# HalluGuard: Evidence-Grounded Small Reasoning Models to Mitigate Hallucinations in Retrieval-Augmented Generation

**Loris Bergeron**[*1,4] **Ioana Buhnila**[*2,3] **Jérôme François**[4] **Radu State**[4]
[1]Banque de Luxembourg [2]Center for Data Science in Humanities, Chosun University
[3]ATILF, University of Lorraine–CNRS [4]SnT, University of Luxembourg
**Correspondence:** loris.bergeron@blu.bank

## Abstract

Large Language Models excel at NLP tasks but remain prone to hallucinations, limiting trust in real-world applications. We present HalluGuard, a 4B-parameter Small Reasoning Model (SRM) designed as a guardrail for Retrieval-Augmented Generation (RAG) pipelines, which classifies document-claim pairs as grounded or hallucinated in closed-book, document-grounded settings and produces evidence-grounded justifications. Our approach combines (i) a domain-agnostic synthetic dataset derived from FineWeb and refined through multi-stage curation and data reformation, (ii) synthetic grounded and hallucinated claims, and (iii) preference-based fine-tuning with Odds Ratio Preference Optimization (ORPO) to distill large-model reasoning into a smaller backbone. On the RAGTruth subset of the LLM-AggreFact benchmark, HalluGuard achieves 84.4% balanced accuracy (BAcc), surpassing specialized models, MiniCheck (7B; 84.0%) and Granite Guardian 3.3 (8B; 82.2%) while using roughly half their parameters. Across the benchmark, it reaches 77.1% BAcc, surpassing larger general-purpose LLMs such as GPT-4o (75.9%). HalluGuard and its datasets are available on Hugging Face[1].

## 1 Introduction

Large Language Models (LLMs) have been used for a variety of Natural Language Processing (NLP) tasks, achieving strong results in summarization, text classification, and question answering (Tan et al., 2023; Singhal et al., 2023).

However, recent research shows that Small Language Models (SLMs) (Schick and Schütze, 2021) can achieve competitive results in specific tasks, especially when fine-tuned on domain-specific data. In addition to being cost and energy efficient, SLMs are practical in resource-constrained settings (Lepagnol et al., 2024) such as on-premise environments, often required in the financial sector and industries with strict compliance requirements.

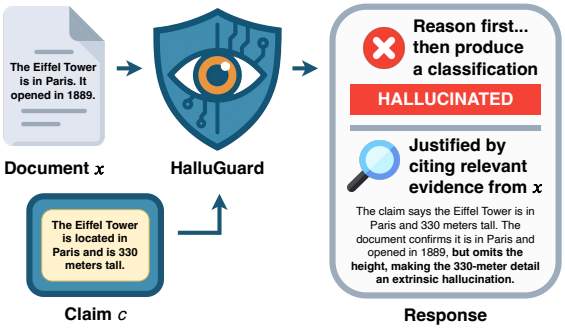


Figure 1: **HalluGuard Concept.** In a closed-book setting, HalluGuard is given a document $x$ and a claim $c$ and only reasons over $x$ to determine whether the claim is `grounded` or `hallucinated`, and produces a justification by citing evidence from the document.

However, a major remaining challenge is that both LLMs and SLMs are prone to hallucinations, outputs inconsistent with the input prompt or factual knowledge (Zhang et al., 2025a; Huang et al., 2023), and are problematic in Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) pipelines, increasingly deployed in companies due to their ability to deliver context-aware responses.

Even when using documents, RAGs remain vulnerable to hallucinations (Niu et al., 2024), compromising trust and explainability (Ni et al., 2025). To address this, models must be able to detect hallucinations and justify their outputs with evidence while being integrated into RAG pipelines.

Recent work emphasizes models designed for reasoning, the ability to perform multi-step inference, follow logical chains, and provide transparent reasoning traces (Wei et al., 2022). Small Reasoning Models (SRMs) are not merely SLMs run with Chain-of-Thought (CoT) prompts. Rather, they are trained to produce structured intermediate reasoning that decomposes complex tasks before generating output, often through distillation from stronger reasoners and reward-guided training (Wang et al.,

[*]These authors contributed equally.
[1]https://hf.co/collections/lrsbrgrn/halluguard



*Findings of the Association for Computational Linguistics: ACL 2026*, pages 16918–16932

2025). This makes SRMs particularly well suited for mitigating hallucinations in RAG pipelines.

Moreover, most of the previous work on hallucination detection uses BERT-based classifiers (Devlin et al., 2019). Although effective, these models do not provide justifications, making them unsuitable when explainability is mandatory.

At the same time, companies deploy custom RAG pipelines with team-specific knowledge (e.g., finance) to improve business efficiency. In these settings, users must see which passages of the retrieved document support or do not support a claim.

To address this gap, we propose HalluGuard, an SRM for document-grounded hallucination detection in a closed-book setting. Although HalluGuard is designed to operate within Retrieval-Augmented Generation (RAG) pipelines, it performs post-generation verification, treating the retrieved document as the only source of knowledge. As shown in Figure 1, given a document $x$ and a claim $c$, HalluGuard only reasons over the content of $x$ to predict whether the claim is `grounded` or `hallucinated`, while producing an evidence-grounded justification by citing relevant passages of the document, fostering end-user trust.

Our contributions are threefold:

- We release HalluGuard, a 4B-parameter Small Reasoning Model (SRM) designed as a closed-book guardrail for Retrieval-Augmented Generation (RAG), detecting document-grounded hallucinations and producing evidence-grounded justifications.
- We release HalluClaim, a large-scale synthetic dataset derived from FineWeb (Penedo et al., 2024) for training and evaluating document-based hallucination detection.
- We demonstrate that HalluGuard achieves competitive performance compared to larger open-source and closed LLMs. Our ablation study highlights the role of reasoning traces, consensus filtering, and Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024) fine-tuning in driving these gains.

## 2 Related Work

Mitigating hallucinations in LLMs has been approached through prompt engineering, Retrieval-Augmented Generation, decoding strategies, supervised fine-tuning, and self-reflection (Ji et al. 2023; Song et al. 2024; Tonmoy et al. 2024; Zhang et al. 2025b). Despite extensive prior work, there is no consensus on hallucination taxonomy, as the distinction between hallucination and factuality remains blurred (Wei et al. 2024; Mallen et al. 2023). Bang et al. (2025) proposed a conceptual separation between intrinsic hallucination (inconsistency with training data), extrinsic hallucination (inconsistency with the input context), and factuality (reliance on world knowledge). We adopt a document-centric variant of this taxonomy (see Section 3).

Several benchmarks focus on hallucination and factuality evaluation. LLM-AggreFact (Tang et al., 2024a) and AVERITEC (Schlichtkrull et al., 2023) provide human-annotated claims paired with evidence, while FEVEROUS (Aly et al., 2021) focuses on Wikipedia-based verification. In contrast, HalluClaim is derived from FineWeb, enabling coverage of a wider range of topics and sources.

Fact-checking models have also been proposed for the detection of hallucinations. MiniCheck (Tang et al., 2024a) introduced a 7B model trained on synthetic data, highlighting the effectiveness of compact models. IBM's Granite Guardian 3.3 (Padhi et al., 2024), an 8B model, detects hallucinations in RAG settings and provides binary predictions with optional reasoning traces through hybrid thinking modes. More recently, HaluCheck (Pandit et al., 2025) applied curriculum-based Direct Preference Optimization (DPO) (Rafailov et al., 2023) for hallucination detection[2].

