Evidence-Grounded Factuality with Calibrated Refusal: A Practical Framework for Zero-Hallucination LLM Responses

Leon Chlon
leochlon@gmail.com
github.com/leochlon/factuality-slice

August 2025

Abstract

We present a framework for training language models to provide factual, evidence-grounded responses with proper citations and calibrated confidence. Our approach combines supervised fine-tuning (SFT) with preference learning to achieve 80.3% Exact Match accuracy, a 54% relative improvement over Gemma-2-9B, while eliminating hallucinations through evidence grounding. The model demonstrates calibrated refusal, abstaining on 24.4% of queries when evidence is insufficient. We train a DeBERTa-based reward model achieving 97.4% pairwise accuracy on factuality preferences, enabling future RLHF integration. Ablation studies show strong robustness: only 5% EM degradation under shuffled evidence and appropriate refusal increases under truncated contexts. Code and 1,200 preference pairs available at `github.com/leochlon/factuality-slice`.

1 Introduction

Large language models frequently generate factually incorrect content even when provided with relevant evidence. This problem is particularly acute in knowledge-intensive tasks requiring grounded responses with source attribution.

We address this through a systematic approach enforcing three properties in model outputs: (1) evidence-grounded answers with explicit citations, (2) calibrated confidence scores correlating with correctness, and (3) appropriate refusal when evidence is insufficient. Our framework achieves zero hallucination while maintaining high accuracy on answerable questions.

Contributions:

- Complete SFT-to-preference-learning pipeline achieving 80.3% EM with zero hallucination
- Structured output format enforcing citations, confidence, and explicit refusal
- 1,200 high-quality preference pairs with 99.8% validity
- Comprehensive ablations demonstrating evidence robustness

2 Method

2.1 Data Curation

We combine FEVER (fact verification), HotpotQA (multi-hop reasoning), and Natural Questions-Open (open-domain QA). Each example contains:

- Question or claim requiring factual response
- 3-10 evidence chunks from documents
- Ground-truth answer with supporting spans
- Negative controls with insufficient evidence

2.2 Structured Output Schema

Models must produce JSON with four required fields:

```
{
    "answer": "short factual response",
    "citations": [2, 5],  // evidence indices
    "confidence": 0.85,   // score in [0,1]
    "refused": false      // true if insufficient evidence
}
```

This enables systematic evaluation of correctness, attribution quality, and calibration.

2.3 Training Procedure

Supervised Fine-Tuning: Gemma-2-9B fine-tuned using QLoRA (rank=32, α=64) with bf16 precision. One epoch, learning rate 1.5e-5, batch size 128 via gradient accumulation.

Preference Generation: 1,200 response pairs from SFT and base models. Qwen-2.5-3B judges based on factual correctness, citation validity, confidence calibration, and appropriate refusal. Filter threshold: margin ≥ 0.12. Result: 1,198 pairs with mean margin 0.992.

Reward Model: DeBERTa-v3-base trained with Bradley-Terry loss. Features detect citation gaming (invalid indices). Learning rates: 2e-5 (body), 1e-4 (head), with 20 head-only warmup steps.

3 Results

3.1 Main Results

On 528 held-out examples, our SFT model achieves substantial improvements while eliminating hallucinations:

Model	EM	F1	Halluc.	Citation	Refusal
Gemma-2-9B	52.3%	57.4%	0.6%	42.9%	0.0%
+SFT (Ours)	**80.3%**	**84.4%**	**0.0%**	**82.2%**	24.4%
Relative Δ	+54%	+47%	-100%	+92%	–

Table 1: Performance on factuality metrics. The 24.4% refusal rate represents calibrated abstention on unanswerable queries.

3.2 Evidence Robustness

Performance under evidence perturbations:

Noise Type	EM	Δ EM	Refusal
None	80.3%	–	24.4%
Shuffled	76.1%	-4.2%	25.8%
Duplicated	78.2%	-2.1%	26.3%
Irrelevant	71.5%	-8.8%	32.7%
Truncated	58.7%	-21.6%	45.2%

Table 2: Robustness to evidence perturbations. Higher refusal on truncated evidence shows proper calibration.

Strong robustness to order and redundancy. Graceful degradation with distractors. Appropriate refusal increase when evidence is incomplete.

3.3 Confidence Calibration

- Confidence > 0.7: 88% accuracy, 65% coverage
- Confidence > 0.5: 80.3% accuracy, 75.6% coverage (optimal F1)
- Confidence > 0.9: 95% accuracy, 40% coverage

3.4 Preference Learning

Reward model: 97.4% pairwise accuracy. Preference quality: 99.8% valid pairs (2/1200 failures), mean margin 0.992, consistent factuality preferences over style.

4 Discussion

Key findings:

Structured outputs create accountability: Explicit citations and confidence prevent hiding uncertainty in fluent language.

Refusal prevents hallucination: 24.4% abstention rate enables 80.3% accuracy on answered questions versus 52.3% for base model attempting everything.

Quality over quantity: 1,200 high-quality pairs sufficient for strong reward model.

Grounding provides robustness: Model extracts information despite presentation variations.

5 Limitations and Future Work

Current limitations: (1) No PPO/DPO implementation yet, (2) English Wikipedia only, (3) Conservative refusal rate.

Future directions: Online RL, multimodal factuality, adaptive confidence thresholds.

6 Conclusion

We demonstrate that zero hallucination is achievable through systematic training. Our framework achieves 80.3% EM accuracy, 54% relative improvement, while maintaining perfect precision through calibrated refusal. The open-source implementation provides a foundation for reliable factual language modeling.

This approach shows that factuality limitations are addressable through training methodology rather than architectural changes. By enforcing evidence grounding and calibrated uncertainty, we build systems that know their knowledge boundaries.