ContextCheck: Sentence-Level Faithfulness Verification with Context-Aware Disambiguation

Yueqin Yin[1,2*] **Yaxi Li**[2] **Xin Liu**[2,3*] **Xun Wang**[2] **Kaiqiang Song**[2] **Simin Ma**[2] **Shujian Liu**[2] **Sathish Reddy Indurthi**[2] **Haoyun Deng**[2] **Pengcheng He**[2] **Mingyuan Zhou**[1†] **Song Wang**[2†]

[1]The University of Texas at Austin [2]Zoom [3]University of Michigan

yueqin.yin@utexas.edu, mingyuan.zhou@mccombs.utexas.edu, song.wang@zoom.us

Abstract

Large language models often hallucinate, producing content that is factually incorrect or not grounded in the sources. Reliable faithfulness verification is critical for trustworthy deployment. In the provided-source (closed-world) setting, existing verifiers either classify whole passages in one step or check sentences independently, overlooking cross-sentence context. We present **ContextCheck**, a framework for sentence-level faithfulness verification with context-aware disambiguation. Each sentence is verified against the grounding document while conditioning on preceding sentences, enabling pronouns and references to be resolved directly in context. This design avoids the separate decontextualization step of rewriting claims into self-contained forms, casting verification as a context-conditioned task. Fine-tuned from Llama-3.1-8B-Instruct, ContextCheck sets a new state of the art on three context-dependent datasets; it improves Macro F1 by over 10 points compared to the strongest baselines, and matches or slightly surpasses the strongest baselines on 14 standard single-sentence datasets compared to prior 8B-scale verifiers (average Macro F1 73.5 vs. 72.8). These results show that ContextCheck offers a practical and effective approach for sentence-level hallucination detection. Source code is available at https://github.com/yinyueqin/ContextCheck-Verifier.

1 Introduction

Large language models (LLMs) have rapidly advanced from fluent text generators to capable task solvers, yet they are persistently plagued by hallucination: generating content that is factually incorrect or unfaithful to a given source. These errors stem naturally from the statistical pressures of pretraining and from evaluation pipelines that reward confident outputs over calibrated abstention (Kalai et al., 2025; Zhang et al., 2025). This fundamental weakness undermines reliability, limits trustworthy deployment, and therefore the ability to reliably identify these hallucinations is essential to develop trustworthy and responsible LLM services (Wang et al., 2025).

The community commonly distinguishes two axes: factuality (consistency with world knowledge) and faithfulness (consistency with a specified source) (Cossio, 2025). Factuality requires open-domain retrieval; faithfulness can be assessed in a closed-world setting by verifying claims against a grounding document (Min et al., 2023; Wang et al., 2024).

In this work, we focus on fine-grained faithfulness verification, where the goal is to assign hallucination labels to each individual sentence within a generated response. Existing sentence-level verifiers (Tang et al., 2024a; Zheng and Lee, 2025; Seo et al., 2025) evaluate sentences independently, and therefore struggle when verification requires resolving references across sentences. For example, in Fig. 1, the claim "Its body is made of titanium" can only be validated by linking "Its" back to "Stellar Pro" mentioned earlier. To address such ambiguity, current pipelines often add a decontextualization step, rewriting sentences into self-contained forms before verification (Wanner et al., 2024b). This, however, introduces additional cost and potential inconsistencies, with performance highly sensitive to decomposition choices (Hu et al., 2025). More recently, whole-passage verifiers such as HallOumi (Webb and Schuler, 2025) avoid explicit rewriting by predicting sentence-level labels for an entire input response, but this design may sacrifice sentence-level precision and perform joint, whole-passage inference.

To address this gap, we introduce ContextCheck, a framework that performs sentence-level faithfulness verification with context-aware disambigua-

[*]Work done during an internship in Zoom GenAI.
[†]Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 13419–13468
July 2-7, 2026