Despite these advances, existing approaches do not consistently produce evidence-grounded justifications for closed-book, document-grounded hallucination detection, motivating transparent and efficient reasoning-based models at a compact scale.

## 3 Problem Formulation

We consider a closed-book setting in which the document $x$ is the only source of truth to evaluate a claim $c$. Under this assumption, the task is to determine the relationship $t(x, c)$ between a document $x$ and a claim $c$, which can take one of three labels:

$$t(x,c) = \begin{cases} \text{Grounded} & \text{if } c \text{ is supported by } x \\ \text{Intrinsic}_{\text{Hallu}} & \text{if } c \text{ contradicts } x \\ \text{Extrinsic}_{\text{Hallu}} & \text{if } c \text{ is not in } x \end{cases}$$

A claim is labeled `Grounded` when it is fully supported by the information explicitly stated in $x$.

[2]Not publicly available at the time our study was conducted.

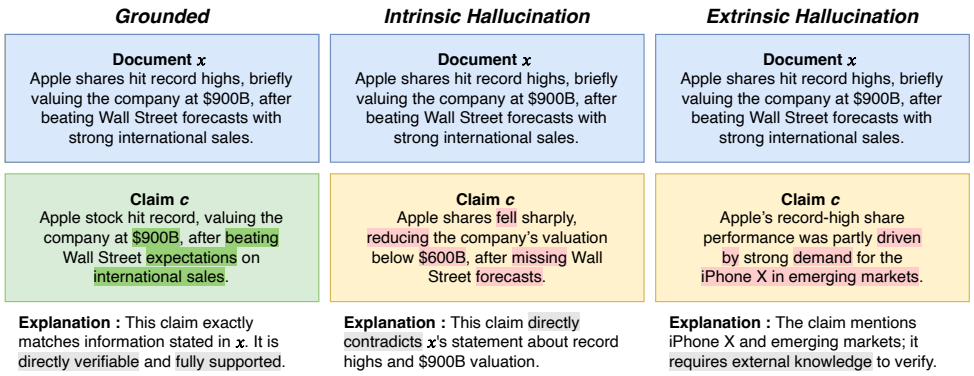


Figure 2: **Examples of Relations.** A grounded claim, an intrinsic hallucination, and an extrinsic hallucination.

Unsupported claims are treated as hallucinations and divided as $\texttt{Intrinsic}_{\texttt{Hallu}}$ if a claim $c$ directly contradicts $x$, or $\texttt{Extrinsic}_{\texttt{Hallu}}$ if it introduces information that cannot be verified from $x$ and requires external knowledge (see Figure 2).

## 4 Method

### 4.1 HalluGuard Overview

HalluGuard is a Small Reasoning Model (SRM) for closed-book, document-grounded hallucination detection, and it can be integrated as a guardrail in Retrieval-Augmented Generation (RAG) pipelines. Given a document-claim pair, HalluGuard only reasons over the provided document to predict whether the claim is `grounded` or `hallucinated`, and produces a justification by citing evidence from the document. This design improves transparency and user trust. HalluGuard supports two inference modes: in the `think` mode, it generates intermediate reasoning traces before the final output, while in `non-think` mode, it skips these traces and outputs directly. The mode is controlled at inference time by adding `/think` or `/no_think` to the prompt.

Our method begins with a domain-agnostic corpus that has been curated for safety, quality, and diversity (see Figure 3). The texts in this corpus are then linguistically reformed in tone and style by the Data Reformer (DR; Qwen3-235B-A22B (Yang et al., 2025)) to improve cross-domain generalization. From these reformed texts, we generate grounded and hallucinated synthetic claims using the Claim Generator (CG; Qwen3-235B-A22B).

To align the model towards high-quality reasoning and justifications, we construct a synthetic preference dataset. For each document-claim pair, we generate two candidate completions: one from the Preference Generator-Large (PG-L; Qwen3-235B-A22B) and one from the Preference Generator-Small (PG-S; Qwen3-0.6B). We designate the output of PG-L as the `chosen` completion and the output of PG-S as the `rejected` one. This creates preference pairs that exploit the empirical quality gap between large and small models, allowing us to build a training dataset without the need for additional human annotation. To further improve reliability, we apply two filtering steps: (i) model-agreement verification, in which the label deduced from the synthetic claim (from CG) is compared with the classification produced by PG-L; and (ii) LLM-based consensus filtering. In this step, two Independent Evaluators (IE-1; gpt-oss-120B (OpenAI, 2025) and IE-2; DeepSeek-V3.1-Terminus (DeepSeek-AI, 2024)) judge both completions. Only pairs in which both evaluators select the `chosen` completion are retained. Then, to avoid the Small Model Learnability Gap (Li et al., 2025), we fine-tuned a Qwen3-4B backbone using LoRA (Hu et al., 2022) and ORPO, which merges Supervised Fine-Tuning (SFT) and preference alignment in a single stage (see Appendix A).

Thus, HalluGuard is an SRM for document-grounded hallucination detection in RAG pipelines.

### 4.2 Structured Claim Dataset Construction

**Domain-Agnostic Corpus.** The performance of LLMs depends on both the size and the quality of the dataset (Gunasekar et al., 2023). Larger and more diverse datasets improve generalization by exposing models to varied contexts. We use FineWeb, an open, domain-agnostic web corpus.

From the 10BT sample[3], we retain only documents with high confidence in being in English (`language_score` $\geq 0.95$) and remove exact duplicates. Then, we randomly sample 300,000 documents to create the baseline dataset, $D_{\text{agnostic}}$.

[3] https://hf.co/datasets/HuggingFaceFW/fineweb

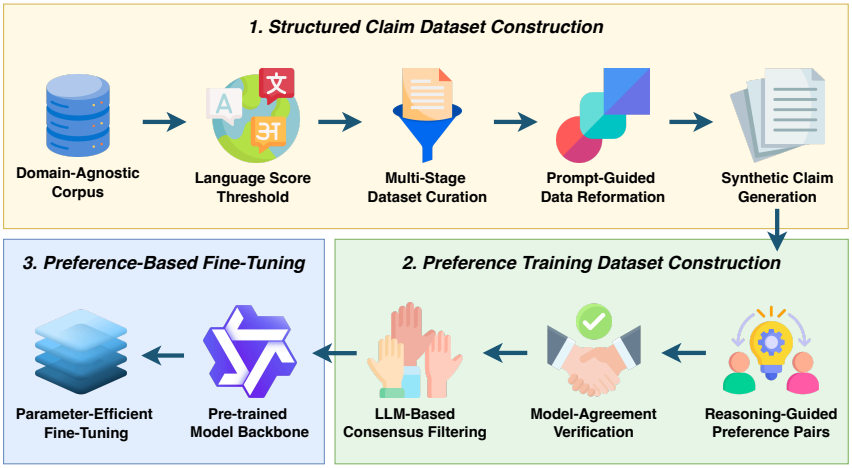


Figure 3: **HalluGuard Training Pipeline.** A domain-agnostic corpus is filtered, reformed, and used to generate three types of synthetic claims (grounded, intrinsic hallucinated, and extrinsic hallucinated). Preference data are built via cross-model generation (Qwen3-235B-A22B and Qwen3-0.6B), model-agreement verification and LLM-based consensus filtering are used to enhance quality and confidence. The Qwen3-4B backbone is then fine-tuned using LoRA and ORPO to mitigate hallucinations and produce evidence-grounded justifications in RAG pipelines.

