

Scaling Down Hallucinations: Mitigation Strategies for Compact LLMs

Salef Gabriel Gamberini Silva(✉) and Renato Porfirio Ishii

FACOM, Universidade Federal de Mato Grosso do Sul, Cidade Universitária, Campo Grande, MS 79070-900, Brazil
sgleft@gmail.com

Abstract. The deployment of generative artificial intelligence on consumer hardware is substantially constrained by the tendency of smaller architectures to hallucinate unfaithful or factually incorrect content. This study investigates and categorizes contemporary mitigation strategies through a systematic literature review, classifying them into four distinct groups comprising Model Optimization, Inference and Output Enhancement, External Knowledge and Interaction, and Latent Space Intervention. After applying reproducibility and computational feasibility criteria, eight representative techniques (including Shadow-FT, DoLa, REPLUG, and ITI) were selected for experimental analysis on two representative small language models, Llama-3.2-3B and Qwen3-4B. Evaluations using deterministic reasoning, generative factuality, and summarization benchmarks revealed variable performance across the two models. Specifically, REPLUG was the strongest stabilizer for the tested Qwen3-4B configuration, with no drift events observed in this evaluation. Meanwhile, ITI and Hermes 3 achieved the best results in generative truthfulness for the standard baseline with gains exceeding 30% points, and DoLa demonstrated consistent improvements in source faithfulness. Furthermore, empirical results indicate that mitigation effectiveness is heavily contingent on the interaction between the method and the underlying model. For instance, retrieval acted as a stabilizer for the dual-mode architecture but introduced distractor sensitivity in the standard instruct counterpart. Consequently, in the tested 3B–4B configurations, no single method dominated across tasks, and reliable edge deployment requires matching mitigation strategies to model behavior and latency constraints.

Keywords: Large language models · hallucination mitigation · natural language processing · comparative analysis · machine learning

1 Introduction

Small open-weight Large Language Models (LLMs) in the 3B–4B parameter range enable private and offline execution on consumer hardware. However, their deployment in professional and safety-critical settings remains constrained by

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026, LNCS 16768, pp. 633–650, 2027.
https://doi.org/10.1007/978-3-032-30488-9_39

hallucination. In this work, hallucination is treated as both an extrinsic failure, where the model fabricates facts unsupported by world knowledge, and an intrinsic failure, where the output contradicts a provided source.

The consequences of such failures can be extreme in high-stakes domains. In healthcare, models have been documented producing incorrect diagnostic criteria and fabricated citations [10,17]. In legal practice, generative models have already produced non-existent case citations yielding professional and judicial repercussions [6], reinforcing the need for reliability prior to widespread adoption.

While the literature proposes a variety of strategies designed to mitigate hallucination, evidence remains fragmented for the edge-feasible regime. Existing evaluations typically target large proprietary models, or study them in isolation, raising the question of what strategies remain effective under the constrained compute and representational capacity of small models. This limitation is particularly consequential because the behavior of different model families under identical mitigation techniques may vary widely, not merely in the magnitude of improvement, but also in architectural stability under stochastic decoding.

This study addresses this gap by building a systematic benchmark targeted specifically at the 3B–4B parameter class. A systematic literature review spanning 2021–2025 is conducted and reproducible, computationally feasible techniques are organized into four categories, referred to as Model Optimization, Inference and Output Enhancement, External Knowledge and Interaction, and Latent Space Intervention. Eight representative techniques are reproduced under a unified protocol and evaluated on two small language models, Llama-3.2-3B and Qwen3-4B, selected because they exhibit different baseline stability profiles under stochastic decoding. The evaluation spans deterministic reasoning and domain knowledge (ARC, OAB), generative factuality (TruthfulQA, NQ-Open), and source-faithful summarization (CNN/DailyMail) with the inclusion of paired bootstrap resampling and multi-run stochastic evaluation where applicable.

The empirical findings indicate that mitigation effectiveness is strongly conditioned by task structure and by the interaction between the method and the model under evaluation. In the tested setting, retrieval techniques were shown to act as a stabilizer for Qwen3-4B by eliminating observed instruction drift events (in which the model strays away from the original task) and substantially improving source faithfulness in summarization; in contrast, the same category of mitigation introduced distractors for Llama-3.2-3B on open-domain fact retrieval. Model optimization and latent-space interventions provided complementary benefits where supervised fine-tuning and inference-time steering yielded the largest improvements in generative truthfulness for Llama-3.2-3B, exceeding 30% points on TruthfulQA. Separately, inference-time logit shaping for Qwen3-4b produced consistent improvements in source faithfulness.

The contribution of this work is a reproducible comparative framework for the small-model regime that bridges taxonomic categorization and empirical performance under resource constraints. By contrasting two compact language

models that exhibit distinct empirical stability characteristics, the study provides configuration-level evidence for selecting mitigation strategies according to retrieval sensitivity, deployment cost, and observed stability.

Structurally, this work is organized as follows: Sect. 2 reviews related work concerning hallucination mitigation including the operational directions of modifying decoding dynamics and supplementing generation with retrieved evidence; Sect. 3 details the methodology, outlining the systematic literature review process, the model baselines, and the evaluation protocol; Sect. 4 presents the experimental results across deterministic reasoning, generative factuality, and source faithfulness benchmarks; Sect. 5 discusses the empirical findings, focusing on computational efficiency and deployment trade-offs, while also addressing the study's limitations; Sect. 6 concludes the study and indicates directions for future investigation.