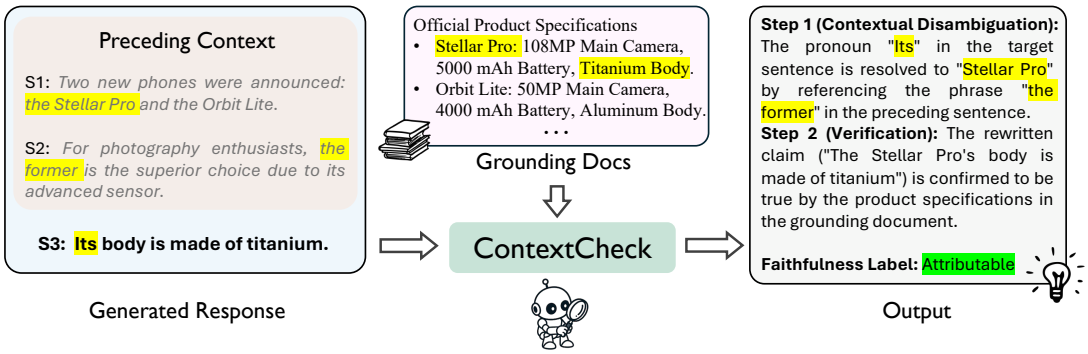

Figure 1: Overview of the ContextCheck framework for sentence-level faithfulness verification. The verifier conditions on both the grounding document and preceding sentences in the response to resolve contextual references (*e.g.*, pronouns) before verification. In this example, "Its body" is resolved to "Stellar Pro" using prior context, and the claim is then verified against the product specifications, yielding the label *Attributable*.

tion. As shown in Fig. 1, when verifying a sentence S_i, the verifier conditions on both the grounding document and all preceding sentences in the response, $C_i = S_1, \cdots, S_{i-1}$, which we define as the *context* in this work. This design resolves pronouns and other references during verification. Our approach thus shifts verification from isolated, context-free claims to claims evaluated within the natural, evolving flow of discourse.

We build training data that combines three context-dependent sources, including dialogues and QA, with standard single-sentence datasets that provide a baseline for basic verification ability. We further introduce a taxonomy of 14 hallucination types derived from a systematic analysis of existing evaluation benchmarks. Our experiments show that a fine-tuned Llama-3.1-8B-Instruct model achieves state-of-the-art performance on both established single-sentence datasets and our newly developed context-dependent datasets, substantially surpassing other 7B/8B-scale verifiers. Furthermore, our fine-grained evaluation offers a detailed diagnosis of model behavior, revealing the strengths and limitations of both ContextCheck and existing baselines across different hallucination types. Our contributions can be summarized as:

- We present ContextCheck, a new framework for faithfulness verification that performs sentence-level verification while conditioning on preceding sentences, allowing contextual references to be resolved during verification.
- We construct context-dependent datasets for both training and evaluation, and introduce a taxonomy of 14 hallucination types to support fine-grained analysis.
- We show that a fine-tuned ContextCheck verifier achieves state-of-the-art results on both standard single-sentence datasets and our new context-dependent evaluation datasets, substantially outperforming comparable scale models.

2 Related Work

Sentence- vs. passage-level faithfulness verifiers Faithfulness verification in the closed-world setting typically falls into sentence-level or passage-level designs. Sentence-level verifiers classify each sentence with respect to a grounding document, often without using discourse context. Examples include MiniCheck (Tang et al., 2024a), HHEM (Vectara, 2024), and FactCG (Lei et al., 2025). Recent models augment sentence-level decisions with chain-of-thought (CoT) rationales, such as Reasoning-CV-8B (Zheng and Lee, 2025) and ClearCheck-8B (Seo et al., 2025), improving transparency but still verifying sentences in isolation. In contrast, HallOumi-8B (Webb and Schuler, 2025) predicts labels jointly for an entire response, avoiding explicit claim rewriting but incurring monolithic inference and potential loss of sentence-level precision. Our approach, ContextCheck, preserves sentence-level granularity while conditioning on preceding sentences, enabling reference resolution during verification without the complexity of whole-passage joint reasoning.

Decompose/Decontextualize-then-Verify vs. Context-Conditioned Verification A complementary line of work mitigates context ambiguity through decomposition or decontextualization. FactScore (Min et al., 2023) extracts atomic claims, while methods such as DnDScore (Wanner et al., 2024b), SUCEA (Liu et al., 2025a), and PASS-FC (Zhuang, 2025) rewrite ambiguous spans into self-contained forms. These strategies can simplify

retrieval but introduce extra cost and errors, with performance sensitive to the chosen granularity (Hu et al., 2025; Lu et al., 2025; Wanner et al., 2024a). ContextCheck instead integrates disambiguation directly into verification: each sentence is judged against the source while conditioning on preceding sentences, eliminating external rewriting and reducing reliance on decomposition choices.

