BiCon-Gate: Consistency-Gated De-colloquialisation for Dialogue Fact-Checking

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

Abstract

Automated fact-checking in dialogue involves multi-turn conversations where colloquial language is frequent yet understudied. To address this gap, we propose a conservative rewrite candidate for each response claim via staged de-colloquialisation, combining lightweight surface normalisation with scoped in-claim coreference resolution. We then introduce BiCon-Gate, a semantics-aware consistency gate that selects the rewrite candidate only when it is semantically supported by the dialogue context, otherwise falling back to the original claim. This gated selection stabilises downstream fact-checking and yields gains in both evidence retrieval and fact verification. On the DialFact benchmark, our approach improves retrieval and verification—with particularly strong gains on SUPPORTS—and outperforms competitive baselines, including a decoder-based one-shot LLM rewrite that attempts to perform all de-colloquialisation steps in a single pass.

1 Introduction

Automated fact-checking determines whether a claim is supported, refuted, or there is not enough information (NEI) to verify it by grounding the claim in evidence from a reliable knowledge source such as Wikipedia (Thorne and Vlachos, 2018; Thorne et al., 2018). Most systems follow a retrieve–verify pipeline: evidence retrieval (Information Retrieval; IR) followed by fact verification (FV).

Dialogue fact-checking extends this setting to multi-turn conversations, where a response claim depends on preceding context C (Kim et al., 2021). In dialogue, colloquial phenomena—contractions, missing punctuation, inconsistent casing, ellipsis, and pronominal references—can blur claim boundaries and entity mentions (Figure 1), making both retrieval and verification brittle (Kim et al., 2021; Chamoun et al., 2023).

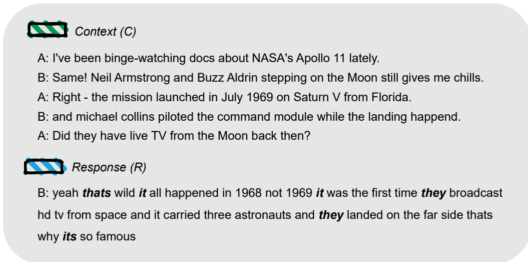

Figure 1: Example multi-turn dialogue illustrating a context-dependent response claim about the *Apollo 11 mission*. In-claim pronouns in the response refer to antecedents introduced in earlier turns.

A common remedy is to rewrite conversational inputs to be more self-contained (e.g., normalisation, de-contextualisation, or incomplete-utterance rewriting), which can improve recall on noisy queries (Sundriyal et al., 2023; Li et al., 2023; Deng et al., 2024; Guo et al., 2024; Cao et al., 2024).

This creates a rewrite-induced IR–FV trade-off: making a claim more explicit can increase retrieval recall, while even small semantic drift (e.g., resolving a pronoun to the wrong entity) can mislead verification and degrade end-to-end (E2E) performance (Gupta et al., 2022; Chamoun et al., 2023). We focus on in-claim pronominal anaphora because it is frequent in dialogue claims and can often be addressed via minimal span substitution rather than broad paraphrasing; in DialFact, 42.1% of claims contain at least one in-claim pronoun (§3.2).

We ask whether we can improve retrieval and verification with minimal, meaning-preserving edits to the claim surface. To this end, we first construct a conservative rewrite candidate via staged de-colloquialisation: de-contraction, punctuation restoration, true-casing, and scoped in-claim pronoun resolution. We then propose **BiCon-Gate**, a bidirectional natural language inference (NLI)-based consistency gate that routes each instance to the rewrite candidate only when it is semantically

Proceedings of the Ninth Fact Extraction and VERification Workshop (FEVER), pages 59–73
March 28, 2026

supported by the dialogue context; otherwise, it falls back to the original claim (Zhang et al., 2023).

Our contributions are three-fold:

- Controlling the IR–FV trade-off. We use a semantics-aware gate as a conservative control mechanism that selectively accepts rewrites to mitigate verification harm from semantic drift.
- Isolating where rewriting helps or hurts. We disentangle retrieval and verification effects via IR-only, FV-only (with gold evidence), and E2E evaluations.
- Demonstrating drift in one-shot LLM rewriting. We compare against a decoder-only, single-prompt rewrite baseline and show that aggressive one-shot rewrites can drift and hurt verification, whereas scoped edits paired with gating yield more reliable gains.

To our knowledge, this is the first work that uses bidirectional NLI-style consistency signals as an instance-wise router for conservative dialogue-claim rewriting.

To structure our evaluation, we test three hypotheses. These hypotheses are stated as mechanism-driven expectations based on how the claim surface is consumed by both stages of the pipeline: IR uses it as a query, while FV uses it as the verifier's hypothesis. In particular, R_1–R_3 make lightweight surface-form edits designed to preserve propositional content, whereas R_4 makes explicit reference substitutions that can change downstream behaviour, motivating H3's test of whether semantic routing improves E2E robustness.

We denote the cumulative pipeline outputs as R_1–R_4, and R_5 as the one-shot decoder rewrite (§4). (H1) Lightweight surface normalisation (R_1–R_3) is largely neutral for both retrieval and verification. (H2) Resolving in-claim pronouns (R_4) can improve fact verification, and BiCon-Gate increases robustness by falling back when a rewrite is not semantically supported. (H3) In end-to-end fact-checking, semantic routing mitigates the tension between retrieval gains and verification robustness.

2 Related Work

2.1 Colloquial noise in retrieval and verification

Informal dialogue phenomena (e.g., contractions, ellipsis, and pronominal references) can blur span boundaries and entity mentions, making retrieval and verification brittle. In dialogue fact-checking, making a claim more explicit can improve evidence retrieval, but incorrect rewrites (e.g., wrong antecedent substitutions) can introduce semantic drift that harms downstream verification. This yields an IR–FV trade-off: aggressive de-colloquialisation may raise recall, while conservative, semantics-preserving rewriting is needed to maintain verification accuracy.

On the retrieval side, Kim et al. (2021) show that converting formal FEVER (Thorne et al., 2018) claims into colloquial variants led to a large drop in document recall (90.00% $\rightarrow$ 72.20%), highlighting how sensitive rankers are to informal phrasing. In dialogue, Gupta et al. (2022) similarly identify colloquiality—especially pronouns and underspecified references—as an obstacle to fact-checking.

A complementary line of work makes inputs more self-contained before retrieval. Sundriyal et al. (2023) show that normalisation and de-contextualisation improve retrieval on noisy social media text, and de-contextualising pronoun-heavy claims improves evidence finding in multi-document settings (Deng et al., 2024). Related approaches rewrite incomplete utterances to mitigate ellipsis-driven failures in dialogue (Li et al., 2023; Guo et al., 2024; Cao et al., 2024). However, in fact-checking the rewritten claim also serves as the verifier's hypothesis; thus, retrieval-oriented rewrites can hurt E2E performance if they drift semantically. This motivates instance-wise safeguards that preserve meaning while still improving retrieval robustness.

Beyond rewriting, a related line of work improves fact-checking by explicitly aligning retrieval with verification objectives. Feedback-based Evidence Retriever (FER) feeds verifier signals back into the retriever, aligning document selection with end-task performance (Zhang et al., 2023). Our work follows a similar principle—using downstream-oriented signals—but applies it to deciding when to trust a rewrite, rather than which documents to retrieve.