**Multi-Stage Dataset Curation.** We further filter $D_{\text{agnostic}}$ to ensure safety, quality, and diversity, following C4-style practices (Raffel et al., 2020). We remove documents containing unsafe terms[4] or failing quality heuristics (e.g., fewer than five sentences, missing terminal punctuation, boilerplate such as `Lorem Ipsum` or cookie notices, or malformed text such as single tokens exceeding 1000 characters). We also discard documents shorter than 50 words and remove near-duplicates based on overlapping three-sentence spans to promote diversity by eliminating redundant content, ensuring a wider range of topics. The resulting dataset, $D_{\text{clean}}$, contains 104,966 documents.

**Prompt-Guided Data Reformation.** Despite multi-stage curation, $D_{\text{clean}}$ remains web-centric in style due to the nature of FineWeb. To increase linguistic diversity and improve generalization to non-web formats (e.g., reports, dialogues), we use DR to rewrite each document, producing a wider range of styles that better reflect real-world variation (Veselovsky et al., 2023; Long et al., 2024).

The reformed dataset is then:

$$D_{\text{reformed}} = \left\{ s_{j(x)}\big(x; T(x)\big) \;\middle|\; x \in D_{\text{clean}} \right\} \quad (1)$$

where $j(x)$ is a random style from $\mathcal{S} = \{s_1, s_2, \ldots, s_{18}\}$ defined in Appendix B, and $T(x)$ the temperature sampled uniformly from $[0.2, 0.7]$.

**Synthetic Claim Generation.** We generate one synthetic claim per document in $D_{\text{reformed}}$. To balance the classification task, we generate half grounded and half hallucinated claims (split evenly between intrinsic and extrinsic).

For each document $x_i \in D_{\text{reformed}}$, we ask CG to generate a claim $c_i$ in structured JSON format (He et al., 2024) (see Appendix C). Each claim is assigned a label $t_i \in \{\texttt{Grounded}, \texttt{Intrinsic}_{\text{Hallu}}, \texttt{Extrinsic}_{\text{Hallu}}\}$ and results in 104,966 document-claim-label triplets:

$$HalluClaim_{\text{base}} = \bigcup_{t \in \mathcal{C}} \{(x_i, c_i, t) \mid x_i \in D_t\} \quad (2)$$

where $D_t$ denotes the subset of document-claim pairs corresponding to label $t$.

### 4.3 Preference Training Dataset Construction

**Reasoning-Guided Preference Pairs.** The balanced dataset $HalluClaim_{\text{base}}$ contains document-claim-label triplets. However, our goal is not only to classify claims correctly, but also to train models to produce evidence-grounded justifications.

We convert $HalluClaim_{\text{base}}$ into the preference dataset format[5], where each instance comprises a prompt and two completions: a `chosen` completion and a `rejected` one (see Appendix D). For each triplet of document-claim-label, we construct a prompt $P_i$ containing: (i) task instructions defining the `Grounded`, $\texttt{Intrinsic}_{\text{Hallu}}$ and $\texttt{Extrinsic}_{\text{Hallu}}$ labels, (ii) the document $x_i$ and (iii) the claim $c_i$ and requiring classification and evidence-grounded justification in a fixed XML format (see Appendix E).

[4] https://github.com/LDNOOBW

[5] https://hf.co/docs/trl/dataset_formats

Thus, we use PG-L and PG-S, with the same prompt $P_i$. Each model $m \in \{\text{PG-L}, \text{PG-S}\}$ produces the response as follows:

$$R_i^{(m)} = \left(y_i^{(m)},\ j_i^{(m)},\ r_i^{(m)}\right) \tag{3}$$

where $y_i^{(m)}$ is the predicted label, $j_i^{(m)}$ is the justification and $r_i^{(m)}$ is the intermediate model reasoning within the `<think>` tags.

Assuming that larger models perform better, we apply a size-based heuristic: marking $R_i^{(\text{PG-L})}$ as `chosen` and $R_i^{(\text{PG-S})}$ as `rejected`.

For each triplet $(x_i, c_i, t_i)$ in $HalluClaim_{\text{base}}$, we produce preference tuples of the form:

$$z_i = \left(P_i,\ R_i^{(\text{PG-L})}{}_{(\text{chosen})},\ R_i^{(\text{PG-S})}{}_{(\text{rejected})}\right) \tag{4}$$

**Model-Agreement Verification.** The size-based heuristic provides a useful starting point, but some `chosen` completions may still misclassify the claim. To correct this, we require agreement between the synthetic label assigned by CG during claim generation and the classification predicted by PG-L. Any tuple where the `chosen` label disagrees with the synthetic label is removed. After verification, the $HalluClaim_{\text{pref}}$ dataset contains 87,486 tuples.

**LLM-Based Consensus Filtering.** To further improve reliability, each tuple is independently evaluated by IE-1 and IE-2 using a dedicated prompt that asks for the selection of the best completion according to three criteria: (i) classification correctness, (ii) coherence of reasoning, and (iii) clarity of justification (see Appendix F). The models receive the full prompt $P_i$ and completions, without being told which one is the `chosen` completion.

A tuple is retained only if IE-1 and IE-2 select the same completion that matches the `chosen` one.

$$\text{IE-1}(P_i) = \text{IE-2}(P_i) = R_i^{(\text{chosen})} \tag{5}$$

The LLM-based consensus step removed 1,886 tuples. The removal rates were uniform between styles (from 1.57% for `dialogue` to 2.90% for `social_media_post`), indicating that there was no disproportionate impact or style-specific bias.

Thus, to address the imbalance between classes (`Grounded`, $\texttt{Intrinsic}_{\text{Hallu}}$, $\texttt{Extrinsic}_{\text{Hallu}}$), we implement a controlled sampling strategy. We first determined the maximum number of samples per hallucinated class that would allow equal representation, and then sampled twice as many `Grounded` examples to maintain balance. The final $HalluClaim$ dataset was shuffled, leading to 76,708 high-quality preference tuples.

### 4.4 Preference-Based Fine-Tuning

After creating a high-quality preference dataset, we use Qwen3-4B (Yang et al., 2025) as the backbone for fine-tuning. Its 32,768-token context window supports document-level reasoning, while the 4B-parameter scale remains compact enough for on-prem deployment. This choice also mitigates the Small Model Learnability Gap (Li et al., 2025) observed in models with at most 3B parameters.

**Parameter-Efficient Fine-Tuning.** We fine-tune Qwen3-4B using ORPO, a fine-tuning technique that increases the gap between `chosen` and `rejected` completions so that the model consistently favors the `chosen` one (see Appendix G). Unlike DPO, ORPO performs an SFT stage during preference alignment, without relying on a reference model. This makes training more efficient and allows HalluGuard to accurately classify claims while generating justifications and reasoning distilled from stronger models.

To apply ORPO in a parameter-efficient manner, we use LoRA (Hu et al., 2022), which freezes most base weights and trains only small adapter layers. This reduces memory and compute costs while mitigating catastrophic forgetting when adapting pre-trained models to specific tasks (Bafghi et al., 2025). Reproducibility and fine-tuning details are provided in the Appendices H and J.

## 5 Experimental Setup

**Benchmark Dataset.** We evaluated on LLM-AggreFact (Tang et al., 2024a), a collection of 11 human-annotated datasets designed to assess whether model-generated claims are supported by evidence documents in a binary `Grounded` vs. `Hallucinated` task across both closed-book and grounded settings. The benchmark spans diverse domains and incorporates real hallucinations from recent LLMs. Importantly, it includes RAGTruth (Niu et al., 2024), which is particularly relevant to our focus on hallucination mitigation in RAG (see Appendix K).