For a more comprehensive review of related work, please refer to Appendix A.1.

3 Main Method

Our methodology for developing ContextCheck has three main components. First, we design a data curation pipeline (Fig. 2) that integrates diverse corpora for fine-tuning effective verifiers on context-dependent hallucination detection. Second, we introduce a unified verification framework that applies structured chain-of-thought steps to handle both self-contained and context-dependent claims. Third, we develop a fine-grained hallucination taxonomy that enables systematic evaluation of verifier performance across different hallucination types.

3.1 Defining Context Dependency

We formally define a sentence as context-dependent if it cannot be fully understood from the grounding document alone and requires information from earlier sentences in the response. Typical cases include unresolved pronouns (*e.g.*, *he*, *it*, *this*), references to entities introduced only in prior sentences, or relative statements that are ambiguous in isolation.

3.2 Generating a Corpus from Context-Rich Dialogues

Our first goal is to build a training corpus tailored for context-dependent hallucination detection. We achieve this by generating challenging instances through a curation and annotation pipeline inspired by the TofuEval benchmark (Tang et al., 2024b).

Dialogue Selection. We sample dialogue documents from two public sources: MediaSum (NPR/CNN interviews) (Zhu et al., 2021) and MeetingBank (city council meetings) (Hu et al., 2023). These datasets were chosen because their multi-party structure produces rich discourse phenomena. Frequent speaker turns and interactive exchanges create many context-dependent cases, such as ambiguous pronouns (*e.g.*, resolving who "he" refers to) and topic shifts. These characteristics are essential for training a verifier that can resolve references internally during verification and are less common in single-author news texts (Tang et al., 2024b).

Topic Generation. For each dialogue, we generate three focal points to guide the summarization task. Using a zero-shot prompt (shown in Appendix K.1), an LLM produces exactly three distinct topics: two labeled as *Main Topics*, which capture the central themes of the conversation, and one labeled as a *Marginal Topic*, which reflects a secondary issue. For example, in a city council meeting, the Main Topics might be *budget approval* and *zoning policy*, while the Marginal Topic could be *community comments on traffic*. These topics are used only to guide the summarization prompt. They keep the summary focused on the selected theme and reduce off-topic content.

Topic-Focused Summarization. After generating three topics for each document, we then prompt six models to generate topic-focused summaries. The set covers both closed-source models (GPT-4o, Claude-Sonnet-3.5) and open-source models (Llama-3.1-8B-Instruct, Llama-3.2-3B-Instruct, Qwen2.5-7B-Instruct, and Mistral-7B-Instruct). For each topic, every model produces one summary at a specified length: short (1–3 sentences), medium (4–5 sentences), or long (6–7 sentences). The motivation is to diversify the summaries in both style and complexity. Critically, our prompts are engineered to elicit strong inter-sentence coherence and co-reference. This naturally produces sentences that depend on prior context, which cannot be verified in isolation and are therefore valuable for training ContextCheck (prompts are shown in Appendix K.2).

Detecting Contextual Dependency. Each generated summary is segmented into sentences using the NLTK toolkit. We then use GPT-4.1 to determine whether each sentence is context-dependent and filters out a set of context-dependent sentences, which are then used for subsequent factuality annotation (Appendix K.5).

Factuality Annotation via Multi–LLM Consensus. After selecting the context-dependent sentences, we assign each one a hallucination label. Following ClearCheck (Seo et al., 2025), we use three hallucination labels: *Attributable*, *Not Attributable*, and *Contradicted*. TofuEval (Tang et al.,

