Claim-Level Selective Grounding for Reducing Partial Hallucinations in Small Language Models

Anonymous authors
Paper under double-blind review

Abstract

Language models often produce partial hallucinations: answers that are largely correct but contain a few unsupported or contradictory statements. These failures are operationally costly because they are difficult to spot; the response looks correct overall while still including specific wrong claims. A common mitigation is prompt-level selective refusal (e.g., instructing the model to answer only when supported and otherwise say "I don't know"). We show that this response-level intervention does not reliably reduce partial hallucinations under mixed-quality retrieval contexts and can worsen them by increasing the number of answers that contain both supported and conflicting claims.

We propose claim-level selective grounding, a post-hoc method that aligns the unit of control with the unit of failure. The method (i) generates an answer using a sub-3B instruction-tuned small language model (Gemma-3-1B-IT), (ii) extracts sentence-level claims, (iii) labels each claim as supported, conflicting, or irrelevant using natural language inference (NLI) against retrieved context, and (iv) recomposes the final answer using only supported claims; if none are supported, it outputs "I don't know". In a fully reproducible pipeline with an external Wikipedia-20231101 corpus and deliberately mixed-quality contexts, claim-level grounding reduces claim-level conflict from about 0.078 to about 0.0014 and lowers partial hallucinations from 12 to 14 down to 3 on a frozen sample of 100 items, while producing outputs for all ids (no silent dropping). These results suggest that hallucination mitigation in retrieval-conditioned QA is fundamentally a claim verification problem rather than a prompt-control problem.

1 Introduction

Retrieval-augmented generation (RAG) is widely used to improve factuality by conditioning generation on external evidence (1). Yet even with retrieval, models often emit partially hallucinated answers: most statements align with evidence, but a small number of claims are unsupported or directly contradict the provided context. These errors matter disproportionately: a single conflicted claim can invalidate an otherwise correct answer, and the high overall plausibility of the response makes human review harder.

A common mitigation is to prompt the model to refuse when evidence is insufficient (e.g., "if you are not sure, say you don't know"). Recent work studies training and prompting for refusal and abstention (6). However, refusal is typically a response-level control: it encourages either complete answering or complete abstention. Partial hallucinations, in contrast, are claim-level failures that occur within an otherwise correct response.

This paper makes three contributions:

- We formalize partial hallucinations as responses containing at least one conflicting claim and at least one supported claim, and we evaluate mitigation at the claim granularity.
- We show that prompt-level selective refusal does not reliably reduce partial hallucinations and can worsen them under mixed-quality retrieval contexts.

- We introduce claim-level selective grounding, a lightweight post-hoc procedure that removes conflicting and irrelevant claims, yielding near-elimination of conflicts without over-refusal.

2 Related Work

Retrieval-augmented generation. RAG combines parametric generation with non-parametric memory through retrieval (1; 2). While retrieval can increase support, it also introduces failure modes: irrelevant passages, decoys, and conflicting evidence (e.g., due to entity ambiguity, temporal drift, or retrieval noise). Such settings are common in practical systems and can trigger partial hallucinations.

Citation and long-form QA benchmarks. ASQA targets ambiguous factoid questions that require long-form synthesis (3). ALCE evaluates end-to-end systems that generate answers with citations and provides automatic evaluation protocols (4). In this work, ASQA/ALCE serve only as sources of QA items and claim targets; retrieval is performed over an external Wikipedia snapshot to control evidence quality.

Refusal and abstention. Prompting or tuning models to refuse when uncertain is a standard approach for safety and factuality (6). However, refusal is usually measured as an answer-vs-abstain decision, which does not directly address the mixed nature of partial hallucinations.

Claim verification with NLI. Natural language inference (NLI) models trained on datasets such as MNLI provide a practical signal for entailment/contradiction between a claim and evidence (5). We use NLI as a post-hoc verifier for extracted claims.

Decomposing generations for verification. FactScore evaluates long-form generations by decomposing them into atomic facts for verification (9). We use sentence-level segmentation for determinism and simplicity and treat granularity as a limitation.

3 Problem Setup

3.1 Partial hallucinations

Given a question q, retrieved context C (a set of passages), and a model-generated answer a consisting of sentences $\{s_i\}$, we treat each sentence s_i as a claim. A claim is:

- **supported** if it is entailed by C,
- **conflicting** if it contradicts C,
- **irrelevant** if C is insufficient to judge or does not address it (neutral).

A partial hallucination occurs when an answer contains at least one conflicting claim and at least one supported claim.

3.2 Why prompt refusal is misaligned

Prompt-level selective refusal encourages the model to abstain globally when uncertainty is detected. Under mixed-quality contexts, a model can still produce an answer with high apparent support while adding a small number of conflicting claims—precisely the partial hallucination regime. This motivates claim-level filtering rather than response-level abstention.

4 Methods

Our pipeline is implemented end-to-end in `scripts/10_run_phases.py` and is reproducible via CLI. We study five phases.

4.1 Phase 1: Answer generation

We generate an initial answer using Gemma-3-1B-IT (7) conditioned on a question q and retrieved context C.

4.2 Phase 2: Claim extraction

We split the answer into sentence-level claims $\{s_i\}$ using deterministic sentence segmentation. This choice prioritizes reproducibility and avoids introducing additional generation artifacts during extraction.

4.3 Phase 3: Claim scoring with NLI

For each claim s_i, we score it against the retrieved context C using an off-the-shelf NLI model trained on MNLI-style labels (5). We map NLI outputs to: support $\leftrightarrow$ entailment, conflict $\leftrightarrow$ contradiction, irrelevant $\leftrightarrow$ neutral. We record a per-claim label and confidence. (The exact verifier checkpoint is specified in the repository configuration for reproducibility.)