2.2 Pronominal Coreference in Conversation

Recent mention-based coreference resolution systems such as Maverick (Martinelli et al., 2024) achieve strong accuracy with efficient span-scoring, but they are trained to link textual mentions and are not designed for discourse-level references or informal dialogue. We focus specifically on in-claim pronominal anaphora because the claim text is used

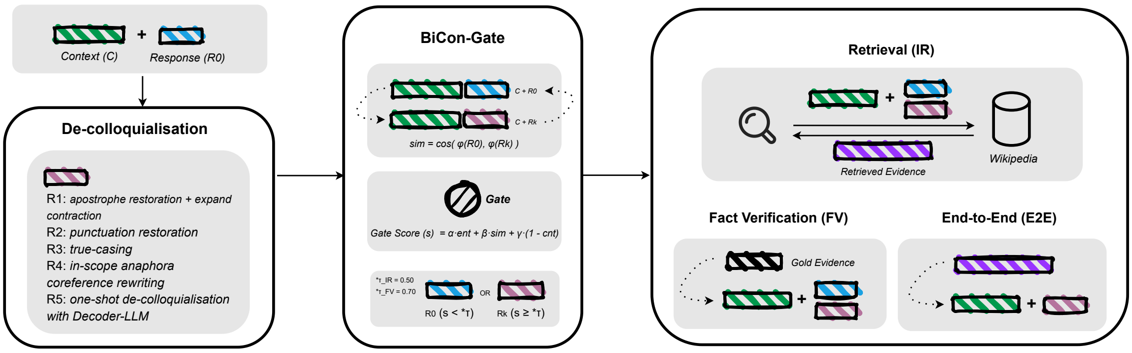

Figure 2: Overview of the staged de-colloquialisation pipeline (R_1-R_5), BiCon-Gate routing, and evaluation protocols (IR-only, FV-only, and E2E). BiCon-Gate combines bidirectional NLI entailment/contradiction signals (*ent*/*cnt*) with embedding cosine similarity (*sim*) and accepts a rewrite only when the gate score exceeds a task-specific threshold (τ); otherwise it falls back to R_0.

directly as (i) the retrieval query and (ii) the verifier's hypothesis; unresolved pronouns therefore degrade both stages. Moreover, in-claim pronouns can often be resolved from the preceding dialogue without rewriting the entire conversation, making them a targeted and controllable intervention.

LLMs demonstrate impressive referential reasoning, but converting that ability into stable gains on standard coreference benchmarks—especially for pronominal mentions in long contexts—remains challenging (Gan et al., 2024; Manikantan et al., 2025). Recent analyses suggest that remaining errors concentrate on pronominal mentions and overlapping mention structures in long contexts, making antecedent substitutions brittle even for strong LLMs.

In dialogue fact-checking, Chamoun et al. (2023) find that naively substituting in-claim pronouns improves document recall (56.85% $\rightarrow$ 67.00%) but slightly harms sentence-level evidence selection (54.19% $\rightarrow$ 53.86%) and fact-verification accuracy (44.06% $\rightarrow$ 42.71%). These findings suggest that imperfect coreference rewrites can propagate errors downstream. Their error analysis attributes this drop to incorrect coreference links that steer evidence selection toward similar sentences in the wrong document and can even change a claim's label with respect to the gold evidence, motivating our instance-wise semantic gate.

Prior work typically studies individual rewrite phenomena in isolation and reports only end-to-end outcomes, which conflates retrieval and verification effects under a fixed backbone. It also rarely provides an explicit, conservative acceptance criterion to prevent semantic drift when applying coreference-based rewrites. We address these gaps by (i) measuring the component-wise impact of cumulative rewrites on the same instances using IR-only, FV-only, and E2E protocols, and (ii) introducing an instance-wise semantic gate that adopts a rewrite only when it is consistent with the original claim.

2.3 NLI-based factual consistency and semantic gating

A recurring theme in factual-consistency research is to compare a candidate text against a reference via NLI-style containment and semantic similarity. SummaC operationalises consistency by checking whether each summary sentence is entailed by, and not contradicted by, the source document (Laban et al., 2022). AlignScore similarly defines alignment as being fully supported by the reference without contradiction (Zha et al., 2023). MENLI further shows that NLI-based entailment and contradiction signals are more robust indicators of correctness than sentence-similarity scores alone (Chen and Eger, 2023).

We adopt this principle in a different setting: rather than using NLI scores purely for evaluation, we use them operationally to decide whether a rewrite should be applied for a given instance. This design directly targets the IR–FV trade-off described above by accepting rewrites only when they are semantically consistent with the original claim. Motivated by NLI-based factual-consistency work (e.g., SummaC, AlignScore, MENLI), BiCon-Gate uses bidirectional entailment/contradiction signals together with semantic similarity to conservatively accept rewrites and otherwise fall back to the original claim.

3 Task & Data

3.1 Task Definition

Given a multi-turn dialogue context C and a response claim R, our system retrieves evidence from a Wikipedia snapshot aligned with DialFact (see § 5.2 for details) and predicts a label in {SUPPORTS, REFUTES, NEI}. Because the claim surface is used both as the retrieval query and as the verifier hypothesis, rewriting can have opposite effects on IR and FV (e.g., higher recall but increased semantic drift). We denote the original claim R as R_0 and the variants produced by our de-colloquialisation pipeline as R_1–R_5.

To isolate rewriting effects, we keep the retriever and verifier fixed and vary only the claim surface presented to them. We report results under three evaluation protocols: (i) IR-only, where the selected surface is used as the retrieval query; (ii) FV-only, where the verifier receives gold evidence and the selected surface is used as the hypothesis; and (iii) E2E, where both retrieval and verification consume the selected surface.

Figure 2 summarises our pipeline and where rewriting interacts with retrieval and verification. Starting from the original claim R_0, we generate progressively de-colloquialised variants R_1–R_4 through lightweight surface normalisation (de-contraction, punctuation restoration, true-casing) followed by scoped in-claim coreference rewriting. We also construct an alternative one-shot LLM rewrite R_5. BiCon-Gate then routes between R_0 and a candidate rewrite (mainly R_4; R_5 in an ablation) using semantic consistency signals derived from bidirectional NLI and embedding similarity; implementation details are provided in § 4.

3.2 Data

We use the official validation and test splits of DialFact (Gupta et al., 2022), whose dialogues are categorised into factual or personal subsets (Table A1). DialFact features colloquial multi-turn contexts, context-dependent claims, and a large proportion of NEI cases. We focus on DialFact because it combines multi-turn dialogue with a subset of sentence-level gold evidence, enabling retrieval-focused evaluation while still supporting E2E fact verification.

Sentence-level gold evidence annotations are provided only for the factual subset; accordingly, we compute IR metrics on that subset, while FV-only and E2E metrics are reported on the full test split. As shown in Table A1, 39.7% of validation claims and 44.3% of test claims contain at least one in-claim pronoun (42.1% overall). This prevalence motivates our focus on in-claim anaphora: unresolved pronominal references are common and directly affect both retrieval queries and verifier hypotheses.

4 Methodology

Our methodology consists of a staged de-colloquialisation pipeline with a semantic consistency gate. We describe the rewriting steps (R_1-R_5) and then the consistency gate and evaluation protocols.

4.1 De-colloquialisation

We adopt a cumulative pipeline that reduces colloquial noise and then resolves in-claim anaphora. Let R_0 be the original response; R_1–R_5 are derived as follows.