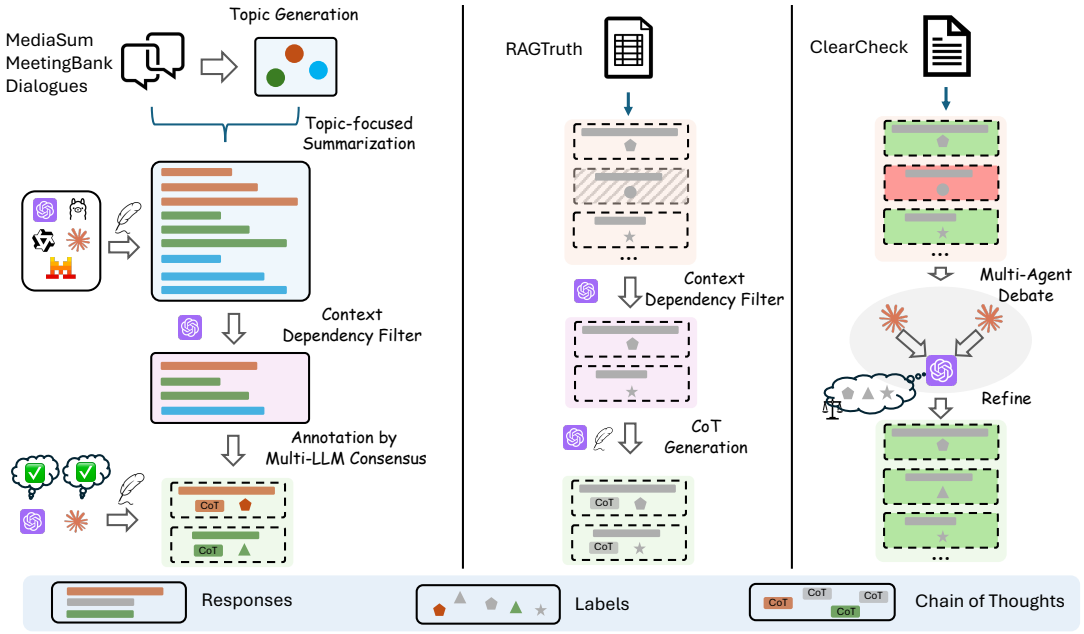

Figure 2: Overview of our data curation pipeline, which combines three sources: dialogue summarization (MediaSum, MeetingBank), RAGTruth, and ClearCheck. Each source is processed with tailored steps such as topic-focused summarization, context-dependency filtering, or multi-agent re-annotation, yielding high-quality datasets for training ContextCheck.

2024b) employs expert linguists to annotate its evaluation datasets, producing high-quality labels but at a cost that does not scale to training-sized corpora. To enable large-scale training, we design an annotation pipeline that uses multiple LLMs as judges. Prior work (Jacovi et al., 2025; Liu et al., 2025b; Li et al., 2024) shows that aggregating judgments from diverse LLMs can reduce individual bias and improve reliability in factuality evaluation. In our pipeline, a panel of strong LLMs (GPT-4.1, GPT-4o, and Claude-Sonnet-4.0) independently labels each sentence using our structured verification prompt (the detailed prompt is shown in Appendix K.6). For each annotation, the grounding document is given along with all preceding sentences from the same response as context. This setup enables the annotator LLMs to resolve references during verification and to produce both a CoT rationale and a final label. We retain only items with unanimous consensus labels across annotators, yielding a context-aware training set for verifier fine-tuning.

3.3 Curating RAGTruth for Context-Aware Training

To broaden our training data beyond dialogue summarization, we incorporate the RAGTruth corpus (Niu et al., 2024), which contains grounding documents, model-generated responses, and fine-grained human-annotated hallucination labels from summarization, question answering and data-to-text tasks. We adapt this corpus by applying our context-dependency filter (Section 3.2). Responses in which all sentences are self-contained are discarded; if at least one sentence is context-dependent, we retain the entire response. Training on such data encourages the model to recognize when preceding context is necessary for judgment and when it can be safely ignored. Since gold labels are already available, we do not re-annotate. Instead, we prompt GPT-4.1 with the existing label to produce a chain-of-thought verification, following our established pipeline (Section 3.5).

3.4 High-Quality Single-Sentence Data Curation

Although our main focus is context-dependent verification, the model must also handle single-sentence claims. We therefore reuse the ClearCheck training set (Seo et al., 2025). This corpus includes ANLI data (Nie et al., 2019) and synthetic multi-hop verification examples. We keep the grounding documents and single-sentence responses. ClearCheck provides hallucination labels, but we found that some of them are incorrect. To refine the dataset, we design a curation pipeline based on a multi-agent debate inspired by (Koupaee et al., 2025). Two annotators are assigned different roles. Agent A is given the original ClearCheck label and, following our prompt format (Section 3.5), explains why the label is correct. Agent B does not see the label. It generates a chain-of-thought explanation and proposes a label. If both outputs match,

we keep the instance. If they disagree, the instance is passed to a third adjudicator model (*e.g.*, GPT-5). The adjudicator reviews the two rationales and makes the final decision. According to (Koupaee et al., 2025), a single round of debate and adjudication can ensure high accuracy. Considering cost, we therefore only adopt this single round setting. The result is a refined single-sentence dataset that provides a reliable foundation for calibrating the verifier's basic verification ability.