2 Related Work

Recent work on hallucination mitigation has developed along two broad operational directions: modifying the model or its decoding dynamics, and supplementing generation with retrieved evidence or extended context. Within this literature, current comparisons focus on the relative utility of parametric adaptation, retrieval quality, and inference-time controls, as well as on the trade-off between factual gains and computational overhead.

2.1 Fine-Tuning vs. Retrieval

A central design choice in hallucination mitigation is whether factuality should be improved by internalizing domain behavior through parameter updates or by conditioning generation on external evidence. Ovadia et al. [18] reported that fine-tuning can adapt stylistic behavior effectively while remaining less reliable at correcting factual errors, whereas retrieval-based methods often perform better on knowledge-intensive tasks. These findings do not make retrieval universally superior, but they do motivate treating external evidence as a strong baseline whenever knowledge freshness and factual grounding are central.

2.2 Sensitivity of Retrieval Systems

Although retrieval augmentation is often effective, its benefits depend strongly on the quality and presentation of the retrieved context. Cuconasu et al. [3] showed that irrelevant or weakly ranked passages can degrade generation, indicating that more context is not necessarily better. This shifts attention from retrieval alone to the alignment between the retrieved evidence, ranking quality, and the model's ability to use that evidence under finite context budgets.

2.3 Inference-Time Optimization

A separate line of work investigates mitigation at decoding time rather than through external knowledge or parameter updates. Li et al. [13] introduced Contrastive Decoding, arguing that hallucinations can emerge when token selection is dominated by the model prior. By contrasting a stronger model with a weaker or less context-sensitive baseline during generation, their framework suppresses tokens that are fluent yet weakly supported. This perspective recasts hallucination mitigation as a decoding-control problem rather than exclusively a training problem.

2.4 The Efficiency Trade-Off

Despite the proliferation of mitigation techniques, computational overhead is often under-reported relative to factual gains. Surveys on efficient LLM inference [21] note that iterative sampling, long retrieved contexts, and ensemble-like decoding can introduce latency and memory costs that are acceptable in server settings but problematic for on-device use. This matters especially for Small Language Models (SLMs), where limited capacity and narrower attention budgets can change not only the magnitude of improvement but also the failure mode induced by a mitigation strategy. The present study therefore focuses on the 3B–4B regime and evaluates methods as deployable configurations rather than assuming direct transfer from much larger backbones.

Consequently, a lack of systematic evaluation remains regarding the transferability of these methods to **Small Language Models (SLMs)**. While the theoretical efficacy of contrastive decoding or retrieval augmentation is well-documented on hyperscale architectures (e.g., Llama-2-70B), it remains an open question how these techniques interact with the compressed feature spaces and limited attention heads of 3B–4B parameter models.

3 Methodology

This section details the systematic framework employed to identify, categorize, and evaluate hallucination mitigation strategies. The experimental design is strictly confined to the 3B to 4B parameter range to address the specific computational constraints of edge deployment. By prioritizing architectures amenable to local execution on consumer-grade hardware, this study evaluates the trade-off between hallucination suppression and the resource efficiency required for non-server environments.

3.1 Scope and Selection Criteria

To ensure representative coverage of the state of the art, a systematic literature review done according to the guidance of Kitchenham and Charters [8]. The search targeted studies that were published from 2021 through 2025 and

followed a PICOC (Population, Intervention, Comparison, Outcome and Context) framework to delimit scope. Jointly, candidate identification and feasibility screening were carried out jointly across arXiv, GitHub, and Hugging Face. Where arXiv was used for paper discovery, GitHub for implementation discovery and reproducibility checks, and Hugging Face for model-lineage tracing, dataset discovery, and compatibility verification for the selected backbones. Backward and forward citation chaining were used to help minimize bias due to keywords. The screening process was completed in December 2025.

From an initial corpus of 216 candidate studies on hallucination mitigation, methods were retained only if they addressed textual generation, were published between 2021 and 2025, provided accessible open-source implementations, contained sufficient methodological detail for reproduction, and remained feasible in the 3B–4B setting. Multimodal approaches, proprietary API-dependent methods, and techniques lacking usable implementations were excluded, yielding eight methods distributed across four canonical categories (Table 1).

Table 1. Search sources and query templates used in the literature review (screening completed December 2025).

Source	Query template and filters
arXiv	`(all:"LLM" AND (ti:"hallucination" OR ti:"hallucinating" OR ti:"factuality" OR ti:"factualness") AND NOT (ti:"vision" OR ti:"visual" OR ti:"multimodal" OR ti:"multi-modal" OR ti:"MLLM" OR ti:"LMM" OR ti:"VLM" OR ti:"image" OR ti:"video"))`. Results were sorted by submission date and filtered to years 2021–2025 during collection.
GitHub	Repository discovery and reproducibility validation were performed using keyword searches aligned with the primary arXiv terms and with method-name searches for shortlisted techniques (e.g., `hallucination`, `factuality`, `faithfulness`, `decoding`, `retrieval`/`RAG`, `steering`/`intervention`, and method identifiers such as `DoLa`, `REPLUG`, `ITI`).
Hugging Face	Model discovery and verification were performed by traversing model trees and base-model lineage for the selected backbones, using the platform metadata when available (e.g., base model fields and linked derivatives) to locate compatible instruct variants and community releases. Dataset repositories were also consulted to identify the corpora used for retrieval and evaluation.