R_1 **De-contraction.** We first restore missing apostrophes with conservative, regex-gated rules (e.g., *im/ive/ill* $\rightarrow$ *I'm/I've/I'll*), then expand contractions (e.g., *it's* $\rightarrow$ *it is*) while protecting dotted acronyms (e.g., Ph.D., U.S.), honorifics, and URLs via placeholders. This stabilises token boundaries and negation cues.

R_2 **Turn-preserving punctuation restoration.** We apply a multilingual punctuation model (Vandeghinste and Guhr, 2024) to each turn and insert only predicted commas and sentence-final marks, leaving existing tokens and punctuation intact. For non-question turns without sentence-final punctuation, we append a period. This improves sentence and NP boundaries for downstream resolution.

R_3 **True-casing.** Using a BERT-based masked LM (`bert-base-cased` (Devlin et al., 2019)), we true-case sentence onsets and proper names with a margin rule: for each alphabetic token, we compare the MLM log-probabilities of upper- vs. lower-initial variants and flip only when the margin exceeds a threshold. We keep spelling, whitespace, and punctuation intact, modifying only token-initial characters to aid coreference resolution without semantic change of the underlying text.

R_4 **Scoped coreference rewriting (with gate).** Scope. We target only *in-scope* pronominal anaphora whose antecedents are present in C (e.g., *he/she/they/you*), excluding deictic (*this/that*) and expletive *it* (e.g., *it is raining*). Detect & propose. On R_3 we detect pronominal anchors (POS patterns) and let Maverick (Martinelli et al., 2024) propose up to 10 candidate antecedent NPs from the true-cased context. Select & rewrite. We rank up to 10 candidates with an instruction-tuned LLM (`Llama-3.1-8B-Instruct` (Grattafiori et al., 2024)) and substitute the selected antecedent span for the pronoun mention in R_3 to obtain a candidate R_4.

Instruction-following LLMs have been shown to work effectively as controllable decision modules for claim matching in automated fact-checking, motivating our use of an instruction-tuned LLM as a lightweight *selector* over a small candidate referent set (Pisarevskaya and Zubiaga, 2025). We then apply BiCon-Gate (§4.2) as an instance-wise router.

R_5 **Decoder-based one-shot reformulation.** A single `Qwen2.5-14B-Instruct` (Qwen et al., 2025) prompt attempts all editing steps (R_1-R_4) in one shot given $\{C+R_0\}$, producing R_5. We use `Qwen2.5-14B-Instruct` as a representative strong instruction-tuned decoder to instantiate a competitive one-shot baseline.

The rewrite is prompted as a constrained editing task (Appendix Table A5); the model is instructed to apply only surface normalisation and unambiguous pronoun substitution based on the provided context, without adding or changing meanings.

For ablation experiments (§5.5) we optionally apply the same semantic gate as a router between R_5 and R_0: for each instance, if $s_i^{(5)} \geq \tau$ we use R_5, otherwise we fall back to R_0. Model identifiers for all third-party components used in our pipeline are listed in the Appendix Table A4.

4.2 Consistency Gate: BiCon-Gate

For an instance i with context C_i and original response $R_{0,i}$, a rewrite $R_{k,i}$ ($k \in \{4,5\}$) is scored by

$$e_i = \min\Big\{p^{\text{ent}}\big(C_i+R_{0,i} \Rightarrow R_{k,i}\big), \\ p^{\text{ent}}\big(R_{k,i} \Rightarrow C_i+R_{0,i}\big)\Big\}, \quad (1)$$

$$c_i = \max\Big\{p^{\text{ctr}}\big(C_i+R_{0,i} \Rightarrow R_{k,i}\big), \\ p^{\text{ctr}}\big(R_{k,i} \Rightarrow C_i+R_{0,i}\big)\Big\}, \quad (2)$$

$$\text{sim}_i = \cos\big(\phi(R_{0,i}),\, \phi(R_{k,i})\big). \quad (3)$$

Here $p^{\text{ent}}(\cdot)$ and $p^{\text{ctr}}(\cdot)$ denote calibrated NLI probabilities for the entailment and contradiction classes, respectively, so that $e_i, c_i \in [0,1]$. ϕ is a sentence encoder and sim_i is the cosine similarity between the original and rewritten claim. We take the minimum bidirectional entailment to require mutual semantic containment, and the maximum bidirectional contradiction to penalise any directional inconsistency.

We use a three-way NLI model and score both directions by swapping the premise and hypothesis between $C_i+R_{0,i}$ and $R_{k,i}$; we take p^{ent} and p^{ctr} from the entailment and contradiction softmax probabilities.

We calibrate the NLI probabilities via temperature scaling. Concretely, we rescale NLI logits $\mathbf{z}$ as $\mathbf{z}/T$ before softmax and learn a single scalar T on the DialFact validation split (used as a calibration set) by minimising NLL. We then apply the learned $T = 4.96$ to all bidirectional NLI scores at test time (Xie et al., 2024; Guo et al., 2017).

The gate score and decision are:

$$s_i^{(k)} = \alpha\, e_i + \beta\, \text{sim}_i + \gamma\,(1-c_i), \quad (4)$$

$$\text{accept}_i^{(k)} = \mathbb{I}\big[s_i^{(k)} \geq \tau\big], \quad (5)$$

with non-negative weights $\alpha, \beta, \gamma \geq 0$ that sum to one ($\alpha+\beta+\gamma = 1$) and a threshold $\tau \in [0,1]$. For $R_{k,i}$, if $\text{accept}_i^{(k)} = 1$ we keep $R_{k,i}$; otherwise we fall back to $R_{0,i}$.

4.3 Protocols and metrics

We evaluate each claim surface R_k under three protocols that disentangle retrieval and verification. Unless stated otherwise, IR/E2E results focus on R_0–R_4, and R_5 is reported only as an FV-only ablation (§5.5); metrics are listed in Table A2.

IR-only. The query is $C+R_k$ and the retriever is fixed. We report document-level Recall@K and nDCG@K at the operating depths used in our retrieval stack (see §5.2).

For gate tuning (§5.1) we additionally report micro Recall@K and $1-\text{ZHR@K}$. ZHR@K denotes the zero-hit rate (the fraction of queries whose top-K retrieved set contains no gold evidence).

FV-only. We isolate the effect of the claim surface R_k on fact verification by providing the verifier with gold evidence sentences as premises and using $C+R_k$ as the hypothesis. We report fact-verification accuracy, macro-F1, and classwise F1 for SUPPORTS/REFUTES/NEI.

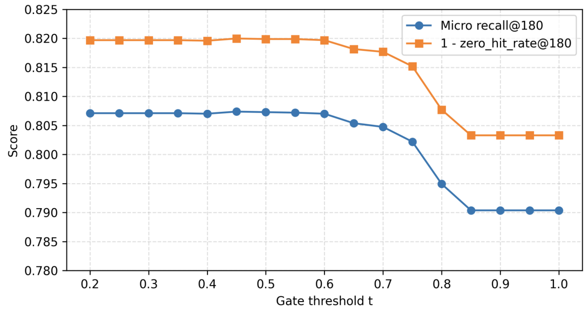

Figure 3: IR gate sweep on the validation split. We report BM25 micro-Recall@180 and $1-$ZHR@180 as a function of the R_4 gate threshold τ (where ZHR@180 is the zero-hit rate). We use $\tau_{\text{IR}} = 0.50$ in the IR-only protocol.