3.5 Prompt Construction and Hallucination Taxonomy

This section details the two-level hallucination taxonomy that defines our verification standards and the structured CoT prompt architecture designed to implement these standards during both data annotation and model inference.

A Two-Level Hallucination Taxonomy. To systematically analyze hallucination patterns, we design a two-level taxonomy that serves two goals: fine-grained error analysis and practical model training. At the first level, we define 14 hallucination subtypes (R1–R14), listed in Table 11. This schema enables detailed error analysis across datasets, as shown by the distribution in Fig. 6. For example, R7 (*Invented Mechanism*) covers cases where a model fabricates a process or explanation, such as stating "The new policy works by using an AI algorithm to monitor citizens" when the document contains no such mechanism. R11 (*Subjective Value Judgment*) captures opinions injected into the output, such as "This decision was a wise and fair move," which also lacks grounding. However, training 7B–8B models to separate all 14 subtypes is difficult. To simplify the learning objective, we group them into seven high-level principles (C1–C7), listed in Table 12. These principles capture core aspects of faithfulness such as scope, causality, modality, and entity integrity, and each maps to one or more subtypes (see Appendix Table 13). For instance, both R7 and R11 fall under the broader principle of *Subjective and Normative Claims* (C5). More examples of this mapping are provided in the Appendix Section F. During training and inference, these principles are embedded in the system prompt to guide the verifier's reasoning.

Structured Chain-of-Thought Prompting. Our verifier prompt is engineered to guide the model through a structured reasoning process, ensuring transparent and diagnosable outputs. This format, detailed in Appendix K.12, unfolds in two main stages. It begins with a *Disambiguation* step, where the model resolves ambiguities (*e.g.*, pronouns) by leveraging the immediate preceding context. Following this, the model enters a two-phase *Analysis*, first generating a verification *Planning* step and then proceeding to *Execution*. During execution, it systematically interleaves evidence extraction from the grounding document with inference, yielding an interpretable reasoning trace. This integrated design enables the model to perform both context-aware disambiguation and rigorous faithfulness analysis within a single forward pass.

3.6 Supervised Fine-tuning Configurations

We use supervised finetuning (SFT) to align our ContextCheck verifier with the high-quality labels and reasoning chains from our curated training dataset. To study the effect of different supervision signals, we train two configurations: `context_direct` and `context_cot`. Direct supervision provides the model with the grounding document, the context (preceding sentences), and the target sentence. The model is trained to output only one of the three hallucination labels: *Attributable*, *Not Attributable*, or *Contradicted*. CoT supervision (`context_cot`) uses the same inputs but requires the model to generate a reasoning trace that justifies its decision before producing the final label.

4 Experiments

4.1 Experimental Setup

Training Data. Our supervised finetuning corpus consists of 96,000 samples drawn from three sources: 6,000 context-dependent dialogues from MediaSum and MeetingBank, 8,000 context-filtered examples from RAGTruth, and 82,000 refined single-sentence cases from ClearCheck (Seo et al., 2025). Full construction details are provided in Appendix Section C.

Evaluation Benchmarks. Our evaluation covers both context-dependent and single-sentence verification. For context-dependent tasks, we constructed three new datasets for evaluation, namely Context_MediaSum, Context_MeetingBank, and Context_RAGTruth. Context_MediaSum and Context_MeetingBank were created using held-out documents following the same construction process as in training. To ensure label quality, we added GPT-5 to the annotation panel and required unanimous agreement from all four models before keeping

a sample. Context_RAGTruth was derived from the original RAGTruth evaluation set by retaining only sentences identified as context-dependent. To further validate the reliability of our automatically labeled test sets, we conducted a human verification study on Context_MediaSum and Context_MeetingBank (see Appendix Section C). For single-sentence verification, we evaluated on standard benchmarks, including LLM-AggreFact (Tang et al., 2024a), as well as CoverBench (Jacovi et al., 2024a), Hover (Jiang et al., 2020), and SciFact (Wadden et al., 2020). These datasets, incorporated by ClearCheck (Seo et al., 2025), provide established grounds for testing foundational attribution ability. A detailed breakdown of evaluation sample counts for all benchmarks is provided in Appendix Table 4, and the detailed distribution of hallucination types across these datasets is analyzed in Appendix Section G.