A single Boolean template could not be transferred uniformly across all three sources because each platform exposes different search behavior and metadata. Candidate identification and feasibility screening were therefore conducted jointly across arXiv, GitHub, and Hugging Face, with GitHub and Hugging Face contributing directly to implementation discovery, compatibility checks, and reproducibility assessment.

3.2 Model Baselines

The evaluation compares two well-established small language models from the contemporary open-weight ecosystem. Rather than relying on a rigid architectural dichotomy, it contrasts models with distinct training and post-training profiles under a common evaluation protocol.

- **Llama-3.2-3B:** This model serves as a conventional instruct-tuned baseline for small-model evaluation. It is included to assess how mitigation strategies interact with a compact model whose behavior is shaped by standard pre-training followed by alignment-oriented post-training.
- **Qwen3-4B:** This model provides a contrasting small-model baseline from the Qwen3 family. It is included to assess whether the same mitigation strategies behave differently on a model with a distinct post-training recipe and response profile. In all experiments reported here, Qwen3-4B was evaluated in its non-thinking configuration.

3.3 Mitigation Configurations

To keep the benchmark grounded in realistic deployment settings, we instantiate one representative technique for each backbone within each mitigation category. The objective is not to provide a symmetric method-by-method comparison, but rather to compare deployable configurations under a common evaluation protocol.

This design reflects the practical constraint that several mitigation methods are mature only within particular implementation stacks. Accordingly, our conclusions are drawn at the level of the evaluated configurations and the trends observed across architectures, rather than as universal claims about an entire mitigation category.

Model Optimization. Model optimization approaches reduce hallucinations by updating parameters during fine-tuning or continued training, aiming to internalize more reliable generation behavior beyond transient inference-time controls.

- **Shadow-FT (Qwen3-4B):** Direct fine-tuning can degrade the behavior of already aligned small models when useful capabilities are overwritten during adaptation. To mitigate this risk, Shadow-FT [22] is employed. This method uses the unaligned base model as a proxy learner, computes weight updates (ΔW) on malleable parameters, and grafts them onto the instruct variant in order to preserve the model's pre-existing alignment.
- **Hermes 3 Strategy (Llama-3.2-3B):** For Llama-3.2-3B, the Hermes 3 methodology [20] is evaluated as a benchmark for rigorous Supervised Fine-Tuning (SFT). This approach uses aggressive data filtering and synthetic reasoning tasks (e.g., `<SCRATCHPAD>` tokens) to condition the model toward more reliable generation behavior during training.

Inference and Output Enhancement. Inference-time enhancements reshape token logits during decoding by contrasting context-conditioned predictions with a context-free prior, downweighting tokens that are plausible under the prior but unsupported by evidence.

- **DoLa (Qwen3-4B):** Decoding by Contrasting Layers (DoLa) [1] contrasts final-layer logits against earlier-layer logits during generation, attenuating tokens that become plausible due to premature semantic drift.
- **Context-Aware Decoding (Llama-3.2-3B):** CAD [19] is implemented by combining logits from a context prompt and a no-context prompt using $\ell_{\text{CAD}} = (1+\alpha)\ell_{\text{ctx}} - \alpha\ell_{\text{nc}}$ with default $\alpha = 0.5$. The no-context input is instantiated as a BOS or empty prompt unless a negative prompt is specified.

External Knowledge and Interaction. External evidence is incorporated at inference time to mitigate limitations of parametric memory. Task-specific corpora are sourced from Hugging Face and indexed for retrieval: `pszemraj/simple_wikipedia_LM` for NQ-Open and TruthfulQA; `allenai/sciq` for ARC; and a curated collection of ten Brazilian legal codes (e.g., Constitution, Civil Code, Penal Code) for OAB-style questions. For CNN/Daily Mail summarization, retrieval is restricted to the input article only (no external corpus).

- **REPLUG (Qwen3-4B):** For Qwen3-4B, $k = 5$ evidence units are retrieved per query (BM25 over a tokenized corpus in the provided implementation). The model is evaluated with independent forward passes conditioned on each retrieved unit, and aggregation is performed across units using a score-combination rule. In log-likelihood based evaluation, the best (max) continuation score across the k contexts is used; in OAB multiple-choice evaluation, the probabilities of the option tokens (A/B/C/D) are averaged across the k contexts. Prompts are truncated to a maximum length of 2048 tokens.
- **Task-Dependent RAG (Llama-3.2-3B):** For Llama-3.2-3B, external retrieval is implemented as a task-dependent RAG configuration with two task-specific variants. In reasoning and QA settings, a naive concatenative variant uses `BAAI/bge-large-en-v1.5` embeddings with L2 normalization and a FAISS `IndexFlatIP` GPU index, retrieves the top-k passages with $k = 3$, filters by similarity threshold 0.50, and "prepends" the retained passages as reference material before the original query. Teacher-forced log-likelihood is then computed on the target conditioned on the query plus retrieved evidence. In summarization, an iterative source-confined variant restricts retrieval to the source document itself and uses retrieval only to bias content selection without introducing external knowledge. In the results tables, these variants are reported as Naive RAG and Iterative RAG, respectively.

