Better Scores, Worse Grounding

Hidden Regressions after Fine-Tuning in Dialogue Fact Verification

Hyunkyung Park | Arkaitz Zubiaga

Queen Mary University of London | {hyunkyung.park, a.zubiaga}@qmul.ac.uk

SIGDIAL 2026

POSTER Session 3

Dialogue Fact Verification Grounding Audit

Atlanta, Georgia

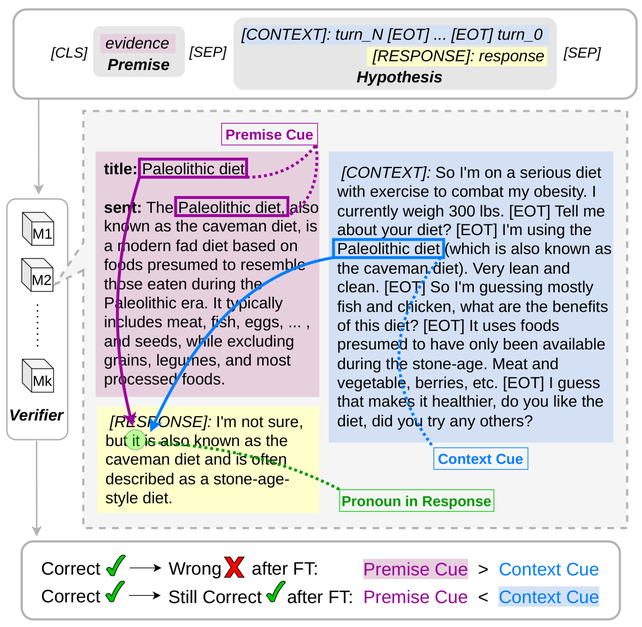
Fine-tuning improved Macro-F1 in all 18 comparisons, yet newly regressed cases showed greater relative premise-side sensitivity under PPS than stable-correct cases in 17 of 18.

18 / 18
Macro-F1 improved

17 / 18
positive NF - PNF PPS gap

1. A score gain can hide a grounding failure

Paper Figure 1: the verifier changes from the correct gold label to an incorrect label after fine-tuning.



1 Context fixes the referent
In the dialogue, "it" refers to the diet being discussed.

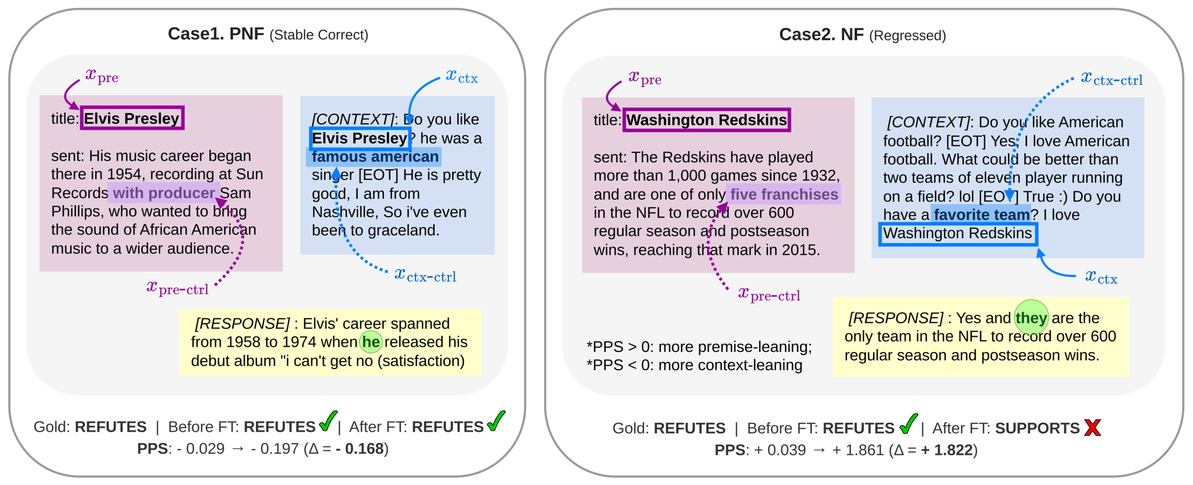
2 Evidence offers a rival cue
The premise repeats "Paleolithic diet", which is locally compatible with the response surface.

3 Fine-tuning flips the answer
The verifier moves from a correct REFUTES decision to an incorrect SUPPORTS decision.

This is a negative flip (NF): correct before FT -> wrong after FT.
The regression is "hidden" when gains on other cases raise the aggregate score.

2. Audit transitions, then measure cue sensitivity

Paper Figure 2: compare newly regressed cases (NF) with cases that remain correct (PNF) under the same masking diagnostic.



NF | correct -> wrong

PNF | correct -> correct

PPS = adjusted premise-mask drop - adjusted context-mask drop
PPS > 0: stronger premise-side sensitivity

PPS keeps the gold label fixed and subtracts same-side matched masking controls.