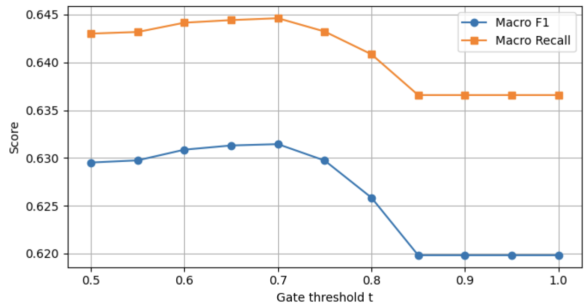

Figure 4: FV-only gate sweep on the validation split (gold evidence). Macro-F1 and macro-Recall as a function of the R_4 gate threshold τ; we use $\tau_{\text{FV}} = 0.70$ in FV-only and E2E.

	BM25 (K=180)		E5 dense (K=10)		BGE-CE (K=1)	
Claim	R@180	nDCG@180	R@10	nDCG@10	R@1	nDCG@1
R_0	73.89	48.37	70.47	61.98	53.51	57.01
R_1	73.22	47.96	69.95	61.56	53.14	56.59
R_2	73.22	47.96	69.94	61.60	53.18	56.64
R_3	73.22	47.96	69.96	61.67	53.21	56.66
R_4	**76.54**	**50.94**	73.18	64.45	55.53	59.15
+Gated	76.53	50.91	**73.19**	**64.46**	**55.58**	**59.19**

Table 1: IR-only retrieval results on the test (factual) split (macro Recall/nDCG, %) for claim variants R_0–R_4. R_4+*Gated* applies BiCon-Gate with $\tau_{\text{IR}} = 0.50$ and falls back to R_0 when the rewrite is rejected.

End-to-End (E2E). IR and FV both consume R_k. Unlike FV-only, the verifier receives the top-1 passage retrieved by the IR component as evidence, so E2E reflects the combined effect of rewriting under retrieval noise. We report the same FV metrics as FV-only: accuracy, macro-F1, and classwise F1.

5 Experiments

Our goal is to isolate the effect of claim decolloquialisation and semantic routing, rather than to optimise the underlying retriever or verifier. Therefore, across all settings we keep the IR and FV backbones fixed and vary only the claim surface (R_0–R_4) and whether BiCon-Gate routes to a rewrite candidate; we additionally include a decoder-based one-shot rewrite (R_5) as an ablation. We evaluate on DialFact because it provides multi-turn dialogue contexts and evidence annotations that support IR-only, FV-only, and E2E protocols.

5.1 Gate parameters

Following §2.3, we fix the gate weights to $(\alpha, \beta, \gamma) = (0.4, 0.2, 0.4)$ and tune task-specific thresholds on the validation split. We set α and γ symmetrically to weight entailment and non-contradiction equally, and down-weight cosine similarity (β) as a secondary signal because similarity alone can be high even under subtle semantic drift.

To minimise hyperparameter tuning while keeping the gate interpretable, we keep these weights fixed across all experiments and tune only τ. We leave weight learning or re-tuning under larger distribution shifts to future work.

Threshold for IR. For IR, we tune the threshold using BM25 retrieval metrics on the validation split. For each $\tau \in [0.20, 1.00]$ in increments of 0.05, we build a gated claim surface exactly as in §2.3, run BM25 with depth $K{=}180$, and compute micro-recall@180 and $1-$ZHR@180.

Both curves exhibit a broad plateau for $\tau \in [0.4, 0.6]$ and start to degrade once $\tau > 0.6$. We set the IR gate threshold to $\tau_{\text{IR}}{=}0.50$, which lies near the centre of this plateau and slightly maximises micro-recall@180 (Figure 3).

Threshold for FV. For FV, we use gold evidence and sweep the gate threshold $\tau \in [0.20, 1.00]$ on the validation split. We use the `gateSweep` routine, which for each threshold τ constructs hypotheses with either R_4 or R_0 according to whether the BiCon-Gate score satisfies $s_i^{(4)} \geq \tau$, and runs the verifier once on the resulting premise–hypothesis pairs. As shown in Figure 4, macro-F1 and macro-Recall both peak at $\tau \approx 0.70$ on valid split; performance improves as the gate starts accepting high-scoring R_4 rewrites, but once τ becomes too large we discard too many well-rewritten R_4 and the curves drop back towards the R_0 baseline.

5.2 IR-only

We study how the claim surface (R_0–R_4) affects evidence retrieval. We index a 37 GB Wikipedia snapshot (`2019-08-01`) into `100`-token passages with a stride of `50`, matching the dump used to construct DialFact (Gupta et al., 2022) and ensuring that all annotated gold evidence is, in principle, retrievable from our index.

Since all passages have roughly the same length, we use a relatively weak length normalisation (b=0.4) and set the term-frequency saturation parameter to k_1=1.5, following common BM25 settings in Pyserini/BEIR-style benchmarks (Lin et al., 2021). Our retrieval pipeline has three steps.

First, BM25 retrieves the top 300 passages per query. Based on an elbow analysis, we fix K=180 as the working cut-off and pass these 180 candidates to the dense retriever. Second, we apply an E5-large bi-encoder (Wang et al., 2024) to the K=180 BM25 candidates, retrieve the top 20 semantically similar passages and keep the top 10. Finally, we apply a BGE-reranker-large cross-encoder (Chen et al., 2024) to these 10 passages and select a single passage as the final evidence for IR-only evaluation.

The original claim R_0 already retrieves most gold evidence (73.89% R@180 and 48.37% nDCG@180 with BM25), and the intermediate normalisation variants R_1–R_3 leave IR almost unchanged: all metrics remain within 0.7 points of R_0 at every stage. Relative to R_0, R_4 improves BM25 recall@180 from 73.89% to 76.54% and nDCG@180 from 48.37% to 50.94%; E5 dense retrieval recall@10 rises from 70.47% to 73.18% and nDCG@10 from 61.98% to 64.45%; and BGE cross-encoder recall@1 increases from 53.51% to 55.53% with a similar gain in nDCG@1. These gains indicate that making colloquial claims more self-contained by resolving in-scope pronouns helps all three retrieval stages focus on the correct entities.

The gated variant uses the IR threshold τ_{IR}=0.50 chosen in §5.1. Figure 3 and Table 1 show that this threshold has almost no effect on retrieval: across BM25, E5, and BGE-CE, *Gated* stays within 0.05 points of R_4 on all recall and nDCG metrics.

Overall, these IR-only results are consistent with H1: minimal surface normalisation (R_1–R_3) is retrieval-neutral across all three retrieval stages. In contrast, scoped pronoun rewriting (R_4) improves retrieval gains, and applying the IR-tuned gate (τ_{IR}=0.50) preserves these gains.

Claim	Acc	Macro-F1	F1(S)	F1(R)	F1(NEI)	ΔF1
R_0	62.67	61.09	45.08	76.77	61.42	–
R_1	62.66	61.08	45.00	76.78	61.46	-0.01
R_2	62.71	61.16	45.33	76.69	61.47	+0.07
R_3	62.80	61.22	45.25	76.75	61.67	+0.13
R_4	63.15	61.85	47.59	76.20	61.77	+0.76
+*Gated*	**63.77**	**62.93**	**50.59**	76.12	**62.10**	**+1.84**
R_5	58.27	56.93	42.79	67.61	60.40	-4.16
+*Gated*	62.23	60.40	42.92	**77.00**	61.28	-0.69

Table 2: FV-only results on test split with gold evidence, comparing claim surfaces R_0-R_5 (Accuracy and macro/class-wise F1, %). R_4+*Gated* and R_5+*Gated* apply BiCon-Gate with τ_{FV} = 0.70 (fallback to R_0), and ΔF1 is the macro-F1 change over R_0.