Latent Space Intervention. Latent-space interventions modify intermediate activations during the forward pass by adding targeted offsets to internal representations, biasing generation toward directions associated with truthfulness without updating model weights.

- **Gaussian Steering (Qwen3-4B):** Steering is implemented with forward hooks on transformer blocks. Layer-specific control directions loaded from a serialized vector file are added to the block output with a Gaussian depth schedule centered at $\mu = 18$. For block index i, the coefficient is $\texttt{strength} \cdot \exp\left(-\frac{(i-\mu)^2}{2\sigma^2}\right)$, applying only to layers with an associated direction and skipping coefficients below 10^{-4}.
- **Inference-Time Intervention (Llama-3.2-3B):** ITI [12] is instantiated by training linear probes over attention-head activations using TruthfulQA multiple-choice features. At inference, activations of the top-K heads with $K = 48$ are shifted along the probe-identified truthful directions with intervention strength $\alpha = 15$.

3.4 Evaluation Instruments

The assessment protocol employs a diverse suite of benchmarks to evaluate reasoning, factuality, and summarization quality.

Deterministic Reasoning. For tasks requiring rigorous logic, log-likelihood evaluation is performed to eliminate sampling variance.

- **ARC-Challenge:** Selected to assess the model's ability to perform complex reasoning in scientific contexts beyond superficial pattern matching [2].
- **OAB Exam:** A dataset derived from the Brazilian Bar Exam [5] is included to evaluate performance in a non-English (Portuguese) professional context. This tests the mitigation methods' ability to handle specialized legal terminology and maintain cross-lingual logical consistency.

Generative Factuality.

- **TruthfulQA:** This benchmark is critical for quantifying "imitative falsehoods" [15]. The generative track is used to assess the model's resistance to common misconceptions, with outputs graded by a specialized fine-tuned judge model.
- **NQ-Open:** To evaluate precise information retrieval, NQ-Open [11] is employed in a 5-shot in-context learning setting. Performance is measured via Exact Match (EM) against human-annotated ground truths.

Summarization Quality. For long-form generation, a stratified sample ($N = 1,000$) of the **CNN/Daily Mail** test set is utilized. Evaluation relies on two complementary metrics:

- **ROUGE-L:** Captures structural fidelity and longest common subsequences [14].

– **BERTScore (F1):** Addresses the limitations of lexical overlap by measuring semantic similarity via contextualized embeddings [23]. The F1 measure is reported exclusively to balance the penalization of hallucinations (precision) with the retention of critical details (recall).

3.5 Statistical Analysis

To quantify performance differences, a paired statistical protocol was adopted at the instance level.

Significance Testing. Traditional parametric tests assume normally distributed errors, an assumption frequently violated in NLP evaluation [4]. Consequently, **Paired Bootstrap Resampling** [9] was adopted with $B = 10{,}000$ resamples. For deterministic benchmarks (ARC and OAB), each evaluation item constitutes one paired observation between a mitigated model and its corresponding backbone baseline. Significance is reported when the 95% confidence interval of the paired performance difference excludes zero.

Stochastic Stability. For generative benchmarks (TruthfulQA, NQ-Open, CNN/DM), single-pass evaluation is insufficient because of decoding stochasticity. Following the motivation of Holtzman et al. [7], $K = 5$ independent inference runs were conducted for each instance using architecture-specific sampling parameters. The per-instance score was defined as the mean over the K runs, $\bar{S}_i$, and significance testing was then applied to the vector of these per-instance means. This preserves the paired structure while reducing variance induced by decoding randomness.

4 Results

Evaluating hallucination mitigation approaches against the benchmarks defined in Sect. 3 reveals distinct performance trade-offs. The following analysis compares these methods quantitatively, providing empirical evidence for their potential and limitations in reducing LLM hallucinations.

4.1 Execution Environment

Experiments were benchmarked on NVIDIA A100 GPUs using the Hugging Face `transformers` library. This setup was chosen to accelerate repeated runs and standardize evaluation, not because the selected methods depend on datacenter-only hardware. The evaluated models remain in the 3B–4B range and the selected configurations were also verified to execute on consumer-grade hardware, so the A100 should be understood as a convenient benchmarking platform rather than as a deployment requirement. For consistency, `lm-evaluation-harness` was used for ARC and NQ-Open, custom scripts were used for OAB and BERTScore,

and TruthfulQA was evaluated with its specialized generative judge. Deterministic benchmarks (ARC and OAB) were run once, whereas stochastic benchmarks (TruthfulQA, NQ-Open, and CNN/DailyMail) were averaged over five independent runs ($K = 5$) to reduce decoding variance.

Table 2. Reasoning and Domain Knowledge Performance. Comparison of reasoning (ARC) and legal domain knowledge (OAB). Best results per model are **highlighted**. ($*$: 95% paired bootstrap CI excludes zero in favor of the method, †: 95% paired bootstrap CI excludes zero in favor of the baseline).

