection rule.

A.4 Example audit records

Field	Validation example (id=97___4-3)	Test example (id=83___4-1)
context_id	97___4	83___4
id	97___4-3	83___4-1
data_type	generated	generated
type_label	factual	factual
context	[0] I think jazz music is awesome! Many see jazz as being America's classical music! [1] I love Jazz! Who's your favorite artist? [2] I love them all! I can't pick just one! The Jazz age was in the 1920s and is now recognized as a major form of musical expression! [3] Ah, do you know where jazz was invented?	[0] I try to use Instagram, but I just have the hardest time "getting" it. I think I'm too old. [1] Lol! When did instagram first come out? [2] it was launched in october 2010. I think it was mostly just for editing photos then. At least thats all I thought it was. [3] What else does instagram do besides edit photos?
response	It started in New Orleans, specifically in the pre - WWII zones	It has a bunch of other features too, like image sharing, commenting, trending, and a whole host of other things.
evidence_list	[0] title=Jazz; url=https://en.wikipedia.org/wiki/Jazz; sent=Jazz is a music genre that originated in the African-American communities of New Orleans, United States, in the late 19th and early 20th centuries, and developed from roots in blues and ragtime.; id=0 [1] title=; url=; sent=Jazz is a music genre that originated in the African-American communities of New Orleans, Louisiana, United States, in the late 19th and early 20th centuries, with its roots in blues and ragtime; id=extra [2] title=New Orleans; url=https://en.wikipedia.org/wiki/New_Orleans; sent=New Orleans (, or ;) is a major United States port and the largest city and metropolitan area in the state of Louisiana.; id=1	[0] title=Instagram; url=https://en.wikipedia.org/wiki/Instagram; sent=Instagram is a mobile, desktop, and Internet-based photo-sharing application and service that allows users to share pictures and videos either publicly, or privately to pre-approved followers.; id=0 [1] title=Instagram; url=https://en.wikipedia.org/wiki/Instagram; sent=The service also added messaging features, the ability to include multiple images or videos in a single post, and a 'stories' feature - similar to its main opposition Snapchat - which allows users to post photos and videos to a sequential feed, with each post accessible by others for 24 hours each; id=1 [2] title=Instagram; url=https://en.wikipedia.org/wiki/Instagram; sent=Users can like photos and follow other users to add their content to a personal feed; id=2
response_label	REFUTES	NOT ENOUGH INFO
(+) pronoun_info	surface=It span=[0,2]	surface=It span=[0,2]
(+) referent_info	mentions: [0] surface=jazz, span=[0,8,12] [1] surface=jazz, span=[0,41,45] [2] surface=Jazz, span=[1,7,11] [3] surface=Jazz, span=[2,46,50] [4] surface=jazz, span=[3,22,26]	mentions: [0] surface=Instagram, span=[0,13,22] [1] surface=instagram, span=[1,14,23] [2] surface=instagram, span=[3,15,24]
(+) label_id	1	2

Table A8: Example DF-Coref records in the audit record format, including pronoun_info, referent_info, and label_id.

Field	Topic example (id=0___2-0)	Random example (id=409___4-0)
id	0___2-0	409___4-0
history	[0] I love candy, what's a good brand? [1] I haven't got a clue how good they are, but Dylan's Candy Bar has a chain of candy shops in various cities. [2] Oh, they do? What kind of candy do they sell? [3] I'm probably not the best system to ask for that information, really, but they also are a supplier of candy. [4] Oh I see, what kind of candy do they offer?	[0] I like science fiction and alien movies and shows. Do those interest you? [1] As I am a bot I don't have preferences. But I know that sci-fi films uses depictions of phenomena based on fictional science not completely accepted by the mainstream science [2] I liked the Star Wars movies and the Alien movies were pretty intense too, [3] Yeah, did you know that star wars came out in 1977? [4] I didn't know it was that old. What else do you know abou it ? [5] Well, I know that it starred Harrison Ford, Mark Hamil, Carrie Fisher among other people [6] Very interesting. I think I saw almost all of those movies [7] Cool. Did you ever see The Empire Strikes Back in 1980? [8] It's been so long I need to see it again to remember what it was about.
knowledge	It stocks 7,000 candies from around the world.	The film is set three years after "Star Wars" .
original_response	it stocks over 7000 candies from across the world	It is set three years after the original Star Wars.
response	It stocks over 7,000 candies from across the world.	I see, Star wars is set 3 years prior to the film
BEGIN	['Entailment']	['Hallucination']
VRM	['Edification']	['Edification']
(+) pronoun_info	surface=it span=[0, 2]	surface=It span=[0, 2]
(+) referent_info	mentions: [0] surface=Dylan's Candy Bar, span=[1, 44, 61]	mentions: [0] surface=The Empire Strikes Back, span=[7, 23, 46]
(+) label_id	0	1

Table A9: Example FD-Topic and FD-Random records in the audit record format, including pronoun_info, referent_info, and label_id.