4.4 Phase 4: Baseline prompt-level selective refusal

We evaluate a prompt-based refusal baseline that instructs the model to answer only if supported by the context and otherwise respond "I don't know." This baseline is representative of response-level abstention without post-hoc verification.

4.5 Phase 5: Claim-level selective grounding (ours)

We construct the final answer by retaining only supported claims and preserving their original order. Conflicting and irrelevant claims are removed. If no supported claims remain, we output "I don't know." The system emits outputs for all item ids, enabling complete audit trails (no silent dropping).

5 Experimental Setup

5.1 Corpora and retrieval

We retrieve evidence from a Wikipedia snapshot dated 2023-11-01 (approximately 200k articles, 981k passages after chunking) (8). Retrieval contexts are deliberately mixed-quality: each item may include passages that are supportive, irrelevant, or explicitly conflicting (decoys). This setting isolates robustness to retrieval noise.

5.2 QA sources

We use ASQA (3) and ALCE (4) only as sources of questions and reference claim targets. They are not used as retrieval corpora.

5.3 Frozen evaluation

We report results on a frozen sample of 100 items (`sample100`) to ensure determinism across reruns of the reproducible pipeline.

5.4 Metrics

We report:

- **Support rate:** fraction of claims labeled supported,
- **Conflict rate:** fraction of claims labeled conflicting,

Table 1: Frozen `sample100` results under mixed-quality retrieval contexts.

Method	Support ↑	Conflict ↓	Partial Halluc. ↓
Phase 3: Baseline (score-only)	0.889	0.078	12
Phase 4: Prompt refusal	0.890	0.077	14
Phase 5: Claim-level grounding	0.998	0.0014	3

- **Partial hallucinations:** count of items with at least one supported and at least one conflicting claim.

6 Results

Table 1 summarizes performance across phases. Prompt-level refusal does not improve conflict rate and increases partial hallucinations in our setting. Claim-level selective grounding nearly eliminates conflicts and reduces partial hallucinations substantially.

7 Discussion

Why refusal can worsen partial hallucinations. With mixed-quality evidence, refusal prompts can shift generation dynamics toward producing "supported-looking" answers while still allowing a small number of conflicting claims to slip through. This increases the count of mixed (supported + conflict) items even if aggregate support/conflict rates remain similar.

Why claim-level grounding works. The method matches the unit of failure (claims) with the unit of control (claim filtering). By explicitly removing conflicts, it prevents small contradictory statements from contaminating otherwise correct responses. The "I don't know" fallback prevents unsupported synthesis while keeping abstention explicit.

Auditability. Because we output results for all item ids and preserve per-claim labels, the pipeline supports systematic debugging: users can inspect which claims were removed and which evidence triggered support vs. conflict.

8 Limitations

- Sentence-level claims are coarse. Some sentences bundle multiple atomic facts; a mixed sentence may be dropped entirely if labeled conflicting.
- Verifier errors. NLI models can be brittle under domain shift, negation, or long contexts. This work treats verifier quality as fixed and focuses on the post-hoc control mechanism.
- Evidence incompleteness. We score claims against the provided context, not against the full corpus; missing evidence can label a true claim as irrelevant.
- Task scope. We evaluate long-form QA under mixed-quality retrieval. Other tasks may require different claim extraction and verification.

9 Conclusion

Partial hallucinations are claim-level failures that are poorly addressed by response-level refusal. We show that prompt-level selective refusal does not reliably reduce partial hallucinations and can worsen them under

mixed-quality retrieval contexts. Claim-level selective grounding—NLI-based claim verification followed by recomposition from supported claims—nearly eliminates conflicts and reduces partial hallucinations while maintaining auditability and reproducibility on a small local model.

Disclaimer

This work and the accompanying code were created by the author in a personal capacity. They are not affiliated with, endorsed by, or representative of any current or past employer. No proprietary datasets, internal systems, or confidential resources were used; all experiments rely on publicly available datasets and public APIs.

Acknowledgments and AI assistance

The author thanks open-source contributors whose tools and datasets made this work possible. The author used AI-assisted tools (ChatGPT and OpenAI Codex) to help draft and edit parts of this manuscript and to scaffold portions of the accompanying codebase (e.g., boilerplate scripts and documentation). All experimental design choices, code changes, results, figures, and final wording were reviewed and edited by the author, who takes full responsibility for the content.

References

[1] Patrick Lewis et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS, 2020. arXiv:2005.11401.

[2] Vladimir Karpukhin et al. Dense Passage Retrieval for Open-Domain Question Answering. EMNLP, 2020. arXiv:2004.04906.

[3] Ivan Stelmakh et al. ASQA: Factoid Questions Meet Long-Form Answers. EMNLP, 2022. arXiv:2204.06092.

[4] Tianyu Gao et al. Enabling Large Language Models to Generate Text with Citations. EMNLP, 2023. arXiv:2305.14627.

[5] Adina Williams, Nikita Nangia, and Samuel Bowman. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. NAACL, 2018. arXiv:1704.05426.

[6] Lixin Cao et al. Learn to Refuse: Making Large Language Models More Controllable and Reliable through Refusal. EMNLP, 2024.

[7] Gemma Team. Gemma 3 Technical Report. 2025. arXiv:2503.19786.

[8] Wikimedia Foundation. Wikipedia database backup dumps. Snapshot: 2023-11-01. `https://dumps.wikimedia.org/`.

[9] Sewon Min et al. FactScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. EMNLP, 2023.