Method	ARC Challenge	ARC Easy	OAB (Legal)
Qwen3-4B			
Baseline	0.5392	0.7816	0.2742
Shadow-FT	0.5341	0.7555†	0.2656
DoLa	0.5427	0.7841	0.2796
REPLUG	0.5444	**0.7988***	**0.4014***
GCS (Steered)	**0.5589***	0.7891*	0.2810
Llama-3.2-3B			
Baseline	0.4266	0.7146	0.2837
Hermes 3	0.4283	0.7184	**0.3534***
CAD	**0.4889***	0.7193	0.2851
Naive RAG	0.4309	**0.7652***	0.3317*
ITI	0.4249	0.7433*	0.2873

4.2 Reasoning and Domain Knowledge

Table 2 illustrates the divergence between abstract logic (ARC) and domain-specific knowledge (OAB). Within Model Optimization, the Hermes 3 strategy significantly improved Llama-3.2-3B on OAB. Due to the predominantly English alignment data, this result suggests possible cross-lingual transfer to Portuguese legal reasoning, but does not establish language-independent factual subspaces. A similar but smaller pattern appears for ITI, which improves OAB despite being learned from English TruthfulQA features. Conversely, Shadow-FT reduced ARC-Easy for Qwen3-4B, indicating that weight grafting can trade off against zero-shot reasoning on simpler science questions.

Inference-time controls were sensitive to task structure. Context-Aware Decoding (CAD) produced the largest gain for Llama-3.2-3B on ARC-Challenge, consistent with the idea that subtracting prior-driven logits can help when shallow associations compete with compositional reasoning. DoLa, by contrast, produced only small changes for Qwen3-4B on the deterministic reasoning tasks,

suggesting limited benefit in this setting rather than a general failure of layer-contrast methods.

Retrieval-based methods showed their clearest gains on OAB. For Qwen3-4B, REPLUG substantially improved legal-domain accuracy. For Llama-3.2-3B, Task-Dependent RAG (naive) improved OAB and ARC-Easy, but did not materially improve ARC-Challenge. These patterns suggest that retrieval is most effective when the target information is externally recoverable, whereas deductive items remain less responsive to added context. Finally, GCS improved Qwen3-4B across the reasoning benchmarks, suggesting that activation steering can provide gains without external evidence.

4.3 Generative Factuality

Table 3. Generative Factuality (Stochastic). Mean accuracy ± standard deviation across $K = 5$ independent runs. Best results per model are highlighted. Asterisk indicates that the 95% paired bootstrap CI excludes zero in favor of the method; dagger indicates that the 95% paired bootstrap CI excludes zero in favor of the baseline.

Method	NQ-Open (EM)	TruthfulQA (Gen)
QWEN3-4B		
Baseline	0.1454 ± 0.0020	0.6911 ± 0.0042
Shadow-FT	0.1511 ± 0.0024*	0.7028 ± 0.0069
DoLa	0.1502 ± 0.0029	0.7053 ± 0.0039
REPLUG	0.1405 ± 0.0020†	0.7266 ± 0.0055*
GCS (Steered)	0.1498 ± 0.0035*	0.7364 ± 0.0044*
LLAMA-3.2-3B		
Baseline	0.1739 ± 0.0044	0.2095 ± 0.0189
Hermes 3	0.1949 ± 0.0058*	0.5356 ± 0.0134*
CAD	0.1824 ± 0.0032*	0.2480 ± 0.0126*
Naive RAG	0.1214 ± 0.0052†	0.4196 ± 0.0104*
ITI	0.1967 ± 0.0051*	0.5302 ± 0.0119*

Table 3 highlights a fundamental disparity in model initialization. Hermes 3 improved Llama-3.2-3B on TruthfulQA by over 30% points. Given the weak factuality of the Llama baseline in this setting, this result is consistent with improved truthful generation under supervised fine-tuning. In contrast, Qwen3-4B began with high native factuality, so optimization gains are smaller but stable.

External knowledge methods present a trade-off. While Naive RAG improved TruthfulQA, it caused a significant degradation in NQ-Open for Llama-3.2-3B.

To determine whether this drop stems from over-reliance on noisy retrieval or from context window limitations, a qualitative audit of 50 RAG failures was conducted.

Context distraction is used as the umbrella label, and it most often manifests as entity contamination and position bias, with a smaller fraction attributable to window truncation.

Table 4. Qualitative RAG error modes for Llama-3.2-3B with Naive RAG on NQ-Open. Retrieved snippets are abbreviated. Underline marks the entity the model latched onto.

Query	Target	Distractor cue excerpt	RAG Output	Mode
who sings she's like the wind lyrics	Wendy Fraser	Wind It Up is a song by American singer Gwen Stefani, mention of Vanessa Williams appears in retrieved text	Vanessa Williams	Entity contamination
who are the hobbits in lord of the rings	Dominic Monaghan	Cast list begins with Elijah Wood as Frodo, later mentions Dominic Monaghan as Merry	Elijah Wood	Position bias [16]
what are the band members names of the rolling stones	Ronnie Wood	The band members were Mick Jagger, Keith Richards, Brian Jones, Bill Wyman, retrieved list omits later members	Mick Jagger, Keith Richards, Brian Jones, Bill Wyman	Incomplete evidence, truncation prone
when was the first mad max released	12 April 1979	Mad Max is a 1979 Australian action film, day and month missing	1979	Formatting under strict EM

Table 4 exhibits why NQ-Open can degrade under Naive RAG for a small model. When retrieval surfaces semantically proximal but off-target text the model often copies a salient named entity from the snippet. When the retrieved passage is long or "list-like", early entities can dominate attention and suppress later relevant mentions. When the retrieved snippet is incomplete or clipped to fit the budget, it becomes misleading evidence that the model treats as exhaustive. These patterns explain how Naive RAG can raise TruthfulQA when retrieval frequently contains directly usable evidence, yet hurt NQ-Open when retrieval noise and presentation effects dominate.