**Evaluation Metric.** Performance is measured using balanced accuracy (BAcc) (Brodersen et al., 2010), defined as $\text{BAcc} = \frac{1}{2}\left(\frac{\text{TP}}{\text{TP+FN}} + \frac{\text{TN}}{\text{TN+FP}}\right)$ where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives. We adopted BAcc to ensure comparability with prior work, as it was also used in the paper that introduced LLM-AggreFact.

| Model | Size | AGGREFACT | | TofuEval | | WiCE | REVEAL | Claim Verify | Fact Check | Expert QA | LFQA | RAG Truth | BAcc Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CNN | XSum | MediaS | MeetB | | | | | | | | |
| Bespoke-Minicheck-7B | 7B | 65.5 | 77.8 | 76.0 | 78.3 | 83.0 | 88.0 | 75.3 | 77.7 | 59.2 | 86.7 | 84.0 | 77.4 |
| Claude-3.5 Sonnet | - | 67.6 | 75.1 | 73.4 | 84.6 | 77.7 | 89.1 | 71.4 | 77.8 | 60.9 | 85.6 | 86.1 | 77.2 |
| HalluGuard-4B | 4B | 70.7 | 75.5 | 73.2 | 79.0 | 80.5 | 87.7 | 77.3 | 74.0 | 59.4 | 86.1 | 84.4 | 77.1 |
| Qwen3-235B-A22B* | 235B | 76.7 | 71.3 | 73.6 | 81.7 | 80.8 | 86.3 | 74.6 | 72.1 | 58.0 | 84.7 | 85.7 | 76.9 |
| Granite Guardian 3.3 | 8B | 67.0 | 74.9 | 74.0 | 78.6 | 76.6 | 89.6 | 75.9 | 76.1 | 59.6 | 86.9 | 82.2 | 76.5 |
| Mistral-Large 2 | 123B | 64.8 | 74.7 | 69.6 | 84.2 | 80.3 | 87.7 | 71.8 | 74.5 | 60.8 | 87.0 | 85.9 | 76.5 |
| gpt-4-turbo-preview | - | 66.7 | 76.5 | 71.4 | 79.9 | 80.4 | 87.8 | 67.6 | 79.9 | 59.2 | 83.1 | 85.3 | 76.2 |
| gpt-4o-2024-05-13 | - | 68.1 | 76.8 | 71.4 | 79.8 | 78.5 | 86.5 | 69.0 | 77.5 | 59.6 | 83.6 | 84.3 | 75.9 |
| Qwen3-4B | 4B | 60.0 | 74.7 | 72.0 | 78.0 | 77.8 | 91.4 | 73.3 | 76.5 | 60.4 | 85.0 | 84.0 | 75.7 |
| Qwen2.5-72B-Instruct | 72B | 63.6 | 73.0 | 71.9 | 80.4 | 80.2 | 88.9 | 70.0 | 77.0 | 60.1 | 84.3 | 81.9 | 75.6 |
| Llama-3.3-70B-Instruct | 70B | 68.7 | 74.7 | 69.5 | 78.4 | 76.6 | 85.5 | 67.4 | 78.5 | 58.3 | 79.8 | 82.6 | 74.5 |
| Llama-3.1-405B-Instruct | 405B | 64.8 | 75.1 | 68.6 | 81.2 | 71.8 | 86.4 | 67.5 | 79.4 | 58.5 | 81.9 | 82.9 | 74.4 |
| gpt-4o-mini-2024-07-18 | - | 61.8 | 73.6 | 71.3 | 79.7 | 76.3 | 85.8 | 69.8 | 76.0 | 58.3 | 80.3 | 81.6 | 74.0 |
| QwQ-32B-Preview | 32B | 57.0 | 71.6 | 69.3 | 78.5 | 72.3 | 86.2 | 67.7 | 75.6 | 60.0 | 78.9 | 72.4 | 71.8 |
| Mixtral-8x22B | 176B | 57.3 | 70.3 | 69.0 | 78.5 | 69.5 | 85.8 | 63.8 | 79.5 | 57.4 | 76.5 | 78.8 | 71.5 |
| Llama-3.1-8B-Instruct | 8B | 54.7 | 68.5 | 71.1 | 75.5 | 72.0 | 83.5 | 66.5 | 72.3 | 57.8 | 77.5 | 73.6 | 70.3 |
| GPT-3.5-Turbo | - | 63.2 | 72.4 | 66.8 | 73.4 | 68.5 | 84.7 | 65.2 | 70.8 | 57.2 | 73.8 | 75.6 | 70.1 |
| Qwen3-0.6B* | 0.6B | 51.5 | 64.7 | 63.6 | 71.0 | 65.4 | 85.8 | 64.8 | 75.0 | 58.3 | 73.8 | 65.5 | 67.2 |
| Llama-3.2-3B-Instruct | 3B | 51.5 | 60.5 | 53.1 | 52.5 | 58.5 | 81.9 | 62.3 | 62.6 | 55.8 | 58.5 | 63.0 | 60.0 |

Table 1: **Evaluation on LLM-AggreFact.** Models are ordered by average balanced accuracy (BAcc Avg.; higher is better). HalluGuard-4B (ours), Qwen3-0.6B, 4B and 235B-A22B were evaluated using our prompt in `think` mode. All other results are taken from the public leaderboard. The higher score between HalluGuard-4B and Qwen3-4B is shaded in dark green. Alternating grey rows improve readability. * Models used within our training pipeline.

## 6 Results

**Evaluation on Benchmark.** We evaluate classification performance by extracting predicted labels regardless of XML validity and align HalluGuard's three-class output with the binary LLM-AggreFact setting by treating both intrinsic and extrinsic hallucination types as a single `hallucinated` class.

As shown in Table 1, HalluGuard achieves 77.1% BAcc, outperforming its Qwen3-4B backbone by +1.4 points (75.7) with notable gains in AggreFact-CNN (+10.7) and ClaimVerify (+4.0). All results use the `think` mode with fixed inference parameters (see Appendix I). HalluGuard is competitive with much larger general-purpose LLMs such as Mistral Large 2 (76.5) and GPT-4o (75.9), and outperforms Granite Guardian 3.3 (76.5) but remains slightly below MiniCheck (77.4) while using only 57% of its parameters (4B vs. 7B).

Although MiniCheck achieves a marginally higher average BAcc, a paired permutation test on all benchmark datasets indicates that HalluGuard achieves a statistically significant advantage in expected BAcc ($\Delta$BAcc = +0.81, $p < 0.001$).

Overall, results show that our method enables a compact backbone to rival general-purpose LLMs and specialized models in hallucination detection.

**RAGTruth Detailed Evaluation.** This subset focuses on RAG settings, evaluating whether claims are supported by retrieved documents.

HalluGuard achieves an average BAcc of 84.4%, improving over its Qwen3-4B backbone (84.0) and surpassing specialized models such as MiniCheck (84.0) and Granite Guardian 3.3 (82.2), while using roughly half of their parameters. Although the gain in BAcc is modest, it does not fully reflect HalluGuard's recall on the `hallucinated` class.

HalluGuard achieves 84.1% recall on hallucinated claims, missing 203 cases, compared to 334 for Granite Guardian 3.3 (73.7), 287 for Qwen3-4B (77.4), and 229 for MiniCheck (81.9) (see Table 2).