3. Controlled audit design

A deliberately narrow, pronoun-resolved, surface-recoverable slice - designed for auditability rather than full-benchmark coverage.

Audit files

File	n
DF-Coref-Valid	467
DF-Coref-Test	315
FD-Random	405
FD-Topic	483

Second review: 319 / 348 correct (91.7%)

Six encoder-only verifiers

NLI-pretrained
ModernBERT-L
DeBERTa-B-long
DeBERTa-v3-L

Plain base
ModernBERT-B
DeBERTa-v3-B
RoBERTa-B

Prediction transitions

Group	Before -> after
NNF	wrong -> wrong
PF	wrong -> correct
NF	correct -> wrong
PNF	correct -> correct

Main contrast: NF vs PNF

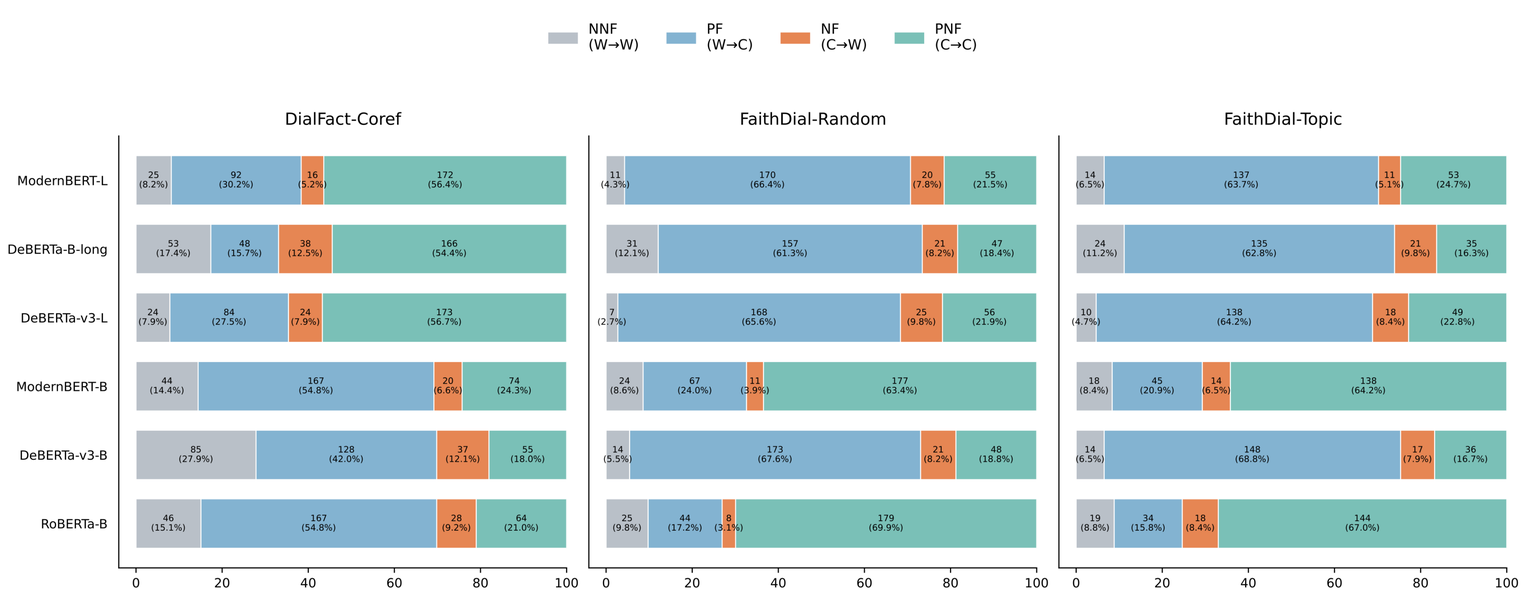
Matched evaluation

Source-specific fine-tuning
DialFact models -> DF-Coref
FaithDial models -> FD-Random / Topic

Reported signals
Macro-F1: full file
Transitions + PPS: fixed joint-diagnosable subset
Main seed 7 + validated seed 42 rerun