Latent-space interventions were among the more consistent methods in this setting. GCS improved Qwen3-4B across both benchmarks, and ITI approached the gains of Hermes 3 for Llama-3.2-3B. ITI modifies activations only at inference time, so these results suggest that useful truthfulness-related representations can sometimes be elicited without permanent weight updates.

4.4 Source Faithfulness

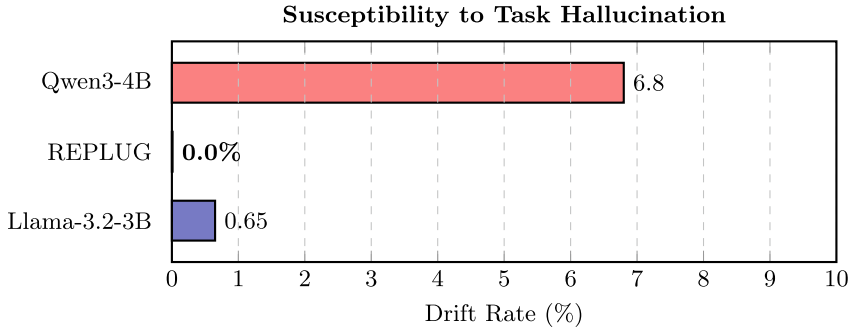

Fig. 1. Instruction drift rates under repeated non-deterministic decoding for CNN/DailyMail summarization ($N = 1000$ sources, $R = 5$ runs per source). Llama-3.2-3B is included as a stability reference. Qwen3-4B exhibits measurable drift, whereas REPLUG yields zero observed drift events in 5000 generations, implying a non-zero upper confidence bound under a naive Wilson approximation.

The assessment of source faithfulness (Table 5) indicates that mitigation effects depend strongly on baseline stability. Figure 1 reports instruction drift for the unmitigated Qwen3-4B, REPLUG-augmented Qwen3-4B, and Llama-3.2-3B as a stability reference rather than as a method-matched control. The unmitigated Qwen3-4B drifted in 6.80% of generations (340/5000). Under a naive independence approximation, the corresponding 95% Wilson interval is $[6.13, 7.53]$ percent. Given that the 5000 generated samples are obtained from $N = 1000$ distinct sources with $R = 5$ runs per source, the resulting interval should be regarded as descriptive and is likely to be optimistically narrow.

In the same setting, REPLUG produced zero observed drift events (0/5000), corresponding to 0.00% with a 95% Wilson upper bound of 0.08% under the same approximation. This pattern is consistent with REPLUG acting as a stabilizing mechanism for the tested Qwen3-4B configuration (although the zero observed drift rate should not be interpreted as a guarantee of zero drift in deployment). The ensemble over retrieved units constrains the decoding trajectory and reduces run-to-run variability.

Llama-3.2-3B showed high baseline stability (0.65% drift). In that context, Task-Dependent RAG (iterative) behaved less like a stabilizer and more like a content-selection bias. It reduced ROUGE-1 while improving BERTScore, suggesting a shift away from reference-like lexical phrasing toward semantically closer but stylistically less matched summaries.

Regarding inference enhancement, DoLa was effective for the less stable Qwen3-4B in this setting, yielding consistent improvements across the reported summarization metrics. This pattern is consistent with layer contrast helping suppress tokens associated with semantic drift during generation. Conversely,

Table 5. Source Faithfulness on CNN/DailyMail. Comparison of lexical (ROUGE) and semantic (BERTScore) alignment. Best methods per model are **highlighted**. (∗: 95% paired bootstrap CI excludes zero in favor of the method, †: 95% paired bootstrap CI excludes zero in favor of the baseline).

	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore (F1)
Qwen3-4B				
Baseline	0.2388 ± 0.0028	0.0818 ± 0.0018	0.1643 ± 0.0015	0.1704 ± 0.0041
DoLa	0.2617 ± 0.0008*	0.0877 ± 0.0007*	0.1748 ± 0.0004*	0.2074 ± 0.0008*
REPLUG	**0.2881** ± 0.0010*	**0.0975** ± 0.0004*	**0.1909** ± 0.0006*	**0.5863** ± 0.0003*
Shadow-FT	0.2329 ± 0.0006†	0.0795 ± 0.0005†	0.1547 ± 0.0006†	0.1834 ± 0.0011*
GCS (Steered)	0.2013 ± 0.0008†	0.0640 ± 0.0008†	0.1384 ± 0.0006†	0.1079 ± 0.0014†
Llama-3.2-3B				
Baseline	0.2799 ± 0.0101	0.1032 ± 0.0046	0.1929 ± 0.0063	0.5466 ± 0.0100
CAD	0.2193 ± 0.0019†	0.0705 ± 0.0015†	0.1731 ± 0.0018†	0.5183 ± 0.0019†
Hermes 3	0.2960 ± 0.0014*	0.1122 ± 0.0005*	0.1992 ± 0.0010*	0.5888 ± 0.0007*
ITI	**0.2991** ± 0.0015*	**0.1147** ± 0.0005*	**0.2023** ± 0.0012*	**0.5919** ± 0.0009*
Iterative RAG	0.2550 ± 0.0045†	0.0867 ± 0.0015†	0.1723 ± 0.0030†	0.5727 ± 0.0031*