5.3 FV-only

We run FV-only with gold evidence and an NLI verifier (He et al., 2023; Laurer et al., 2024); hypotheses use the last two dialogue turns by default, and R_4+*Gated* applies BiCon-Gate with the FV-tuned threshold τ_{FV}=0.70 (see §5.1). Table 2 summarises FV-only performance for R_0–R_5 and the gated variant of R_4. The original claim R_0 attains 61.09% macro-F1 and 62.67% accuracy.

Cumulative claim de-colloquialisation (R_1–R_3) leaves FV almost unchanged when premises are gold evidence: macro-F1 stays within ±0.2 points of R_0, and class-wise F1 for S/R/NEI shifts by at most 0.3 points.

In contrast, scoped coreference rewriting R_4 yields a noticeable FV gain. R_4 improves macro-F1 from 61.09% to 61.85% (+0.76) and accuracy from 62.67% to 63.15%. Most of this gain comes from the SUPPORTS class: F1(S) rises from 45.08% to 47.59%, while F1(NEI) also slightly improves (61.42% → 61.77%), and F1(R) decreases slightly (76.77% → 76.20%).

Making the claims more self-contained by replacing in-scope pronouns therefore helps the verifier distinguish supported facts without hurting NEI. This behaviour contrasts with observations by Chamoun et al. (2023), who report that generic coreference resolution and claim rewriting can improve document recall for conversational claims, but tend to degrade sentence-level evidence selection and claim verification on DialFact due to rewriting and resolution errors. In our setup, the scoped R_4 rewrites already improve FV over R_0.

The gated variant attains the best FV-only performance with 62.93% macro-F1 (+1.84 over R_0) and 63.77% accuracy, while F1(S) increases further to 50.59%. For comparison, the decoder one-shot rewrite R_5 is markedly worse than R_0 (56.93%

Claim	Acc	Macro-F1	F1(S)	F1(R)	F1(NEI)	ΔF1
R_0	34.85	22.60	2.45	15.99	49.37	0.00
R_1	34.86	22.62	2.50	16.01	49.37	+0.02
R_2	34.96	22.71	2.55	16.13	49.45	+0.11
R_3	35.02	22.88	2.89	**16.31**	49.44	+0.28
R_4	34.10	20.73	3.03	9.57	49.60	−1.87
+*Gated*	**35.54**	**23.22**	**4.56**	14.84	**50.24**	**+0.62**

Table 3: End-to-end results on the full test split using the top-1 retrieved passage as evidence (Accuracy and macro/class-wise F1, %). R_4+*Gated* uses BiCon-Gate with τ_{FV}=0.70.

macro-F1), and even with BiCon-Gate it recovers only to 60.40% macro-F1, motivating the ablation discussion in §5.5. These trends are robust to context window size: Appendix Table A6 and Figures A1–A2 summarise ΔMacro-F1 and class-wise ΔF1 over k.

Importantly, we keep the FV threshold fixed at $\tau_{\text{FV}} = 0.70$ when varying the context window size k. The consistent gains of R_4+*Gated* across k in Appendix Table A6 therefore suggest that this operating point is reasonably stable under this moderate distribution shift in dialogue context length.

Overall, these FV-only results are consistent with H1 and H2. With gold evidence provided as premises, light de-colloquialisation (R_1–R_3) remains largely verification-neutral, while scoped rewriting is the most reliable when filtered by BiCon-Gate.

5.4 End-to-End

For each claim surface R_k, we run the IR pipeline from §5.2 and take the BGE cross-encoder's top-1 passage as the premise. We then apply the NLI verifier from §5.3, constructing the hypothesis exactly as in FV-only (the last two turns in context followed by R_k). Since DialFact provides 1.32 gold evidence items per claim on average (test factual subset), top-1 is a conservative but practical choice that isolates claim-surface effects under retrieval noise; however, it may underestimate gains achievable with multi-evidence aggregation (Table 3).

Table 3 shows that using R_0 yields 22.60 macro-F1 and 34.85% accuracy, substantially lower than in the FV-only setting with gold evidence, reflecting the difficulty of relying on a single retrieved passage. Light de-colloquialisation (R_1–R_3) has only a small impact on the full pipeline. Macro-F1 remains within 0.3 points of the R_0 (22.60 → 22.88 for R_3), and accuracy increases only slightly from 34.85% to 35.02% for R_3.

Class-wise F1 follows a similar pattern across $R_{0\text{-}3}$: NEI stays high and stable (≈ 49.3-49.5), while SUPPORTS and REFUTES remain much lower (F1(S) ≤ 2.9, F1(R) ≈ 16). This suggests that, once retrieval noise is introduced, the main bottleneck is deciding SUPPORTS vs. REFUTES under noisy top-1 evidence, rather than the exact surface form of the claim.

Scoped coreference rewriting R_4 shows a different trade-off. Although R_4 improves document-level IR metrics (§5.2), it hurts E2E performance: macro-F1 drops from 22.60 to 20.73 and accuracy from 34.85 to 34.10. Most of this loss is due to REFUTES: F1(R) falls from 16.31 for R_3 to 9.57 for R_4, while F1(NEI) stays essentially unchanged (49.60%).

Applying BiCon-Gate on top of R_4 largely recovers these losses: the gated variant (R_4+*Gated*) uses the FV-tuned threshold from §5.1 to decide, for each example, whether to use R_4 (if $s_i^{(4)} \geq \tau_{\text{FV}}$) or fall back to R_0 otherwise. We retrieve evidence using the same claim surface selected by the gate. This simple policy yields the best E2E performance: macro-F1 rises to 23.22 (+0.62 over R_0) and accuracy to 35.54% (+0.69). Class-wise, the gate boosts F1(S) to 4.56 and raises F1(R) back to 14.84. Taken together, these E2E results are consistent with H3: by selectively accepting rewrites and otherwise falling back to R_0, BiCon-Gate mitigates the rewrite-induced IR–FV trade-off and yields the best E2E performance.

5.5 Analysis

This section provides three analyses of BiCon-Gate: (i) gate activation rates at the IR and FV/E2E operating points, clarifying how often the pipeline applies the scoped rewrite (R_4) rather than falling back to the original claim (R_0); (ii) an ablation comparing the decoder-based one-shot rewrite (R_5) to our scoped coreference rewrite (R_4) to isolate the impact of aggressive reformulation on verification; and (iii) qualitative error patterns that explain when rewriting helps and when it introduces semantic drift.

Gate activation rates. We report the routing (activation) rates at the IR and FV/E2E operating points. At τ_{IR}=0.50, the gate routes 52.59% of factual test queries to R_4.

At τ_{FV}=0.70, it routes 46.56% of test instances to R_4. In the R_5 ablation at the same FV threshold, the gate accepts only 0.92% of R_5 candidates (mean score 0.52), so R_5+*Gated* almost always

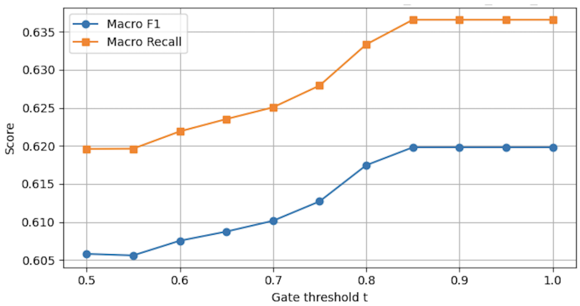

Figure 5: Effect of gating on the decoder one-shot rewrite (R_5) in FV-only on the validation split. As τ increases (rejecting more R_5 instances), macro-F1 and macro-Recall approach the R_0 baseline, indicating that the gate mitigates noisy rewrites primarily via fallback to R_0.

falls back to R_0. BiCon-Gate acts as a semantic router, sending high-confidence scoped rewrites to R_4 and routing drift-prone paraphrases back to R_0.