4. Aggregate gains coexist with negative flips

PF share > NF share in 18 / 18 slices



What the transition bars reveal

18 / 18
Macro-F1 gains

3.1-12.5%
NF share by slice

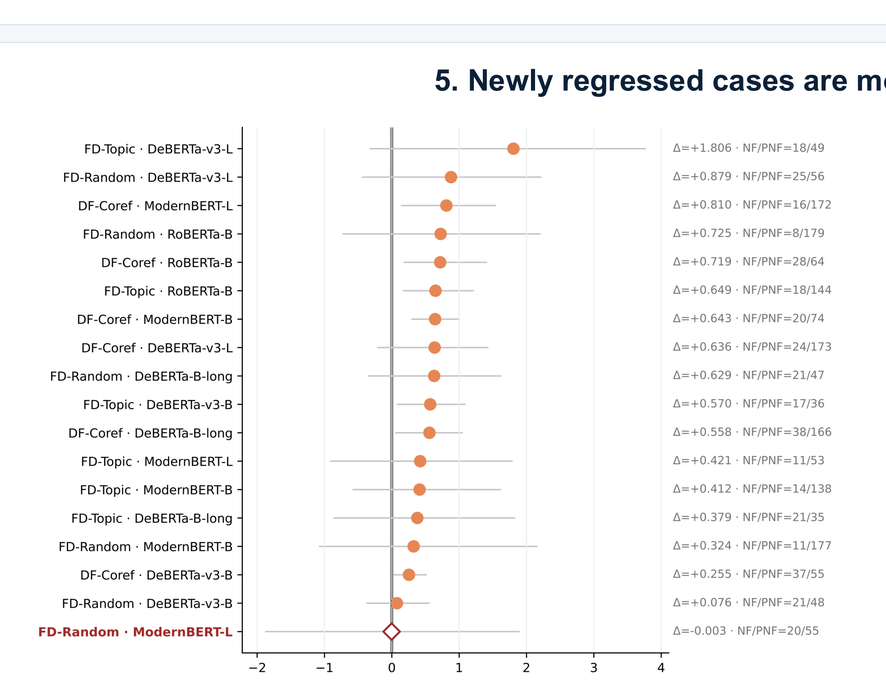
1 What the audit subset shows
Positive-flip (PF) share exceeds negative-flip (NF) share in every joint-diagnosable model-set slice.

2 Why the regression is hidden
Macro-F1 aggregates across the full file; it does not isolate cases that were correct before fine-tuning.

Bars reproduce the paper's transition-share analysis on the joint-diagnosable subset.

5. Newly regressed cases are more premise-sensitive in nearly every setting

Post-FT PPS gap = PPS(NF) - PPS(PNF)



Repeated pattern

17 / 18
positive in main run

6 / 6
positive on DF-Coref-Test

18 / 18
positive at seed 42

Some 95% CIs cross zero; the paper treats this as a recurring descriptive audit pattern, not a universal causal estimate.

Table 3 (paper): exact cross-set results

Post-FT Macro-F1 (Delta post - pre)

Model	DF-Coref-Test	FD-Random	FD-Topic
ModernBERT-L	86.6 (+25.7)	81.9 (+54.9)	85.4 (+59.5)
DeBERTa-B-long	69.0 (+1.3)	73.1 (+51.1)	72.4 (+51.8)
DeBERTa-v3-L	84.1 (+20.4)	77.9 (+47.5)	82.7 (+52.5)
ModernBERT-B	78.7 (+59.7)	79.9 (+31.0)	79.5 (+31.9)
DeBERTa-v3-B	57.8 (+42.0)	80.7 (+59.2)	79.4 (+60.3)
RoBERTa-B	75.9 (+60.1)	80.5 (+38.4)	79.3 (+36.0)

Post-FT NF - PNF PPS gap [95% bootstrap CI]

Model	DF-Coref-Test	FD-Random	FD-Topic
ModernBERT-L	+0.81 [0.13, 1.55]	0.00 [-1.88, 1.90]	+0.42 [-0.92, 1.80]
DeBERTa-B-long	+0.56 [0.05, 1.05]	+0.63 [-0.36, 1.63]	+0.38 [-0.87, 1.83]
DeBERTa-v3-L	+0.64 [-0.22, 1.44]	+0.88 [-0.44, 2.22]	+1.81 [-0.33, 3.77]
ModernBERT-B	+0.64 [0.29, 1.01]	+0.32 [-1.08, 2.16]	+0.41 [-0.58, 1.62]
DeBERTa-v3-B	+0.26 [0.02, 0.52]	+0.08 [-0.38, 0.56]	+0.57 [0.08, 1.10]
RoBERTa-B	+0.72 [0.17, 1.41]	+0.73 [-0.73, 2.21]	+0.65 [0.16, 1.22]

Rows 1-3: NLI-pretrained | Rows 4-6: plain-base checkpoints

Exploratory ablation: PPS is diagnostic, not a repair

Held-out DF-Coref-Test intervention selection (Table 4)

PPS subset	NF rec. Single	NF rec. Strat.	PNF harm Single	PNF harm Strat.
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

NF recovery stayed 17 / 163; PNF harm fell from 82 / 704 to 51 / 704.
When avoiding PNF harm is the priority, the no-intervention baseline remains safest.

Interpret carefully

1 Narrow scope
Pronoun-resolved, surface-recoverable dialogue cases - not dialogue grounding in full.

2 Comparative diagnostic
PPS is not a coreference score or a full causal attribution of entity use.

3 Descriptive evidence
Repeated direction across models and sets; some confidence intervals cross zero.

Queen Mary University of London