CAD caused a significant regression for Llama-3.2-3B, including a 6.06-point drop in ROUGE-1, and the auxiliary length analysis is consistent with under-generation. Early termination rises to 19.2% as measured by EOS within the first 30 generated tokens, corresponding to the baseline 10th-percentile length. Outputs therefore shift toward shorter summaries with a mean of 67 and a median of 76 tokens, while hitting the 100-token cap remains rare at 1%. In this setting, logit subtraction appears to suppress continuation paths that would otherwise elaborate later source details. By contrast, Llama-3.2-3B responded positively to directional steering through Hermes 3 and ITI, suggesting that the more stable baseline benefited more from internal realignment than from stronger decoding constraints.

5 Discussion

The experimental results suggest that several factors must be considered when mitigating hallucinations within the 3B–4B parameter regime including how much model plasticity can be allowed while still maintaining stability with regard to training instructions. Beyond the raw accuracy metrics analyzed in the previous section, the viability of these methods for edge deployment depends heavily on computational scalability and model-specific interactions.

A critical factor for mobile implementation is the inference cost associated with each mitigation category. The results show that performance gains often come at the expense of latency.

Retrieval-based methods such as REPLUG demonstrate the highest computational overhead. Since the ensemble strategy requires processing multiple retrieved contexts, the decoding compute scales approximately linearly with the number of retrieved passages. This yields an approximate $\sim k\times$ model-compute multiplier for k retrieved contexts, with $k = 5$ implying up to $\sim 5\times$ the base decoding cost, in addition to retriever latency and the attention cost induced by longer prompts.

For smartphones, this overhead may be unsuitable for interactive use unless k is reduced, retrieval is cached, or REPLUG is reserved for high-risk or non-real-time queries. In contrast, Latent Space Interventions such as ITI and GCS offer a more favorable efficiency profile. These methods operate by modifying activation vectors during a single forward pass and therefore add only a small constant-factor overhead, remaining near-$1\times$ relative to baseline inference.

Inference-time decoding interventions further illustrate the cost spectrum. DoLa operates within a single forward pass with lightweight additional logit operations, whereas CAD requires both a context and a no-context computation and therefore incurs an approximate $\sim 2\times$ forward-pass overhead. For environments where response time is critical, activation-level steering and single-pass decoding controls provide more scalable mitigation than multi-context ensembling.

Deployment should therefore be model and latency specific. In the present experiments, Qwen3-4B exhibited higher plasticity together with substantially greater instruction drift, making external stabilization particularly useful despite its higher inference cost. In that setting, lightweight interventions alone were less reliable at preventing deviation into hallucinatory patterns. Conversely, Llama-3.2-3B showed stronger baseline stability but also greater sensitivity to irrelevant retrieved context. For that configuration, internal alignment techniques such as supervised fine-tuning or inference-time steering made better use of the existing stability without introducing additional distractors, which in turn makes them more suitable for latency-sensitive edge use than naive retrieval.

5.1 Limitations

Several constraints limit the generality of these findings.

The study relies on automated evaluation pipelines, which in turn introduce a degree of uncertainty into the resulting metrics. In particular, the TruthfulQA generative judge is itself a learned model and may encode preferences that only approximate factual robustness.

The compared mitigation strategies intervene at different stages, with some requiring training-time updates and others operating purely at inference. As a result, the experiments characterize realistic deployment options rather than a fully controlled comparison under identical training data and optimization budgets. The observed patterns should therefore be interpreted as configuration-level evidence, not as definitive rankings of mitigation categories.

In addition, the evaluation emphasizes atomic facts and short-horizon deductions in order to isolate hallucination behavior from longer context-based reten-

tion effects. However, for summarization, ROUGE and BERTScore measure lexical and semantic alignment. They do not evaluate whether the summaries are factually consistent. Therefore, an increase in score indicates an improvement in source alignment rather than an absolute guarantee that subtle contradictions will not occur.

Finally, cross-lingual evaluation was limited to Portuguese legal questions. The OAB results suggest transfer in this setting, but do not establish similar behavior for other languages, scripts, or domains.

6 Conclusions

This study combined a systematic literature review with quantitative evaluation to analyze hallucination mitigation in 3B–4B language models.

In response to the categorization inquiries, the review identified a diverse landscape of approaches classified into four primary categories: 1) Model Optimization; 2) Inference and Output Enhancement; 3) External Knowledge and Interaction; and 4) Latent Space Intervention. Consistent with previous observations, there remains a prevalence of techniques in the Inference and External Knowledge categories, reflecting a research trend toward modular solutions that avoid the computational expense of retraining.

