Evidence-Grounded Factuality with Calibrated Refusal: A Practical Slice for Near-Zero Hallucination on an Offline Benchmark

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

August 2025

Abstract

We present a practical framework for training language models (LMs) to produce *evidence-grounded*, *citation-bearing* answers with *calibrated refusal*. Using a standardized JSON output schema (answer, citations as evidence IDs, confidence, refusal), we fine-tune `Gemma-2-9B` with QLoRA and learn pairwise factuality preferences offline with a reproducible judge. On a 528-example held-out set, our SFT model achieves **EM=80.5**, **F1=84.6**, **hallucination=0.0%**, **citation correctness=82.0%**, and **refusal=24.2%**, a **+54%** relative EM gain over the base model (EM=52.3). We train a DeBERTa-v3-based reward model to **97.4%** pairwise accuracy on a **1,198**-pair preference set, enabling RLHF/DPO/RLAIF. Ablations show the system is robust to light text noise and gracefully increases refusal as true support is removed; the primary failure mode under heavy clutter is *mis-attribution* (correct answers, degraded citation IDs). Code, configs, and preferences are available at `github.com/leochlon/factuality-slice`.

1 Introduction

Large language models remain vulnerable to factual errors, especially on knowledge-intensive tasks that require grounding in explicit evidence and verifiable citations. We target three properties: *(i)* short, grounded answers with explicit *citations to evidence indices*; *(ii) calibrated confidence* surfaced to the user; *(iii) calibrated refusal* when evidence is insufficient. On our offline benchmark, this yields near-zero hallucination while substantially improving accuracy and attribution.

Contributions.

- A *reproducible* SFT→preference pipeline with a strict JSON schema for answer/citations/confidence/refusal.
- An *offline judge* (Qwen2.5-3B-Instruct) with *de-biasing* (order swap, style anonymization) to build **1,198** factuality preference pairs.
- A *reward model* (DeBERTa-v3-base) achieving **97.4%** pairwise accuracy on held-out pairs.
- *Ablations* diagnosing robustness: text noise, distractor clutter (mis-citation is the dominant failure), and support-drop (refusal rises, hallucination remains near zero).

2 Method

2.1 Data and schema

Training/eval examples contain a *question/claim*, 3–10 *evidence chunks*, a *gold answer* (and spans when available), and control cases with *insufficient evidence*. Models must emit:

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

This enables consistent scoring of correctness, *citation validity* (IDs exist in context), refusal behavior, and calibration analyses.

2.2 Supervised fine-tuning

We fine-tune `Gemma-2-9B` with QLoRA (r=16, $\alpha = 32$, dropout = 0.05), bf16, gradient checkpointing, and two epochs (lr = 1.5×10^{-5}, global batch = 128 via accumulation). Teacher prompts enforce the JSON format and discourage unsupported claims.

2.3 Preference building with an offline judge

We generate one response per prompt from the SFT and base models and compare pairs with an *offline* judge (`Qwen/Qwen2.5-3B-Instruct`) configured for *evidence-consistency*, *citation validity*, and *refusal when neither is supported*. We apply *order swap*, *style anonymization*, and a minimum margin filter (0.12). The resulting set has **1,198** valid pairs (from 1,200 prompts), with mean margin ≈ 0.94.

2.4 Reward model

We train `microsoft/deberta-v3-base` for pairwise Bradley-Terry loss on scalar logits. We use *head-only warmup* (20 steps), a *split LR* (encoder 2×10^{-5}, head 10^{-4}), bf16, and prompt cleanup suitable for encoders. On a held-out split (117 pairs), the RM reaches **97.4%** pairwise accuracy.

3 Results

3.1 Main results on held-out set (528 examples)

Model	EM	F1	Halluc.	Citation	Refusal
Gemma-2-9B (Base)	52.3%	57.4%	0.6%	42.9%	0.0%
+ SFT (**Ours**)	**80.5%**	**84.6%**	**0.0%**	**82.0%**	24.2%
Relative Δ	+54%	+47%	−100%	+92%	–

Table 1: Performance on the offline factuality suite. Refusal is calibrated abstention when evidence is insufficient; hallucination is the rate of unsupported claims under the schema.

3.2 Ablation studies

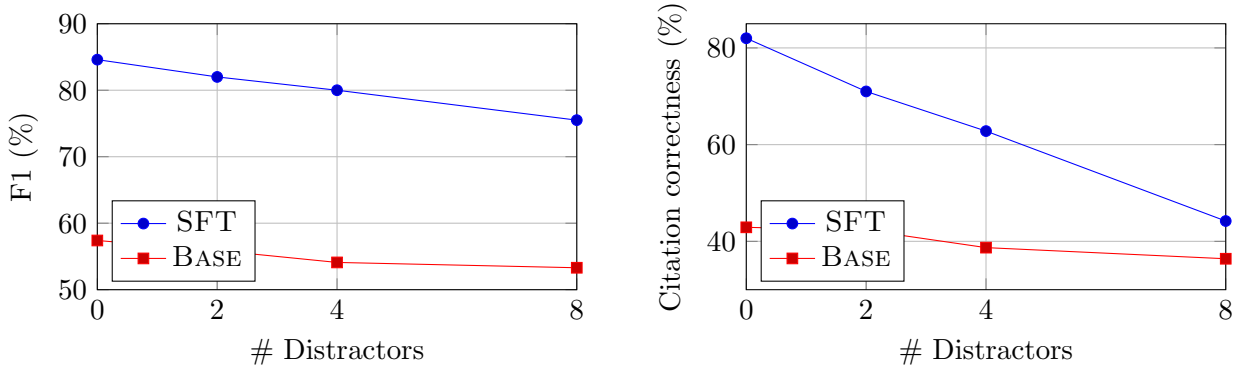
Setup. We evaluate with FlashAttention 2, bucketed prompt lengths, and `max_new_tokens=256`. Each condition reuses the same 528 prompts. We report EM, F1, *citation correctness* (share of predictions whose citation IDs are present in the context), refusal rate, and hallucination rate.