Decoder-based rewriting vs. scoped coreference We compare the decoder one-shot rewrite R_5 (Table A5) against the scoped pronoun rewrite R_4, which performs a targeted antecedent substitution only when a pronoun is in-scope (i.e., its antecedent appears in the dialogue context C). Across context lengths ($k \geq 2$), R_4 consistently improves FV performance (best with R_4+*Gated*), whereas R_5 degrades FV (Appendix Table A6).

Table 2 highlights the contrast clearly: while scoped rewriting R_4 improves FV (61.85% macro-F1) and its gated version further gains to 62.93%, the decoder rewrite R_5 collapses to 56.93% and its gated recovers only 60.40%. Consistent with the low acceptance rate for R_5 at τ_{FV}=0.70 (§5.5), this gap suggests that one-shot rewrites frequently introduce semantic drift, whereas targeted edits can be useful when selectively accepted.

Class-wise analysis supports this interpretation. Appendix Figure A2 shows that R_5's harm is driven primarily by a large drop in REFUTES across k, while R_4+*Gated* improves mostly via SUPPORTS gains. Further, Figure 5 shows that R_5 appears to improve mainly as the threshold increases and the system increasingly falls back toward R_0, whereas R_4 exhibits a clear interior optimum (Figure 4), indicating that a non-trivial subset of high-confidence scoped rewrites is genuinely beneficial.

Qualitative Error Analysis. Appendix Table A3 provides representative R_0/R_4/R_5 triples; we summarise the recurring patterns here. R_4 is typically a minimal, semantics-preserving edit: it replaces an in-scope pronoun with its referent from the preceding text C (e.g., *it* $\rightarrow$ *Heartbreak Hotel*, *He* $\rightarrow$ *Elvis*), making the hypothesis more self-contained without broader paraphrasing. This aligns with the consistent SUPPORTS gains observed for R_4+*Gated*.

In contrast, R_5 often performs surface-level normalisation (e.g., true-casing, punctuation, quoting) while leaving context-dependent pronouns unresolved; when it does paraphrase, it can blur cues that matter for contradiction, consistent with its REFUTES degradation (Appendix Figure A2). Gate scores reflect this difference: in these examples, R_4+*Gated* receives consistently higher scores than R_5+*Gated* (mean 0.79 vs. 0.54; Appendix Table A3), suggesting that BiCon-Gate favours verification-oriented de-contextualisation over cosmetic or potentially drifting rewrites.

6 Conclusions

We study how colloquial, context-dependent dialogue claims degrade both retrieval and verification, and propose a conservative de-colloquialisation pipeline—lightweight surface normalisation and scoped in-claim pronominal rewriting—paired with BiCon-Gate, a bidirectional NLI-based router that accepts rewrites only when semantically supported and otherwise falls back to the original claim.

On DialFact, surface-level normalisation is largely neutral, while scoped pronoun rewriting improves document retrieval across sparse, dense, and cross-encoder stages; yet in a strict top-1 E2E setting, ungated coreference rewrites can hurt verification despite better retrieval, highlighting a rewrite-induced IR–FV trade-off under retrieval noise.

BiCon-Gate mitigates this trade-off by selectively accepting high-confidence rewrites, yielding the strongest FV-only gains (notably on SUPPORTS) and the best top-1 E2E performance among the claim variants, whereas one-shot decoder rewrites are less reliable and often harm verification.

Overall, our results support H1–H3: minimal normalisation yields stable IR and FV with only marginal changes; scoped pronominal rewriting improves FV and BiCon-Gate further improves robustness by filtering rewrites that are not semantically supported; and semantic gating mitigates the IR–FV trade-off, making conservative rewriting a robust, controllable component in retrieval–verification pipelines.

Limitations

First, we evaluate only on the DialFact dataset using an English Wikipedia snapshot. The extent to which the observed gains transfer to other dialogue genres, longer contexts, or languages with different pronominal and morphological systems remains to be validated.

Second, BiCon-Gate depends on multiple off-the-shelf components (a coreference resolver, a retriever, NLI model, and sentence encoders). Their biases, errors, and calibration properties can affect gate decisions; moreover, gate thresholds tuned on DialFact may require retuning under distribution shifts, and the resulting multi-model pipeline increases complexity, latency, and inference cost.

Third, the scope of rewriting is intentionally narrow—limited to light normalisation and in-scope pronominal resolution—leaving other colloquial phenomena (e.g., deixis, ellipsis, or filler words) unaddressed. In addition, incorrect antecedent substitutions can introduce errors that propagate to downstream retrieval and verification.

Finally, our IR is passage-level and E2E setting uses a fixed retriever stack (BM25 $\rightarrow$ E5 $\rightarrow$ BGE-CE) and only the top-1 retrieved passage as evidence. Results may differ with multi-passage evidence aggregation, alternative retriever-verifier architectures, joint training, or recent evidence collections.

Ethics Statement

We use a publicly available dataset (DialFact) and an English Wikipedia snapshot, along with publicly released pretrained models, and we do not involve new data collection or human-subject studies. Our rewriting step—especially pronoun resolution—may introduce factual or attribution errors, which can lead to incorrect retrieval and verification outcomes. We therefore report aggregate benchmark results and caution against using outputs as definitive factual judgments without human oversight and transparent access to supporting evidence.

References

Zhiyu Cao, Peifeng Li, Yaxin Fan, and Qiaoming Zhu. 2024. Incomplete utterance rewriting with editing operation guidance and utterance augmentation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7225–7238, Miami, Florida, USA. Association for Computational Linguistics.

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.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Yanran Chen and Steffen Eger. 2023. Menli: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia. ELRA and ICCL.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330.

Xinnan Guo, Qian Zhu, Qiuhui Shi, Xuan Lin, Liubin Wang, DaqianLi DaqianLi, and Yongrui Chen. 2024. Context-aware tracking and dynamic introduction for incomplete utterance rewriting in extended multi-turn dialogues. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2138–2148, Bangkok, Thailand. 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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Byeongchang Kim, Hyunwoo Kim, Seokhee Hong, and Gunhee Kim. 2021. How robust are fact checking systems on colloquial claims? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1535–1548, Online. 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.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Building efficient universal classifiers with natural language inference. *Preprint*, arXiv:2312.17543.

Zitong Li, Jiawei Li, Haifeng Tang, Kenny Zhu, and Ruolan Yang. 2023. Incomplete utterance rewriting by a two-phase locate-and-fill regime. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2731–2745, Toronto, Canada. Association for Computational Linguistics.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: An easy-to-use python toolkit to support replicable ir research with sparse and dense representations. *Preprint*, arXiv:2102.10073.

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.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Dina Pisarevskaya and Arkaitz Zubiaga. 2025. Zero-shot and few-shot learning with instruction-following llms for claim matching in automated fact-checking. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9721–9736.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Megha Sundriyal, Tanmoy Chakraborty, and Preslav Nakov. 2023. From chaos to clarity: Claim normalization to empower fact-checking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6594–6609, Singapore. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Vincent Vandeghinste and Oliver Guhr. 2024. Fullstop: Punctuation and segmentation prediction for dutch with transformers. *Language Resources and Evaluation*, 58(4):1335–1354.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Johnathan Xie, Annie Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18128–18138.