Addressing comparative efficacy, the results revealed that performance is not merely a function of the mitigation category, but of the interaction between the method and the model under evaluation. Model Optimization proved most effective for restoring lost capabilities and enabling cross-lingual transfer in domain-specific tasks such as the OAB exam. Conversely, External Knowledge and Interaction functioned as a effective stabilizer for the Qwen3-4B configuration, producing zero observed drift events in this evaluation setting, but acted as a distractor for the Llama-3.2-3B configuration, occasionally degrading performance in closed-book query tasks. Furthermore, Latent Space Interventions appeared promising in this benchmark because they improved truthfulness with low added inference overhead, contrasting sharply with the high computational cost of ensemble retrieval methods.

The primary contribution of this work lies in connecting mitigation taxonomy with empirical evaluation under small-model constraints. By contrasting two representative compact language models under a unified protocol, this study provides evidence that mitigation strategies should be selected according to model behavior, task requirements, and computational budget. This offers practical guidance for balancing accuracy, stability, and efficiency in edge-oriented deployments.

Based on these findings, several directions for future investigation are indicated. First, the phenomenon where retrieval augmentation degrades the performance of stable small models warrants a deeper mechanistic analysis to understand how 3B-parameter models filter context noise. Second, while this study focused on textual hallucinations, the framework should be extended to Multimodal Large Language Models (MLLMs) to evaluate if latent space interventions remain robust across visual-textual modalities. Finally, while this study

addressed statistical significance through bootstrap resampling, future work could expand the scope to include "in-the-wild" human evaluation to correlate these automated metrics with perceived user utility in real-time mobile applications.

Acknowledgments. We acknowledge the support of UFMS (Universidade Federal de Mato Grosso do Sul), FUNDECT, and Finep. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of UFMS, FUNDECT, or Finep.

Disclosure of Interests. The authors have no financial or proprietary interests in any material discussed in this article.

References

1. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J., He, P.: DOLA: decoding by contrasting layers improves factuality in large language models. arXiv (2024)
2. Clark, P., et al.: Think you have solved question answering? Try ARC, the AI2 reasoning challenge (2018). https://arxiv.org/abs/1803.05457
3. Cuconasu, F., et al.: The power of noise: redefining retrieval for rag systems. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, pp. 719–729. ACM (2024). https://doi.org/10.1145/3626772.3657834
4. Dror, R., Baumer, G., Shlomov, S., Reichart, R.: The hitchhiker's guide to testing statistical significance in natural language processing. In: Gurevych, I., Miyao, Y. (eds.) Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1383–1392. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1128. https://aclanthology.org/P18-1128/
5. Garcia, E.: OAB exams (2024). https://huggingface.co/datasets/eduagarcia/oab_exams
6. Hemrajani, R.: Evaluating the role of large language models in legal practice in India (2025). https://arxiv.org/abs/2508.09713
7. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration (2020). https://arxiv.org/abs/1904.09751
8. Kitchenham, B., Charters, S.: Guidelines for performing systematic literature reviews in software engineering. EBSE Technical report (2007)
9. Koehn, P.: Statistical significance tests for machine translation evaluation. In: Lin, D., Wu, D. (eds.) Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 388–395. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-3250/
10. Koranteng, E., et al.: Empathy and equity: key considerations for large language model adoption in health care. JMIR Med. Educ. **9** (2023). https://api.semanticscholar.org/CorpusID:266571217
11. Lee, K., Chang, M.W., Toutanova, K.: Latent retrieval for weakly supervised open domain question answering. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 6086–6096. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1612. https://aclanthology.org/P19-1612/

12. Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M.: Inference-time intervention: eliciting truthful answers from a language model (2024). https://arxiv.org/abs/2306.03341
13. Li, X.L., et al.: Contrastive decoding: open-ended text generation as optimization (2023). https://arxiv.org/abs/2210.15097
14. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-1013
15. Lin, S., Hilton, J., Evans, O.: Truthfulqa: measuring how models mimic human falsehoods. arXiv:2109.07958 (2022)
16. Liu, N.F., et al.: Lost in the middle: how language models use long contexts (2023). https://arxiv.org/abs/2307.03172
17. Naik, H.R., Prather, A.D., Gurda, G.T.: Synchronous bilateral breast cancer: a case report piloting and evaluating the implementation of the ai-powered large language model (LLM) chatgpt. Cureus **15** (2023). https://api.semanticscholar.org/CorpusID:258161464
18. Ovadia, O., Brief, M., Mishaeli, M., Elisha, O.: Fine-tuning or retrieval? Comparing knowledge injection in LLMs. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pp. 237–250. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.emnlp-main.15. https://aclanthology.org/2024.emnlp-main.15/
19. Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L., tau Yih, S.W.: Trusting your evidence: hallucinate less with context-aware decoding (2023). https://arxiv.org/abs/2305.14739
20. Teknium, R., Quesnelle, J., Guang, C.: Hermes 3 technical report (2024). https://arxiv.org/abs/2408.11857
21. Wan, Z., et al.: Efficient large language models: a survey (2024). https://arxiv.org/abs/2312.03863
22. Wu, T., et al.: Shadow-ft: tuning instruct model via training on paired base model (2025). https://arxiv.org/abs/2505.12716
23. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: evaluating text generation with BERT. arXiv:1904.09675 (2020)