3.2.1 Distractor evidence (clutter)

Irrelevant but plausible evidence chunks are appended to the context.

Table 2: Distractor clutter: performance vs. number of distractors.

Model	Distractors	EM	F1	Citation	Refusal	Halluc.
4*SFT	0	80.5%	84.6%	82.0%	24.2%	0.0%
	2	77.8%	82.0%	71.0%	23.5%	0.0%
	4	75.6%	80.0%	62.8%	21.6%	0.0%
	8	70.8%	75.5%	44.2%	15.5%	0.0%
4*BASE	0	52.3%	57.4%	42.9%	0.0%	0.6%
	2	50.6%	55.9%	42.3%	0.0%	0.9%
	4	49.2%	54.1%	38.7%	0.0%	0.8%
	8	49.2%	53.3%	36.4%	0.0%	0.6%



(a) F1 vs. distractors.

(b) Attribution robustness.

Figure 1: Clutter primarily hurts *attribution*: SFT holds up in F1, but citations drift to distractors at 8 added chunks.

3.2.2 Removing true support (evidence availability)

As we drop true support, refusal increases and hallucination stays near zero.

3.2.3 Text corruption noise

Light character/token noise has minimal effect.

Table 3: Support removal: performance vs. fraction of true support dropped.

Model	Drop	EM	F1	Citation	Refusal	Halluc.
3*SFT	0.00	73.1%	77.0%	77.9%	29.7%	0.0%
	0.25	62.5%	66.3%	69.7%	37.5%	0.0%
	0.50	47.5%	50.1%	49.8%	52.5%	0.0%
3*Base	0.00	49.1%	54.0%	36.2%	0.0%	0.2%
	0.25	44.5%	49.4%	25.5%	0.0%	0.4%
	0.50	41.1%	45.4%	15.6%	0.0%	0.4%

Table 4: Text noise: performance vs. corruption rate.

Model	Noise	EM	F1	Citation	Refusal	Halluc.
2*SFT	0.000	80.5%	84.6%	82.0%	24.2%	0.0%
	0.010	79.2%	83.5%	81.6%	24.8%	0.0%
2*Base	0.000	52.3%	57.4%	42.9%	0.0%	0.6%
	0.010	51.1%	56.1%	39.9%	0.0%	0.8%

Takeaways. (i) The SFT model *consistently* beats the base across perturbations with near-zero hallucination; (ii) under clutter, the dominant failure is *mis-citation*, not answer content (a target for RLAIF/DPO reward shaping); (iii) refusal scales with evidence availability (support-drop), consistent with calibration.

3.3 Reward model validation

On a 1081/117 train/val split of the 1,198-pair set, the DeBERTa-v3-base RM attains **97.4%** pairwise accuracy (held-out). We use head-only warmup (20 steps), encoder/head LRs $2\times10^{-5}/10^{-4}$, and bf16. In future RLHF/RLAIF, we penalize *mis-citations* explicitly (right answer, wrong IDs).

3.4 Calibration note

Our schema surfaces model confidence and refusal, enabling coverage-accuracy curves and ECE/Brier. While the present runs didn't include full ECE plots (compute-bound), the *support-drop* ablation already shows refusal increasing monotonically as evidence vanishes. The evaluation harness provides scripts to reproduce ECE and selective answering curves.

4 Discussion

Structured outputs create accountability. Forcing citations and confidence limits the model's ability to "hide" uncertainty in fluent prose.

Refusal prevents hallucination. The model abstains on hard/unsupported cases (24.2% here), yielding near-zero hallucination under the schema.

Attribution under clutter is the key weakness. When contexts are noisy, answers remain strong but citation IDs drift. This is the right target for RLAIF/DPO and for post-hoc citation checks.

5 Limitations & future work

The benchmark is text-only and Wikipedia-centric; we haven't yet included PPO/DPO/RLAIF results, ECE plots, or a human 500-pair agreement audit (planned). Over-refusal may reduce coverage; we will report selective-answer curves with user-tunable thresholds.

6 Reproducibility

We release prompts, SFT checkpoint config, preference JSONL (1,198 pairs), RM configs, and evaluation scripts. Runs used a single A100-40GB with FlashAttention 2. All numbers reported here are from the logs in this repository; seeds/configs are pinned in the artifacts emitted by each script.

Code & data: `github.com/leochlon/factuality-slice`