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.

Hengran Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2023. From relevance to utility: Evidence retrieval with feedback for fact verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6373–6384, Singapore. Association for Computational Linguistics.

A Appendix

A.1 DialFact Dataset Statistics

Split	Type	#	S	R	NEI	Ev.Item	Turns(C)	hasPronouns
Valid	factual	8,691	3,342	3,363	1,986	1.13	4.58	**3,395**
	personal	1,745	0	0	1,745	1.01	4.36	747
	total	10,436	3,342	3,363	3,731	1.11	4.54	4,142
Test	factual	10,420	3,939	3,935	2,546	1.32	4.35	**4,549**
	personal	1,389	0	0	1,389	1.10	3.80	679
	total	11,809	3,939	3,935	3,935	1.29	4.28	5,228

Table A1: DialFact validation/test statistics and label counts. IR metrics are computed on the *factual* subset; FV/E2E use the full label distribution unless stated otherwise. "hasPronouns" counts samples containing *at least one in-claim pronoun*.

A.2 Evaluation Metrics

Protocol	Metrics
IR (gate tuning)	micro-Recall@K, $1 - \mathrm{ZHR}@K$ (BM25@180)
IR (doc-level)	macro-Recall@K, nDCG@K (BM25@180, E5@10, BGE-CE@1)
FV-only	Accuracy, macro-F1, classwise F1 (S/R/NEI)
End-to-End	Same FV metrics as FV-only, using IR-produced evidence

Table A2: Evaluation metrics: IR (gate tuning) metrics are computed on the valid split to select τ_{IR}, whereas IR (doc-level), FV-only, and E2E metrics are reported on the test split. ZHR is the zero-hit rate; the proportion of queries for which no gold passage is retrieved in the top-K list.

A.3 Pronoun Scope

We rewrite only *in-scope* anaphoric pronouns whose antecedents appear in dialogue context C. We exclude deictic uses without textual antecedent and expletive *it*.

- **Anaphora (coreferential, *in-scope*)** A pronoun refers back to a preceding entity in C.
  Ex. "I have heard that Louis C.K. performed there in the past." / "I did not know that. He was pretty funny." (*He* $\rightarrow$ Louis C.K.)
- **Deictic (*out-of-scope*)** Reference relies on extra-linguistic context, not on C.
  Ex. "This is amazing." (no textual antecedent)
- **Expletive (*out-of-scope*)** Non-referential "it" in weather/raising/extraposition.
  Ex. "It is raining." / "It seems that she left." / "It was John who called." / "It is important to exercise."

A.4 Examples of Rewrites

`Example 1 (ID: 877___8-1)`	
C	Elvis is great, he was born in 1935. … yeah he really brought rock and roll to the masses, thanks elvis. THANKS A LOT DUDE.
R_0	Remember when heartbreak hotel came out? The public hated **it** at first!
R_4	Remember when heartbreak hotel came out? The public hated **heartbreak hotel** at first! (gate(R_4) $= 0.8178$)
R_5	Remember when **"Heartbreak Hotel"** came out? The public hated **it** at first! (gate(R_5) $= 0.5445$)
`Example 2 (ID: 445___6-2)`	
C	I love The Walking Dead, I've seen every episode since it premiered on October 31, 2010. … Do you know what network I can find The Walking Dead on?
R_0	**It** premiered on amc in the us on october 31, 2010, but you can probably find **it** on any basic cable channel like fox or hulu.
R_4	**The show** premiered on amc in the us on October 31, 2010, but you can probably find **The show** on any basic cable channel like Fox or hulu. (gate(R_4) $= 0.7619$)
R_5	**It** premiered on **AMC** in the **US** on October 31, 2010, but you can probably find **it** on any basic cable channel like Fox or **Hulu**. (gate(R_5) $= 0.5368$)
`Example 3 (ID: 617___2-0)`	
C	were you aware that the famous musician Elvis' middle name was Aaron? no, i wasn't ! … i don't know much about Elvis. Where is he from?
R_0	**He** was born in Tupelo Mississippi, but relocated to Memphis when **he** was 13.
R_4	**Elvis** was born in Tupelo Mississippi, but relocated to Memphis when **Elvis** was 13. (gate(R_4) $= 0.7891$)
R_5	**He** was born in Tupelo, Mississippi, but relocated to Memphis when **he** was thirteen. (gate(R_5) $= 0.5490$)

Table A3: Representative DialFact instances used for qualitative analysis. Each block shows an excerpt of the dialogue context (C) and the resulting claim surface under the original claim (R_0), the scoped antecedent-substitution rewrite (R_4), and the decoder one-shot rewrite (R_5). Edited spans are marked with **...**; BiCon-Gate scores for R_4 and R_5 are shown in parentheses.

A.5 Model identifiers

Component	Model / identifier
Punctuation restoration (R_2)	`oliverguhr/fullstop-punctuation-multilang-large` (Vandeghinste and Guhr, 2024)
True-casing (R_3)	`bert-base-cased` (Devlin et al., 2019)
Coreference resolver (R_4)	Maverick coreference resolver (Martinelli et al., 2024)
Antecedent selector (R_4)	`meta-llama/Llama-3.1-8B-Instruct` (Grattafiori et al., 2024)
Decoder rewrite (R_5)	`Qwen/Qwen2.5-14B-Instruct` (Qwen et al., 2025)
NLI verifier (FV-only) / BiCon-Gate (NLI scorer)	`MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli` (Laurer et al., 2024)
BiCon-Gate embedding encoder	`intfloat/multilingual-e5-large` (Wang et al., 2024)
Sparse retriever (IR)	BM25 (Pyserini)
Dense retriever (IR)	E5-large (Wang et al., 2024)
Cross-encoder reranker (IR)	`BAAI/bge-reranker-large` (Chen et al., 2024)

Table A4: Model identifiers for third-party components used in our de-colloquialisation pipeline, retrieval, and verification experiments.

A.6 Prompts for Decoder-Based Rewriting

Prompt for decoder-based rewrite R_5

System
Follow the instructions exactly. Do not add or change facts.

User
You are an expert editor who rewrites informal, chatty utterances into well-formed declarative English without changing their meaning.
You will receive: (i) **Context**, a list of previous dialogue turns; and (ii) **Response**, the claim text to be normalised.
Task: Rewrite `Response` into `New_Response` by applying only the following operations.
(1) Add missing sentence-ending punctuation and fix spacing around punctuation.
(2) Fix capitalisation at sentence starts and for proper nouns.
(3) Insert missing apostrophes (e.g., *dont*→*don't*, *cant*→*can't*, *im*→*I'm*).
(4) Expand all contractions to full forms (e.g., *isn't*→*is not*, *aren't*→*are not*, *won't*→*will not*, *wouldn't*→*would not*, *I'm*→*I am*, *it's*→*it is*, *they're*→*they are*, *don't*→*do not*, *can't*→*cannot*).
(5) If a pronoun in `Response` (*this/that/it/he/she/they/these/those*) has a unique, clear antecedent in `Context`, replace it with that antecedent phrase; if ambiguous, leave it unchanged.
(6) Do not add, remove, or correct any facts, numbers, names, or dates; preserve the claim's semantics exactly.
(7) Output only the rewritten text as one or more sentences, with no explanations, lists, or markdown.