These results support HalluGuard's use as an effective guardrail mitigating hallucinations in RAG.

| Predicted | Actual | |
|---|---|---|
| | Hallucinated | Grounded |
| `Grounded` | 203 | 12,807 |
| $\texttt{Intrinsic}_{\text{Hallu}}$ | 466 | 403 |
| $\texttt{Extrinsic}_{\text{Hallu}}$ | 600 | 1,882 |
| **Total** | 1,269 | 15,092 |

Table 2: **Confusion Matrix.** Rows show HalluGuard's predictions; columns show binary ground-truth labels.

**Fine-Grained Hallucination Analysis.** Because RAGTruth provides only binary labels, Table 2 should be read as a behavioral analysis of HalluGuard's predictions rather than a direct evaluation of subtype correctness. Among ground-truth `hallucinated` instances, HalluGuard predicts `Grounded`, $\texttt{Intrinsic}_{\text{Hallu}}$, and $\texttt{Extrinsic}_{\text{Hallu}}$ in 16.0%, 36.7%, and 47.3% of cases. Among ground-truth `grounded` instances, it predicts `Grounded` in 84.9% of cases; its false positives are more often $\texttt{Extrinsic}_{\text{Hallu}}$ than $\texttt{Intrinsic}_{\text{Hallu}}$ (12.5% vs. 2.7%), suggesting greater sensitivity to unsupported content than to explicit contradictions.

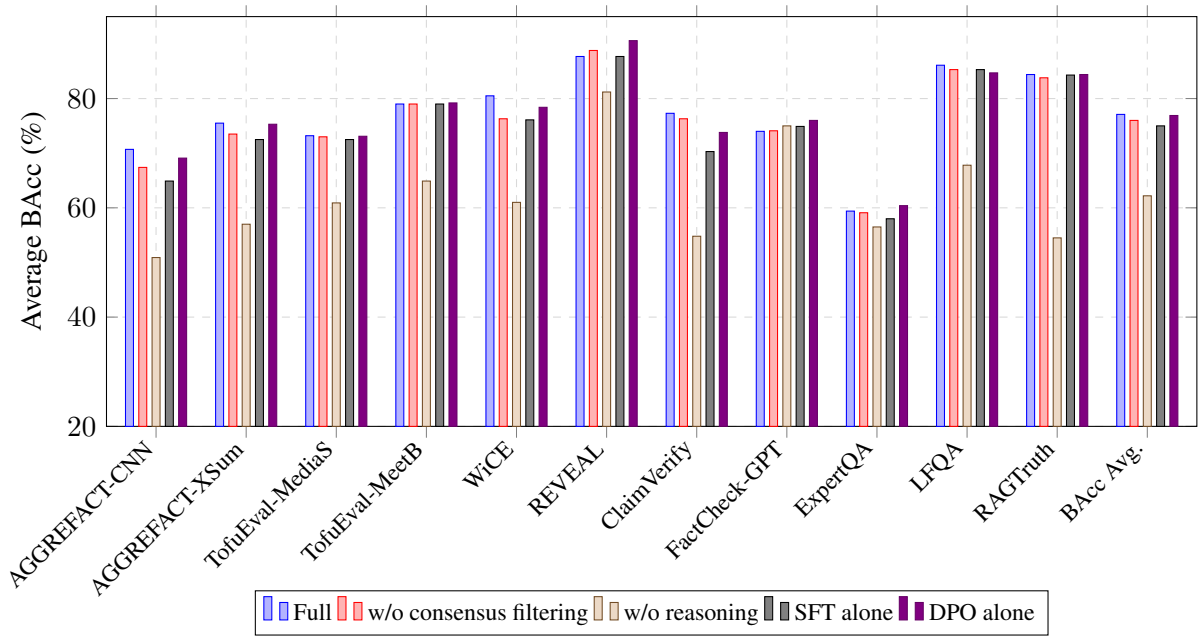


Figure 4: **Ablation of HalluGuard.** Comparison of the full model and four variants on LLM-AggreFact.

**Computational Efficiency.** To compare latency and hardware requirements, we measure inference throughput and GPU memory usage with a vLLM backend, a batch size of 1 on an NVIDIA A100 GPU using our prompt (see Appendix E) and each model's default context window.

As shown in Table 3, HalluGuard achieves an average throughput of 120.93 tokens per second (TPS) over 100 samples, representing improvements of 49% and 68% over MiniCheck and Granite Guardian 3.3, respectively. HalluGuard requires at most 8.1 GB of VRAM, compared to 32.2 GB and 38.6 GB for the baselines, corresponding to a 75-79% reduction in memory footprint.

| Model | Throughput (TPS) | Peak VRAM (MB) |
|---|---|---|
| HalluGuard-4B | 120.93 | 8,098 |
| MiniCheck-7B | 81.15 | 32,211 |
| Granite Guardian 3.3 | 72.07 | 38,585 |

Table 3: **Efficiency Comparison.** Inference throughput (tokens/s) and peak GPU memory usage across models.

On average, HalluGuard takes 10.4 seconds to process each document-claim sample. In a RAG pipeline, this introduces additional post-generation latency. However, this overhead provides a critical performance gain, as shown in Section 7.

Together with its competitive performance, these results indicate a favorable efficiency-performance trade-off for resource-constrained environments.

**Justification Quality.** Reference-based metrics such as ROUGE (Lin, 2004) are problematic for our task, as they do not assess factual grounding and are poorly correlated with human judgments (Wang et al., 2023). We therefore use G-Eval (Liu et al., 2023b) with DeepSeek-V3.1-Terminus as evaluator. For each document, we scored the justifications of PG-L, HalluGuard, and PG-S on Relevance, Consistency, Coherence (1-5) and Fluency (1-3).

Table 4 presents several disparities. Qwen3-235B-A22B outperforms Qwen3-0.6B in all dimensions, while HalluGuard, although nearly two orders of magnitude smaller, achieves comparable scores. This suggests that ORPO transfers strong justification behavior to a compact backbone. Fluency remains uniformly high, indicating that gains arise primarily from improved factual grounding and reasoning rather than surface fluency. Finally, these results show that HalluGuard can match the quality of justification of a 235B-parameter model.

| Model | Rel | Coh | Con | Flu |
|---|---|---|---|---|
| Qwen3-235B-A22B | 4.89 | 4.63 | 4.52 | 3.00 |
| HalluGuard-4B | 4.86 | 4.40 | 4.56 | 2.99 |
| Qwen3-0.6B | 4.52 | 3.78 | 4.42 | 2.96 |

Table 4: **G-Eval Results.** Evaluation of justification quality using four dimensions: Relevance (Rel), Coherence (Coh), Consistency (Con), and Fluency (Flu).

**Justification-Accuracy Alignment.** We evaluate whether justification quality predicts classification reliability by computing Spearman correlations

per-item ($\rho$) between G-Eval scores and classification correctness. HalluGuard exhibits a strong correlation between Consistency and correctness ($\rho = 0.527$, $p < 0.0001$), while Qwen3-4B shows no such relationship ($\rho = 0.011$, $p = 0.6660$). This result suggests that fine-tuning better aligns model reasoning with correct predictions, leading to stronger justification-accuracy alignment. A weaker but significant correlation is also observed for Relevance ($\rho = 0.215$, $p < 0.0001$), further linking HalluGuard's improved accuracy to more consistent and relevant justifications.