Evaluation Metrics. To address class imbalance in faithfulness evaluation, we report Macro F1 and Balanced Accuracy. Macro F1 gives equal weight to each class, while Balanced Accuracy averages recall across classes to avoid inflated scores on skewed datasets. Definitions, formulas, and details of our bootstrapping procedure for stabilized scores are provided in Section D.

Implementation Details. We fine-tuned Llama-3.1-8B-Instruct (Grattafiori et al., 2024) as the backbone for ContextCheck using the LLaMA-Factory[1] framework. For evaluation, we adhered to the settings used for baselines by employing greedy decoding (temperature set to 0) for both Direct and CoT variants. During the evaluation process, we fixed the random seed at 2048. Additional implementation details are provided in Section C.

4.2 Main Results

We compare ContextCheck against a suite of recent open-source fact-checking models. These include LettuceDetect (Kovács and Recski, 2025), representing encoder-based approaches, and MiniCheck-7B (Cai et al., 2024), a binary classifier fine-tuned from InternLM2.5-7B for verifying isolated sentences. We also benchmark against three models fine-tuned from Llama-3.1-8B-Instruct: Reasoning-CV-8B and ClearCheck-8B, which perform sentence-level CoT verification, and HallOumi-8B, which processes an entire response in a single forward pass and outputs sentence-level labels by jointly predicting a decision for each sentence. Finally, we include GPT-4.1 as a reference to represent the performance of proprietary frontier models.

Results. *(1) ContextCheck demonstrates state-of-the-art performance on single-sentence benchmarks.* To evaluate foundational verification capabilities, we test our framework alongside comparisons on 14 established single-sentence benchmarks. As shown in Table 1, our models achieve the highest average performance. ContextCheck-CoT emerges as the top-performing model with an average Macro F1 of 73.46%, closely followed by ContextCheck-Direct at 73.20%. These models outperform other strong 8B parameter verifiers like ClearCheck-8B (72.77%) and Reasoning-CV-8B (70.59%), as well as the MiniCheck-7B (71.79%). Beyond aggregate scores, a fine-grained analysis by hallucination type (see Appendix Fig. 7) shows that both variants of ContextCheck consistently obtain the highest per-type recall across nearly all 14 hallucination categories. In particular, ContextCheck-CoT achieves notable gains on quantitative hallucinations (R2 in our taxonomy) that require arithmetic or interpretation of structured data. This advantage aligns with its superior performance on CoverBench which requires handling complex numerical reasoning and tabular calculations. We attribute these improvements to the structured CoT design: the explicit `[Planning]` stage before `[Execution]` compels the model to articulate intermediate steps, thereby enhancing its ability to reason faithfully with structured evidence.

(2) ContextCheck significantly outperforms all baselines on context-dependent benchmarks. ContextCheck establishes a new state-of-the-art on our newly curated context-dependent datasets, surpassing baselines by a substantial margin. For instance, as detailed in Table 2, **ContextCheck-Direct** achieves a Macro F1 of 91.40% on Context_MediaSum. This result not only exceeds the next-best baseline, HallOumi-8B (80.08%), by over 11 points but also rivals the performance of the frontier model GPT-4.1 (91.66%). We attribute this success to our method's ability to bridge the gap between sentence-level precision and context awareness. Unlike whole-passage verifiers that may miss fine-grained errors, or isolated-sentence verifiers that lack necessary context, ContextCheck explicitly conditions on preceding sentences to effectively resolve ambiguities. For a detailed statistical signif-

