

Reducing Hallucinations in Zen Models: Retrieval Augmentation, Confidence Calibration, and Citation-Grounded Generation

Technical Report v2025.08

Antje Worring, Zach Kelling
Zen LM Research Team
research@zenlm.org

August 2025

Abstract

Hallucination — the generation of plausible but factually incorrect content — remains a central challenge for large language models. We report a comprehensive study of hallucination in the Qwen3-based Zen models (dense 0.6B/4B/8B/32B and the Qwen3-30B-A3B mixture-of-experts variant, all Apache-2.0) and introduce three complementary interventions: (1) **Retrieval-Augmented Factuality Training (RAFT)**, which fine-tunes models to condition on retrieved evidence when generating factual claims; (2) **Uncertainty-Aware Decoding (UAD)**, which modulates token probabilities by a calibrated confidence estimate to suppress low-confidence continuations; and (3) **Citation-Grounded Generation (CGG)**, which trains models to emit inline citations linking claims to source documents. Together, these methods substantially reduce hallucination rates on TruthfulQA, FactScore, and HaluEval relative to the unmodified base models, with negligible perplexity degradation. We report the *relative* contribution of each component rather than restating third-party benchmark figures for the base models.

Contents

1	Introduction	3
1.1	Contributions	3
2	Hallucination Measurement	3
2.1	Benchmark Descriptions	3
2.2	Baseline Hallucination Rates	4
2.3	Domain Stratification	4
3	Retrieval-Augmented Factuality Training (RAFT)	4
3.1	Motivation	4
3.2	Training Procedure	4
3.3	RAFT Loss	4
4	Uncertainty-Aware Decoding (UAD)	5
4.1	Confidence Calibration	5
4.2	Temperature Scaling with Semantic Adjustment	5
4.3	Abstention Mechanism	5

5	Citation-Grounded Generation (CGG)	5
5.1	Citation Format	5
5.2	Citation Training	5
5.3	Citation Confidence Score	6
6	Experiments	6
6.1	Main Results (Component Ablation)	6
6.2	Scale Dependence	6
6.3	Domain-Stratified Results	6
6.4	Citation Accuracy	7
6.5	Abstention Rate	7
6.6	Perplexity Impact	7
7	Related Work	7
8	Conclusion	7

1 Introduction

Language models learn to predict the next token by compressing statistical patterns in training text. When queried about facts outside their training distribution, or about subtle factual nuances within it, they may generate fluent but incorrect content with high confidence. This phenomenon — hallucination — erodes user trust and limits deployment in high-stakes domains such as medicine, law, and scientific research.

Hallucination manifests in several forms:

- **Factual hallucination:** asserting a false proposition about the world (e.g., incorrect dates, attributed quotes, fabricated citations).
- **Knowledge boundary hallucination:** confidently answering questions that are genuinely unanswerable from training data.
- **Reasoning hallucination:** correct premises leading to incorrect conclusions through flawed inference steps.

We address all three categories through our RAFT + UAD + CGG system, evaluated on the Qwen3-based Zen models across the 4B, 8B, and 32B dense scales (and the Qwen3-30B-A3B MoE variant).

1.1 Contributions

1. A systematic measurement of hallucination rates across domains and model scales, using TruthfulQA, FactScore, HaluEval, and a new domain-stratified hallucination benchmark (ZenHallu).
2. Retrieval-Augmented Factuality Training (RAFT): a fine-tuning recipe that teaches models to identify when retrieval is needed and how to synthesize evidence faithfully.
3. Uncertainty-Aware Decoding (UAD): a training-free decoding modification that calibrates output confidence and suppresses hedging-unfaithful generations.
4. Citation-Grounded Generation (CGG): a training procedure for models to emit structured inline citations with confidence scores.
5. Ablation and analysis showing complementary contributions of each component.

2 Hallucination Measurement

2.1 Benchmark Descriptions

TruthfulQA [1]: 817 questions spanning conspiracy theories, misconceptions, fiction, and factual recall. Models must answer truthfully or abstain. We use the multiple-choice formulation.

FactScore [2]: Measures the factual precision of long-form biography generation by decomposing outputs into atomic claims and verifying each against Wikipedia. Score $\in [0, 1]$, higher is better.

HaluEval [3]: A large-scale hallucination evaluation benchmark with 35,000 samples across QA, dialogue, and summarization. Each sample contains a human-annotated hallucinated response and a correct response; models are evaluated on classification accuracy.

ZenHallu (ours): A domain-stratified benchmark covering medicine, law, science, history, mathematics, and geography. Each domain contains 500 factual questions with verified ground truth. Hallucination rate is measured as the fraction of generated responses containing at least one verifiable factual error.

2.2 Baseline Hallucination Rates

Measuring the unmodified Qwen3-based Zen models (4B, 8B, and 32B dense) on the four benchmarks above, we observe the expected scale trend: larger models hallucinate less, scoring higher on TruthfulQA and FactScore and lower on the ZenHallu error rate, with the 32B dense model the strongest of the dense variants. These baseline measurements establish the reference point against which the RAFT + UAD + CGG interventions are evaluated in Section 6; we report intervention effects as improvements relative to this baseline rather than as standalone capability claims.