**Human Alignment Evaluation.** We evaluated whether our heuristic preference construction (Section 4.3) aligns with human judgments. We sampled 100 preference tuples $z_i$ (balanced between grounded and hallucinated claims). Two independent NLP expert annotators compared the two completions in each tuple using the same criteria as those used in dataset construction (classification correctness, reasoning coherence, and justification clarity). Annotators were blind to labels (`chosen` or `rejected`) and saw completions in random order.

At the item level, `chosen` was preferred in 71 of 75 tuples with full agreement (94.7%; $p = 3.4 \times 10^{-17}$, binomial test). At the annotation level, 83.5% of all 200 judgments favored `chosen` ($p = 4.7 \times 10^{-23}$; see Table 5).

Overall, these results indicate that our heuristic provides a reliable proxy for human preferences.

| Evaluation level | Pref. for chosen |
|---|---|
| Item level ($n = 75$) | 94.7% |
| Annotation level ($n = 200$) | 83.5% |

Table 5: **Human Alignment Results.** Annotators preferred the `chosen` completions (94.7% of the 75 fully agreed items; 83.5% of the 200 individual judgments).

## 7 Ablation Study

**Impact of Consensus Filtering.** Applying LLM-based consensus filtering using independent evaluators (IE-1 and IE-2) to preference tuples provides a decisive improvement. With filtering, HalluGuard reaches 77.1% BAcc, compared to 76.0% without it (-1.1%). Although the absolute gain appears modest, its impact is substantial: without consensus filtering, HalluGuard no longer remains competitive, falling behind not only Mistral Large 2 (76.5%) but also Granite Guardian 3.3 (76.5%). This highlights consensus filtering as essential for preserving HalluGuard's competitive performance.

**Contribution of Reasoning.** Disabling reasoning by using `/no_think` in the prompt leads to a decrease in performance. In `think` mode, HalluGuard reaches a BAcc of 77.1%, whereas in `/no_think` mode the BAcc decreases to 62.2% (-14.9%). This represents the largest drop in our ablation study, highlighting the critical role of reasoning in detecting hallucinations.

This is even more marked on RAGTruth, where reasoning improves BAcc (+29.9%), with consistent gains across all other datasets (see Figure 5).

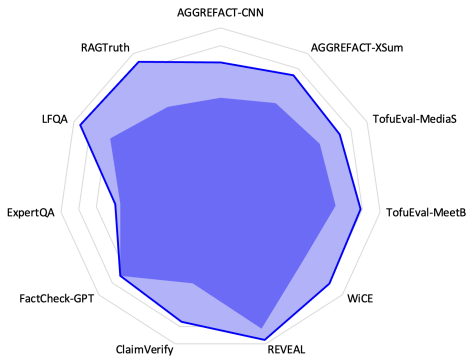


Figure 5: **Effect of Model Reasoning.** Radar plot comparing HalluGuard in `think` mode (lighter blue) vs. in `/no_think` mode (darker blue).

**ORPO Training Strategy.** Fine-tuning with SFT alone results in a notable performance drop, with BAcc decreasing from 77.1% to 75.0% (-2.1%). However, fine-tuning with DPO alone, without prior SFT, achieves a BAcc of 76.9%, outperforming SFT alone but still below ORPO. In general, ORPO achieves the highest balanced accuracy (77.1%), confirming the effectiveness of jointly integrating SFT and preference alignment within a single fine-tuning stage.

**Ablation Results.** Figure 4 and Table 6 present the results of the ablation study for HalluGuard. The full model consistently achieves the highest BAcc against ablated variants, demonstrating that observed performance improvements are the result of the association of all components.

| Model Variant | BAcc | Δ |
|---|---|---|
| HalluGuard-4B (full) | 77.1 | - |
| w/o consensus filtering | 76.0 | -1.1 |
| w/o reasoning (`/no_think`) | 62.2 | -14.9 |
| SFT alone | 75.0 | -2.1 |
| DPO alone | 76.9 | -0.2 |

Table 6: **Ablation Results.** BAcc for the full model and ablated variants. Δ is relative to the full model.

## 8 Conclusion

We presented HalluGuard, a 4B-parameter small reasoning model designed as a guardrail for Retrieval-Augmented Generation, detecting closed-book, document-grounded hallucinations while producing evidence-grounded justifications.

Built on a domain-agnostic synthetic dataset with multi-stage curation and preference-based fine-tuning using ORPO and LoRA, HalluGuard transforms a compact backbone into a model that rivals or surpasses much larger LLMs and recent specialized hallucination detection models. HalluGuard achieves competitive performance on LLM-AggreFact while producing justifications of comparable quality to those of a 235B-parameter model.

Ablation studies demonstrate that carefully aligned small reasoning models can offer a reliable alternative solution for enterprise RAG pipelines, narrowing the gap with frontier LLMs while fostering end-user trust through transparent and explainable hallucination detection.

## 9 Future Work

While HalluGuard achieves strong performance, several directions remain for future work. First, we plan to extend its reasoning capabilities to multimodal settings, allowing it to verify claims against charts and tables. This extension would improve its applicability to a broader range of documents.

Second, we aim to optimize the length of internal reasoning traces to reduce post-generation latency while maintaining a favorable balance among performance, scalability, and deployment constraints.

Finally, we plan to release additional variants fine-tuned from newer pretrained backbones.

## Limitations

**Synthetic Data.** Although multiple filters are applied, the synthetic claims used for training may not fully capture the complexity of real-world hallucinations, potentially limiting generalization when faced with data beyond the synthetic distribution.

**Hallucination Taxonomy.** Multiple fine-grained hallucination classifications can be explored, like factual misalignment, invented world knowledge, grammatical errors, logical failures, etc. In this study, we used a three-label taxonomy (grounded, intrinsic, and extrinsic). More research needs to be conducted on other types of hallucinations according to real world data applications.

**Language and Domain Generalization.** HalluGuard is trained and evaluated on English data. Its performance in other languages or domain-specific settings (e.g., legal or finance) remains unverified.

## Ethical Considerations

As with any hallucination detection model, HalluGuard must be used with caution. Overflagging grounded claims may reduce user trust, while failing to detect hallucinations can lead to harmful errors further down the line. For this reason, HalluGuard should be used as a decision-support tool rather than as a fully autonomous system and should always be paired with human oversight. We therefore encourage responsible deployment in sensitive domains when integrating HalluGuard into real-world RAG pipelines.

## References


Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and verification over unstructured and structured information (feverous) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER) at EMNLP*, volume 2021, pages 1–13.

Reza Akbarian Bafghi, Carden Bagwell, Avinash Ravichandran, Ashish Shrivastava, and Maziar Raissi. 2025. Fine tuning without catastrophic forgetting via selective low rank adaptation. *arXiv preprint arXiv:2501.15377*.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. 2010. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE.

Hung-Ting Chen, Fangyuan Xu, Shane Arora, and Eunsol Choi. 2023. Understanding retrieval augmentation for long-form question answering. *Preprint*, arXiv:2310.12150.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep

bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Suriya Gunasekar, Yi Zhang, and Jyoti Aneja. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189, Miami, Florida, USA. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Alon Jacovi, Yonatan Bitton, Bernd Bohnet, Jonathan Herzig, Or Honovich, Michael Tseng, Michael Collins, Roee Aharoni, and Mor Geva. 2024. A chain-of-thought is as strong as its weakest link: A benchmark for verifiers of reasoning chains. *Preprint*, arXiv:2402.00559.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, and Sophie Rosset. 2024. Small language models are good too: An empirical study of zero-shot classification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14923–14936.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yuetai Li, Xiang Yue, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Bill Yuchen Lin, Bhaskar Ramasubramanian, and Radha Poovendran. 2025. Small models struggle to learn from strong reasoners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25366–25394, Vienna, Austria. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Yang Liu, Dan Iter, and Yichong Xu. 2023b. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint*. ArXiv:2303.16634 [cs].