[1] https://github.com/hiyouga/LLaMA-Factory

Dataset	LettuceDetect	MiniCheck-7B	Reasoning-CV-8B	ClearCheck-8B	HallOumi-8B	ContextCheck-Direct	ContextCheck-CoT	GPT-4.1
CNN	59.51 / 61.54	68.27 / 66.87	64.93 / 65.62	70.31 / 69.34	64.23 / 64.72	69.01 / 68.51	**71.05 / 69.83**	68.48 / 71.21
XSum	68.71 / 68.63	**77.69 / 77.50**	70.69 / 69.66	74.60 / 74.53	73.49 / 73.40	75.68 / 75.44	76.56 / 76.17	77.65 / 77.24
WiCE	83.08 / 81.76	**83.20 / 82.20**	77.76 / 75.05	83.11 / 82.04	80.58 / 78.79	81.80 / 81.92	79.31 / 78.40	81.27 / 81.35
REVEAL	77.07 / 79.43	87.95 / 80.95	86.07 / 77.19	86.91 / 79.85	87.10 / 81.01	87.88 / **83.06**	**88.99** / 82.78	88.75 / 85.26
ClaimVerify	72.70 / 74.80	75.56 / 76.31	67.89 / 69.94	76.40 / 77.55	70.97 / 72.49	**78.93 / 78.89**	75.75 / 76.59	77.28 / 77.54
FactCheck	**83.08 / 81.76**	77.51 / 77.15	77.14 / 70.73	80.90 / 79.12	79.33 / 78.86	76.98 / 78.07	79.06 / 77.99	76.64 / 78.57
ExpertQA	59.98 / 49.97	59.14 / 51.70	58.94 / **58.93**	**60.00** / 52.60	59.72 / 55.03	59.87 / 49.70	59.38 / 52.43	60.65 / 52.22
LFQA	80.79 / 81.08	86.66 / 87.14	73.79 / 74.50	86.33 / 86.88	84.00 / 85.03	**88.67 / 88.15**	86.80 / 86.94	89.05 / 89.01
MediaS	68.95 / 72.12	75.90 / 75.93	72.05 / 75.04	71.62 / 75.31	69.54 / 72.91	75.15 / 76.26	**76.27 / 77.36**	75.26 / 77.48
MeetB	72.20 / 73.00	78.34 / 76.52	77.98 / 78.48	76.95 / 79.02	76.04 / 78.44	**81.10** / 78.54	80.27 / **79.07**	83.09 / 82.27
RAGTruth	82.37 / 75.37	81.76 / 66.40	75.46 / 71.61	79.98 / 71.74	74.78 / 68.07	**86.78 / 78.64**	85.84 / 76.24	89.05 / 89.01
Coverbench	69.73 / 64.93	62.48 / 59.87	70.74 / 68.85	66.88 / 65.28	59.35 / 57.49	66.43 / 63.89	**71.33 / 69.53**	77.32 / 75.21
Hover	49.92 / 33.27	51.43 / 37.26	**55.77 / 45.19**	51.58 / 37.49	51.66 / 36.92	51.32 / 36.28	52.86 / 38.62	51.69 / 35.79
SciFact	65.50 / 65.15	**89.31 / 89.21**	87.17 / 87.49	88.08 / 87.98	83.38 / 83.65	87.72 / 87.46	87.16 / 86.45	87.03 / 87.33
Average	71.00 / 68.77	75.37 / 71.79	72.60 / 70.59	75.26 / 72.77	72.44 / 70.49	76.24 / 73.20	**76.47 / 73.46**	77.37 / 75.68

Table 1: Balanced Accuracy / Macro F1 results across datasets. The **best** and second-best results for each metric are highlighted (excluding GPT-4.1 from ranking). Our ContextCheck variants are shaded for clarity.

Dataset	LettuceDetect	MiniCheck-7B	Reasoning-CV-8B	ClearCheck-8B	HallOumi-8B	ContextCheck-Direct	ContextCheck-CoT	GPT-4.1
Context_MediaSum	61.93 / 50.10	81.55 / 78.34	71.20 / 67.30	74.63 / 68.64	83.77 / 80.08	**91.60 / 91.40**	87.89 / 87.98	93.05 / 91.66
Context_MeetingBank	56.43 / 46.14	74.01 / 71.63	68.46 / 66.84	65.90 / 61.69	81.07 / 79.72	**92.00 / 92.24**	88.42 / 88.48	90.12 / 89.88
Context_RAGTruth	81.02 / 79.49	82.16 / 76.05	74.08 / 74.57	79.33 / 76.46	80.29 / 76.89	**86.97 / 83.57**	84.21 / 80.22	88.95 / 85.06
MedHallu	85.02 / 79.73	82.21 / 77.14	64.48 / 51.84	86.00 / 81.40	70.76 / 66.14	**91.70 / 91.40**	89.60 / 90.80	97.08 / 96.89

Table 2: Model performance on our context-dependent evaluation datasets and the medical-domain generalization set (MedHallu). Metric: Balanced Accuracy / Macro F1. The **best** and second-best results for each metric are highlighted (excluding GPT-4.1 from ranking). ContextCheck variants are shaded for clarity. ContextCheck significantly outperforms all baselines on context-dependent datasets and demonstrates strong generalization ability on the unseen medical domain (MedHallu), despite not being specifically trained on medical data.

icance analysis of these performance gains, please refer to Section D.5.