2.3 Domain Stratification

Stratifying ZenHallu by domain (each domain contains 500 verified factual questions), hallucination rates on the 32B baseline are highest in **medicine** and **law**, intermediate in **science** and **history**, and lowest in **geography** and **mathematics**. This ordering reflects the relative density of verifiable, frequently-updated factual claims in each domain: specialized, fast-changing knowledge (clinical facts, statutes) is harder to recall reliably than stable, heavily-represented knowledge (geographic facts, arithmetic).

3 Retrieval-Augmented Factuality Training (RAFT)

3.1 Motivation

Standard retrieval-augmented generation (RAG) appends retrieved context to the input prompt at inference time. This helps if the model knows to trust the retrieved context, but baseline models often ignore retrieved evidence or confabulate details not present in it. RAFT trains the model to faithfully condition on retrieved evidence.

3.2 Training Procedure

RAFT fine-tunes on a dataset of (question, retrieved documents, answer) triples, where the answer is grounded in the retrieved documents. We construct the RAFT training set using:

1. **Retrieval**: For each factual question q , retrieve the top- k documents $\{d_1, \dots, d_k\}$ from a 100B-token knowledge base using a dense retriever.
2. **Distractor injection**: With probability $p_{\text{dist}} = 0.4$, replace one retrieved document with a semantically similar but factually incorrect distractor. This trains the model to critically evaluate rather than blindly copy.
3. **Chain-of-thought grounding**: The target answer includes an explicit chain of evidence linking claims to specific document spans, formatted as “[Source: d_j , span: ...] $\Rightarrow$ claim”.

3.3 RAFT Loss

The RAFT fine-tuning loss combines standard language modeling with a faithfulness penalty:

$$\mathcal{L}_{\text{RAFT}} = \mathcal{L}_{\text{CE}} + \beta \cdot \mathcal{L}_{\text{faith}} \quad (1)$$

The faithfulness loss $\mathcal{L}_{\text{faith}}$ penalizes generated claims that cannot be grounded in the retrieved documents. It is computed using an NLI classifier h that scores the entailment between each generated claim c_t and the document set:

$$\mathcal{L}_{\text{faith}} = \frac{1}{T} \sum_{t=1}^T \max(0, \gamma - \max_j h(d_j, c_t)) \quad (2)$$

where γ is a faithfulness margin and T is the number of generated claims.

4 Uncertainty-Aware Decoding (UAD)

4.1 Confidence Calibration

A well-calibrated model should express uncertainty proportional to the likelihood that its output is correct. We measure calibration via the Expected Calibration Error (ECE):

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)| \quad (3)$$

where B_m partitions predictions by confidence into M bins.

The baseline Zen-32B model shows a substantial Expected Calibration Error, indicating significant overconfidence that UAD is designed to correct.

4.2 Temperature Scaling with Semantic Adjustment

UAD applies a two-stage calibration: first, global temperature scaling T^* learned on a held-out calibration set to minimize ECE; second, a semantic confidence adjustment that modulates the effective temperature based on the model’s uncertainty about the semantic content of the response:

$$p_{\text{UAD}}(y_t \mid y_{<t}, x) \propto \exp\left(\frac{\log p(y_t \mid y_{<t}, x)}{T^* \cdot (1 + \kappa \cdot u_t)}\right) \quad (4)$$

where $u_t \in [0, 1]$ is a per-token uncertainty estimate derived from the model’s entropy over the top- V vocabulary candidates, and κ controls the strength of the adjustment. High u_t increases the effective temperature, producing a flatter distribution that prevents overconfident hallucinated tokens.

4.3 Abstention Mechanism

When the aggregate uncertainty $\bar{u} = \frac{1}{T} \sum_t u_t$ exceeds a threshold θ_{abs} , the model emits an abstention token “I am not confident enough to answer this question” rather than a potentially hallucinated response. This is calibrated to achieve a false abstention rate below 5% on answerable questions.

5 Citation-Grounded Generation (CGG)

5.1 Citation Format

CGG trains models to interleave inline citations within generated text:

The boiling point of water at sea level is 100°C [cite:doc_42, p.7]
with a boiling point elevation of 0.512°C·kg/mol [cite:doc_18, p.14].

Each citation specifies a document ID and page/paragraph range from the retrieved context.

5.2 Citation Training

CGG adds a citation prediction head to the model that, at each position where a factual claim is completed, predicts the most relevant source span. The head is a lightweight cross-attention module over the retrieved document embeddings:

$$\alpha_j = \text{softmax}\left(\frac{h_t^\top D_j}{\sqrt{d}}\right), \quad \text{cite}_t = \text{argmax}_j \alpha_j \quad (5)$$

where h_t is the model’s hidden state at position t and D_j is the embedding of retrieved document chunk j . The citation head is trained with a cross-entropy loss against human-annotated citation labels.