The input is formatted as: `Context (earliest→latest): {context_lines}`, `Response: {response_text}`, `Output: (New_Response only; no explanations)`.

Table A5: Prompt used with `Qwen2.5-14B-Instruct` to generate the decoder-based one-shot rewrite (R_5). The prompt constrains the model to apply only surface normalisation and unambiguous pronoun substitution based on the provided context, without changing claim's meaning.

A.7 Additional FV-only Results: Context Window Sensitivity

#Turns	Claim	FV-only (gold evidence)					
		Acc	Macro-F1	F1(S)	F1(R)	F1(NEI)	ΔF1
0	R_0	65.31	63.56	46.47	80.37	63.84	–
	R_1	65.20	63.43	46.17	80.27	63.84	-0.13
	R_2	65.27	63.53	46.50	80.26	63.84	-0.03
	R_3	65.36	63.58	46.34	80.35	64.04	+0.02
	R_4	65.02	63.41	47.38	79.38	63.48	-0.15
	+*Gated* (R_4)	65.90	**64.83**	51.15	79.06	64.28	**+1.27**
	R_5	60.84	59.50	44.91	70.82	62.76	-4.06
	+*Gated* (R_5)	65.31	63.56	46.47	80.38	63.84	+0.00
2	R_0	62.67	61.09	45.08	76.77	61.42	–
	R_1	62.66	61.08	45.00	76.78	61.46	-0.01
	R_2	62.71	61.16	45.33	76.69	61.47	+0.07
	R_3	62.80	61.22	45.25	76.75	61.67	+0.13
	R_4	63.15	61.85	47.59	76.20	61.77	+0.76
	+*Gated* (R_4)	63.77	**62.93**	50.59	76.12	62.10	**+1.84**
	R_5	58.27	56.93	42.79	67.61	60.40	-4.16
	+*Gated* (R_5)	62.23	60.40	42.92	77.00	61.28	-0.69
4	R_0	62.90	61.73	49.67	74.51	61.12	–
	R_1	62.94	61.76	49.51	74.64	61.13	+0.03
	R_2	62.89	61.72	49.58	74.55	61.03	-0.01
	R_3	62.86	61.66	49.46	74.38	60.66	-0.07
	R_4	63.37	62.40	51.63	74.13	61.45	+0.67
	+*Gated* (R_4)	63.91	**63.29**	54.07	73.98	61.82	**+1.56**
	R_5	58.54	57.61	46.81	65.84	60.19	-4.12
	+*Gated* (R_5)	62.83	61.58	48.66	74.86	61.22	-0.15
6	R_0	62.77	61.75	50.83	73.89	60.67	–
	R_1	62.82	61.80	50.79	73.93	60.67	+0.05
	R_2	62.81	61.78	50.85	73.85	60.64	+0.03
	R_3	62.73	61.72	50.91	73.69	60.57	-0.03
	R_4	63.33	62.52	53.04	73.57	60.95	+0.77
	+*Gated* (R_4)	63.94	**63.44**	55.44	73.47	61.42	**+1.69**
	R_5	58.21	57.46	47.73	65.07	59.58	-4.29
	+*Gated* (R_5)	62.88	61.77	49.90	74.32	61.07	+0.02
8	R_0	62.69	61.71	50.98	73.87	60.56	–
	R_1	62.78	61.81	50.98	73.82	60.62	+0.10
	R_2	62.72	61.73	50.95	73.67	60.59	+0.02
	R_3	62.63	61.66	51.00	73.53	60.45	-0.05
	R_4	63.21	62.43	53.16	73.32	60.82	+0.72
	+*Gated* (R_4)	63.64	**63.18**	55.40	73.23	60.91	**+1.47**
	R_5	58.13	57.41	47.95	64.86	59.42	-4.30
	+*Gated* (R_5)	62.61	61.57	50.08	73.80	60.83	-0.14

Table A6: FV-only context-window sensitivity (gold evidence). Results for $k \in \{0, 2, 4, 6, 8\}$ context turns, reporting Accuracy and macro/class-wise F1 (%) for each claim surface; ΔF1 denotes the macro-F1 change relative to R_0 at the same k.

A.8 FV-only Macro-F1 Δ Heatmap

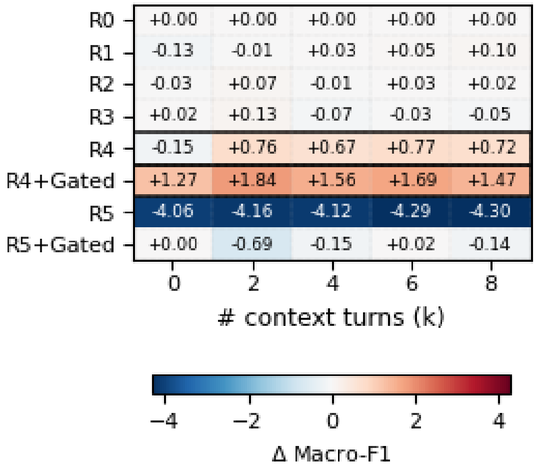

Figure A1: FV-only heatmap of Δmacro-F1 across context turns $k \in \{0, 2, 4, 6, 8\}$, computed relative to R_0 at the same k (Table A6). Boxed rows highlight R_4 and R_4+*Gated*, and the main setting is k=2.

A.9 FV-only Class-wise ΔF1 Heatmaps

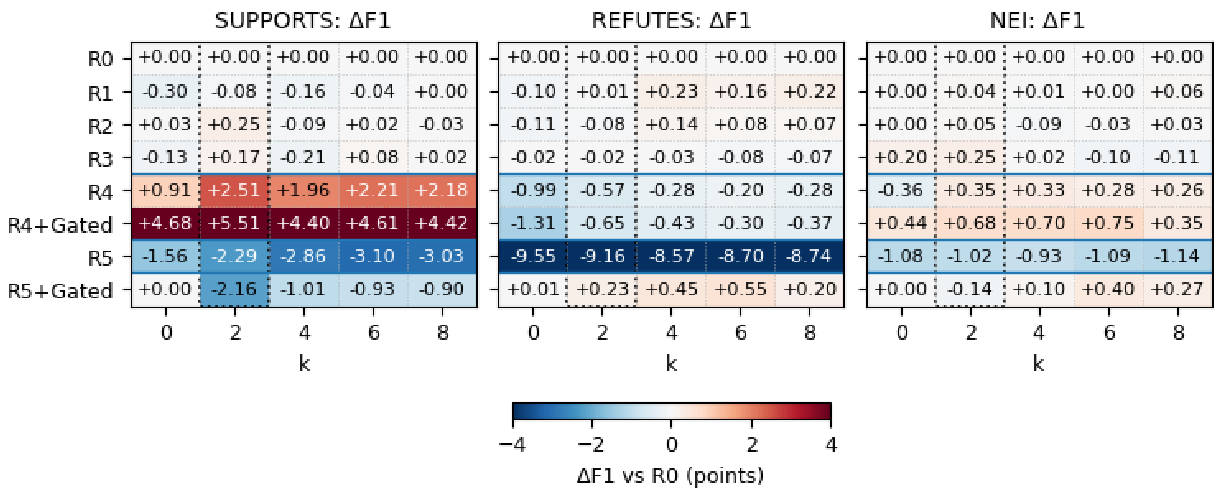

Figure A2: Class-wise FV-only ΔF1 over context turns $k \in \{0, 2, 4, 6, 8\}$, computed from Table A6. Colors are clipped for readability; annotated values show the true deltas, and the dotted box marks the main setting (k=2).