Lin Long, Rui Wang, and Ruixuan Xiao. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. *arXiv preprint*. ArXiv:2406.15126 [cs].

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822.

Bo Ni, Zheyuan Liu, Leyao Wang, Yongjia Lei, Yuying Zhao, Xueqi Cheng, Qingkai Zeng, Luna Dong, Yinglong Xia, Krishnaram Kenthapadi, and 1 others. 2025. Towards trustworthy retrieval augmented generation for large language models: A survey. *arXiv preprint arXiv:2502.06872*.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented

language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2025. gpt-oss-120b gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Inkit Padhi, Manish Nagireddy, Giandomenico Cornacchia, Subhajit Chaudhury, Tejaswini Pedapati, Pierre Dognin, Keerthiram Murugesan, Erik Miehling, Martín Santillán Cooper, Kieran Fraser, and al. 2024. Granite guardian. *Preprint*, arXiv:2412.07724.

Shrey Pandit, Ashwin Vinod, Liu Leqi, and Ying Ding. 2025. Teaching with lies: Curriculum dpo on synthetic negatives for hallucination detection. *arXiv preprint arXiv:2505.17558*.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. Averitec: A dataset for real-world claim verification with evidence from the web. *Advances in Neural Information Processing Systems*, 36:65128–65167.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. Rag-hat: A hallucination-aware tuning pipeline for llm in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558.

Yiming Tan, Dehai Min, Yu Li, Wenbo Li, Nan Hu, Yongrui Chen, and Guilin Qi. 2023. Can chatgpt replace traditional kbqa models? an in-depth analysis of the question answering performance of the gpt llm family. In *International Semantic Web Conference*, pages 348–367. Springer.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8818–8847, Miami, Florida, USA. Association for Computational Linguistics.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024b. TofuEval: Evaluating hallucinations of LLMs on topic-focused dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Akhil Arora. 2023. Generating Faithful Synthetic Data with Large Language Models: A Case Study in Computational Social Science. *arXiv preprint*. ArXiv:2305.15041 [cs].

Chengyu Wang, Taolin Zhang, Richang Hong, and Jun Huang. 2025. A short survey on small reasoning models: Training, inference, applications and research directions. *arXiv preprint arXiv:2504.09100*.

Jiaan Wang, Yunlong Liang, and Fandong Meng. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. *arXiv preprint*. ArXiv:2303.04048 [cs].

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. *Preprint*, arXiv:2311.09000.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025a. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Ziyao Zhang, Chong Wang, Yanlin Wang, Ensheng Shi, Yuchi Ma, Wanjun Zhong, Jiachi Chen, Mingzhi Mao, and Zibin Zheng. 2025b. Llm hallucinations in practical code generation: Phenomena, mechanism, and mitigation. *Proceedings of the ACM on Software Engineering*, 2(ISSTA):481–503.

# A ORPO Objective Function

The objective function consists of the following two components:

1. Supervised Fine-Tuning loss:

$$L_{\text{SFT}} = -\frac{1}{m}\sum_{k=1}^{m}\sum_{i=1}^{|V|} y_i^{(k)} \log\left(p_i^{(k)}\right)$$

2. Odds Ratio loss:

$$L_{\text{OR}} = -\log\sigma\left(\log\frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)}\right)$$

where $\sigma$ is the sigmoid function.

Therefore, $L_{\text{ORPO}} = L_{\text{SFT}} + \lambda L_{\text{OR}}$, where $\lambda$ controls the pretrained language model to adapt to the specific subset of desired output and discourages the generation of the `rejected` answers. Log odds ratio loss is wrapped in the sigmoid function so that it can be minimized by increasing the log odds between $y_w$ and $y_l$.

# B Prompt: Style Reformation

| Style | Instruction |
|---|---|
| `paraphrase` | Paraphrase the following text while retaining its original meaning. |
| `summarize` | Provide a concise summary of the following text. |
| `expand` | Expand on the following text by adding more details and context. |
| `news_article` | Rewrite the following information as a news article. |
| `blog_post` | Transform the following text into an engaging blog post. |
| `report` | Convert the following information into a formal report. |
| `story` | Rewrite the following text as a narrative story. |
| `dialogue` | Transform the following text into a dialogue between two characters. |
| `letter` | Rewrite the following text as a formal letter. |
| `social_media_post` | Transform the following text into a social media post. |
| `script` | Transform the following text into a script for a short video or play. |
| `interview` | Rewrite the following text as an interview between an interviewer and an expert. |
| `product_description` | Transform the following text into a product description. |
| `review` | Rewrite the following text as a review of a product or service. |
| `news_summary` | Summarize the following article into a concise news brief. |
| `formalize_news` | Rewrite the following content in a formal journalistic style. |
| `meeting_summary` | Rewrite the following text as if it were a summary of a meeting. |
| `meeting_dialogue` | Rewrite the following content as a conversation between multiple meeting participants. |

Table 7: Each style is randomly applied to reform FineWeb raw data and increase stylistic diversity.

# C Prompt: Claim Generation

**`Grounded`**

```
{
    "instructions": [
        "Generate a claim that is fully
    supported by the document. All relevant
    parts are directly verifiable from the
    document.",
        "Only return the claim as the answer.
    Do not include any additional text,
    explanation, or formatting."
    ],
    "document": <document>,
    "answer": ""
}
```

**Intrinsic$_{\text{Hallu}}$**

```
{
    "instructions": [
        "Generate a claim that contradicts what
    the document states or clearly implies.",
        "Only return the claim as the answer.
    Do not include any additional text,
    explanation, or formatting."
    ],
    "document": <document>,
    "answer": ""
}
```

**Extrinsic$_{\text{Hallu}}$**

```
{
    "instructions": [
        "Generate a claim that includes
    information that is not stated or implied
    in the document and cannot be verified
    using only the document (it requires
    external knowledge).",
        "Only return the claim as the answer.
    Do not include any additional text,
    explanation, or formatting."
    ],
    "document": <document>,
    "answer": ""
}
```