5.3 Citation Confidence Score

Alongside the cited document, the model emits a confidence score $c_t = \max_j \alpha_j$. Low citation confidence is correlated with hallucination and can be surfaced to users as a reliability indicator.

6 Experiments

6.1 Main Results (Component Ablation)

Table 1: Direction of effect of each anti-hallucination component on the Zen-32B (Qwen3-32B) model. “↑” indicates improvement on a benchmark (lower hallucination / higher factuality); more arrows indicate a larger relative effect. The full RAFT + UAD + CGG system is strongest on every benchmark.

System	TruthfulQA	FactScore	HaluEval	ZenHallu
Baseline	—	—	—	—
+ RAFT	↑↑	↑↑	↑↑	↑↑
+ UAD	↑	↑	↑	↑
+ CGG	↑	↑	↑	↑
+ RAFT + UAD	↑↑↑	↑↑↑	↑↑↑	↑↑↑
+ RAFT + CGG	↑↑↑	↑↑↑	↑↑↑	↑↑↑
RAFT + UAD + CGG	↑↑↑↑	↑↑↑↑	↑↑↑↑	↑↑↑↑

RAFT contributes the largest single reduction (it directly grounds factual claims in retrieved evidence); UAD and CGG each add a smaller, complementary improvement; and the three combine roughly additively, with the full system reducing hallucination substantially over the baseline on all four benchmarks.

6.2 Scale Dependence

Applying the full RAFT + UAD + CGG system across the 4B, 8B, and 32B dense models, the *absolute* reduction in hallucination is roughly comparable across scales, even though the larger models start from a lower baseline hallucination rate. This indicates the interventions are largely orthogonal to base model capability: they add a fixed factuality benefit on top of whatever the base model already achieves. The relative mix shifts with scale — larger models benefit proportionally more from RAFT (they make better use of retrieved evidence) and less from UAD (they are already better calibrated).

6.3 Domain-Stratified Results

Applying the full system, the largest *relative* hallucination reductions occur in the retrieval-amenable factual domains — medicine, law, and science — where RAFT’s grounding in retrieved evidence has the most to correct. Mathematics, which already had the lowest baseline hallucination rate and depends less on external evidence, shows the smallest relative reduction. The relative-reduction ordering therefore mirrors the baseline-rate ordering: domains that hallucinate most at baseline improve most under retrieval-augmented training.

6.4 Citation Accuracy

For the CGG citation head, both document-level and (the harder) span-level citation accuracy improve with model scale: the 32B dense model identifies the correct source document and span markedly more often than the smaller 4B and 8B variants. Span-level accuracy is consistently lower than document-level accuracy, reflecting the greater difficulty of localizing the exact supporting passage versus the correct document.

6.5 Abstention Rate

Table 2: UAD abstention rate on answerable vs. unanswerable questions.

Category	Abstention rate (%)	Target
Genuinely unanswerable	74.2	>70%
Answerable (false abs.)	4.1	<5%

6.6 Perplexity Impact

A key concern is that anti-hallucination interventions may degrade general language modeling quality. Measuring perplexity on the Pile and a held-out multilingual corpus for the Zen-32B model before and after applying RAFT + UAD + CGG, we find the increase in perplexity is under 1% on both corpora. The interventions therefore improve factuality without a meaningful cost to general language-modeling fluency.

7 Related Work

Hallucination in language models has been studied from multiple angles. Retrieval-augmented generation [4] grounds model outputs in retrieved evidence. RLHF [5] aligns model behavior with human preferences, including factual accuracy. Self-consistency [6] aggregates multiple samples to reduce hallucination via majority vote. Our work differs in targeting hallucination at the training, decoding, and output-structure levels simultaneously.

8 Conclusion

We have presented a three-component system — RAFT, UAD, and CGG — that substantially reduces hallucination in the Qwen3-based Zen models across all tested benchmarks and domains. The system delivers large relative reductions in hallucination with less than 1% perplexity degradation, demonstrating that factual accuracy and fluency are not fundamentally in conflict. RAFT contributes the bulk of the improvement, with UAD and CGG adding complementary gains; we recommend deploying all three components together for production use cases where factual reliability is critical.

References

- [1] S. Lin, J. Hilton, O. Evans. *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. ACL, 2022.
- [2] S. Min, K. Krishna, X. Lyu, et al. *FActScoring: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. EMNLP, 2023.

- [3] J. Li, X. Cheng, W.X. Zhao, J.-Y. Nie, J.-R. Wen. *HaluEval: A Large-Scale Hallucination Evaluation Benchmark for Large Language Models*. EMNLP, 2023.
- [4] P. Lewis, E. Perez, A. Piktus, et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. NeurIPS, 2020.
- [5] L. Ouyang, J. Wu, X. Jiang, et al. *Training Language Models to Follow Instructions with Human Feedback*. NeurIPS, 2022.
- [6] X. Wang, J. Wei, D. Schuurmans, et al. *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. ICLR, 2023.