# D Preference Tuple: Full Example

```
{
  "prompt": "{
    "instructions": [
        "You will be given a document and a
    claim.",
        "Decide whether the claim is
    'GROUNDED', 'HALLUCINATED_INTRINSIC', or
    'HALLUCINATED_EXTRINSIC' based ONLY on the
    document.",
        "Definitions:",
        "  - GROUNDED: The claim is fully
    supported by the document. All relevant
    parts are directly verifiable from the
    document.",
        "  - HALLUCINATED_INTRINSIC: The claim
    contradicts what the document states or
    clearly implies.",
        "  - HALLUCINATED_EXTRINSIC: The claim
    includes information that is not stated or
    implied in the document and cannot be
    verified using only the document (it
    requires external knowledge).",
        "Justification requirements:",
        "  - Your justification MUST be
    evidence-grounded.",
        "  - Explicitly refer to the relevant
    parts of the document (by quoting or
    paraphrasing them).",
        "  - Explain how these parts SUPPORT,
    CONTRADICT, or FAIL TO SUPPORT the claim.",
        "  - Do NOT use any external knowledge;
    rely only on the provided document.",
        "Answer format (VERY IMPORTANT):",
        "  - You MUST respond using EXACTLY the
    following XML structure:",
        "    <answer>",
        "
    <classification>CATEGORY</classification>",
        "      <justification>Your reasoning
    here</justification>",
        "    </answer>",
        "  - CATEGORY must be ONE of: GROUNDED,
    HALLUCINATED_INTRINSIC,
    HALLUCINATED_EXTRINSIC.",
        "  - The <justification> must briefly
    explain your reasoning and cite evidence
    from the document.",
        "  - Do NOT add any other text before
    or after the <answer>...</answer> block.",
        "  - Do NOT add any extra tags or
    attributes.",
    ],
    "document": "'As I stepped into the
    charming St. Blasius parish... (truncated)
    ...'",
    "claim": "'St. Blasius Church is over 850
    years old and was built during...
    (truncated) ...'",
  }",
  "chosen": [
    {
      "role": "assistant",
      "content": "<think>Okay, let's tackle
    this. The user wants to determine if the
    claim about St.    Blasius Church...
    (truncated) ... Therefore, the claim is
    GROUNDED</think>
      <answer>
      <classification>GROUNDED</classification>
      <justification>The claim that 'St.
    Blasius Church is over 850 years old and
    was built
      during the reign of King Stephen' is
    directly supported by the document...
    (truncated) ...</justification>
      </answer>
    }
  ],
  "rejected": [
    {
      "role": "assistant",
      "content": "<think>Okay, let's tackle
    this. The user wants me to determine if the
    claim... (truncated) ... Since the document
    contradicts the claim, the claim is
    intrinsically hallucinated.</think>
    <answer>
    <classification>HALLUCINATED_INTRINSIC
    </classification>
    <justification>The document explicitly
    states: 'The church itself was over 850
    years old, built during the reign of King
    Stephen ... (truncated) ...</justification>
    </answer>
    }
  ]
}
```

# E Prompt: Synthetic Pair Generation

```
{
    "instructions": [
        "You will be given a document and a
    claim.",
```

```
    "Decide whether the claim is
    'GROUNDED', 'HALLUCINATED_INTRINSIC', or
    'HALLUCINATED_EXTRINSIC' based ONLY on the
    document.",
        "Definitions:",
        "  - GROUNDED: The claim is fully
    supported by the document. All relevant
    parts are directly verifiable from the
    document.",
        "  - HALLUCINATED_INTRINSIC: The claim
    contradicts what the document states or
    clearly implies.",
        "  - HALLUCINATED_EXTRINSIC: The claim
    includes information that is not stated or
    implied in the document and cannot be
    verified using only the document (it
    requires external knowledge).",
        "Justification requirements:",
        "  - Your justification MUST be
    evidence-grounded.",
        "  - Explicitly refer to the relevant
    parts of the document (by quoting or
    paraphrasing them).",
        "  - Explain how these parts SUPPORT,
    CONTRADICT, or FAIL TO SUPPORT the claim.",
        "  - Do NOT use any external knowledge;
    rely only on the provided document.",
        "Answer format (VERY IMPORTANT):",
        "  - You MUST respond using EXACTLY the
    following XML structure:",
        "    <answer>",
        "
    <classification>CATEGORY</classification>",
        "      <justification>Your reasoning
    here</justification>",
        "    </answer>",
        "  - CATEGORY must be ONE of: GROUNDED,
    HALLUCINATED_INTRINSIC,
    HALLUCINATED_EXTRINSIC.",
        "  - The <justification> must briefly
    explain your reasoning and cite evidence
    from the document.",
        "  - Do NOT add any other text before
    or after the <answer>...</answer> block.",
        "  - Do NOT add any extra tags or
    attributes.",
    ],
    "document": <document>,
    "claim": <claim>,
}
```

## F Prompt: Consensus Filter

```
{
  "instructions": [
    "You will be given a document and a claim,
    along with two responses (RESPONSE_A and
    RESPONSE_B).",
    "Evaluate which response is better based
    solely on classification correctness,
    coherence, clarity, and justification
    quality.",
    "Your output MUST be exactly one of the
    following two strings: 'RESPONSE_A' or
    'RESPONSE_B'.",
    "Do NOT provide explanations, reasoning,
    punctuation, extra text, or any other
    content.",
  ],
  "document": <document>,
  "claim": <claim>,
  "RESPONSE_A": <response_a>,
  "RESPONSE_B": <response_b>,
  "best_response": ""
}
```

## G Reward Margin Evolution


Figure 6: The gap between `chosen` and `rejected` completions increases over training, showing that the model progressively learns to prefer `chosen` examples while assigning lower rewards to `rejected` ones.

## H Technical Reproducibility Details

For reproducibility, we report the experimental setup. HalluGuard was fine-tuned for 16 hours on a single NVIDIA H100 PCIe GPU (80GB memory, TDP 350W). Training consumed approximately 7.35 kWh, estimated using the Machine Learning Impact Calculator (MLIC) (Lacoste et al., 2019). Experiments were conducted on a Linux server with CUDA 12.4.1 and PyTorch 2.4.0, using default random seeds and settings.

## I Inference Parameters

| Parameter | Non-Thinking | Thinking |
|---|---|---|
| temperature | 0.7 | 0.6 |
| min_p | 0.0 | 0.0 |
| top_p | 0.8 | 0.95 |
| top_k | 20 | 20 |

Table 8: Inference parameters used in our experiments, following the recommended Qwen settings for non-thinking and thinking modes.

## J Fine-Tuning Configuration

| Parameter | Value |
|---|---|
| lora_layers_attn | `q_proj`<br>`k_proj`<br>`v_proj` |
| lora_layers_ffn | `gate_proj`<br>`up_proj`<br>`down_proj` |
| lora_rank | 16 |
| lora_alpha | 16 |
| lora_dropout | 0 |
| precision | `bfloat16` |
| epochs | 1 |
| batch_size | 2 |
| grad_accumulation | 4 |
| effective_batch_size | 8 |
| optimizer | AdamW (8-bit) |
| learning_rate | $1 \times 10^{-6}$ |
| lr_schedule | linear |
| orpo_beta | 0.1 |
| max_length | 32768 |
| max_prompt_length | 32768 |
| max_completion_length | 32768 |

Table 9: The setup trains ∼33M parameters (0.81% of the full model) using LoRA for 1 epoch.

## K Benchmark Dataset Details

LLM-AggreFact includes the following datasets: AGGREFACT (Tang et al., 2023), a factual consistency benchmark for summarization; TofuEval (Tang et al., 2024b), a dialogue summarization benchmark with LLM summaries annotated for factual consistency; WiCE (Kamoi et al., 2023), a textual entailment dataset of Wikipedia claims and cited sources; REVEAL (Jacovi et al., 2024), which evaluates reasoning chains in open-domain QA with sentence-level attribution labels against retrieved Wikipedia passages; ClaimVerify (Liu et al., 2023a), which assesses generative search engine responses by verifying check-worthy sentences against cited documents with binary factuality labels; FactCheck-GPT (Wang et al., 2024), which decomposes LLM responses to search queries into atomic facts; ExpertQA (Malaviya et al., 2024), consisting of expert-curated queries across 32 domains where system responses are verified against evidence documents; LFQA (Chen et al., 2023), where LLM long-form answers conditioned on retrieved or random documents are labeled; and RAGTruth (Niu et al., 2024), a retrieval-augmented generation benchmark where outputs grounded in retrieved passages are annotated.