(3) ContextCheck generalizes effectively to specialized domains. To examine the generalization ability of ContextCheck beyond standard benchmarks, we additionally evaluated the model on MedHallu (Pandit et al., 2025), a publicly available medical dataset derived from PubMedQA. From this dataset, we extracted 629 samples that specifically require context-dependent sentence-level verification. As shown in Table 2, ContextCheck-Direct achieves a Macro F1 of 91.40%, surpassing the strongest baseline, ClearCheck-8B (81.40%), by exactly 10 points. These findings demonstrate that ContextCheck outperforms existing baselines in the medical setting despite not being explicitly trained on medical data, confirming that our context-conditioning mechanism possesses strong generalization ability across specialized, high-stakes domains.

5 Analysis

This section discusses several research questions to better understand our method.

(a): *Is a decontextualization step beneficial?* Traditional verifiers often rely on rewriting ambiguous sentences into self-contained forms with the help of a powerful LLM before verification. Our framework avoids this preprocessing step, but we evaluate against three complementary settings for clarity. In the *Non-debiased* setting, models are evaluated directly on the original ambiguous sentences. In the *Decontextualized* setting, each ambiguous sentence is rewritten by GPT-5 [2] so that all entities and references are made explicit, removing ambiguity and yielding an approximate upper bound on performance. In the *Debiased* setting, we pair each ambiguous sentence with its GPT-5–rewritten counterpart and enforce a strict rule: a prediction is counted as correct only if the model is consistent across both versions. This penalizes "lucky guesses" (see Appendix J.1 for an example) and provides a sharper estimate of genuine reasoning ability. Further details on these procedures are provided in Appendix D.4. For consistency, all main results reported in Table 2 are based on this *Debiased* score. As shown in Fig. 3, baseline verifiers such as ClearCheck-8B show a substantial drop from their *Decontextualized* score (79.8% on

[2] In a preliminary pilot check, we manually inspected a few rewritten cases and found GPT-5 to reliably preserve semantic equivalence.

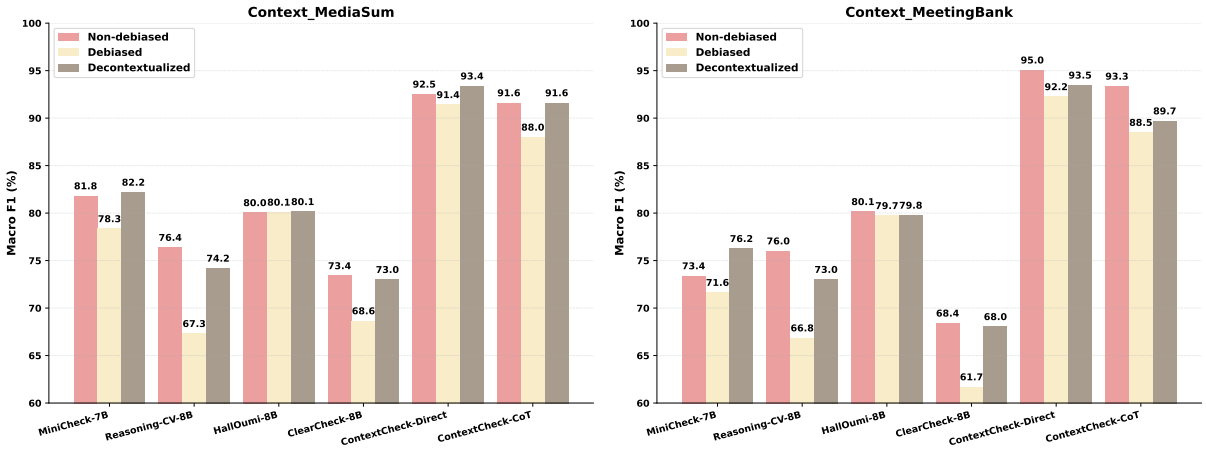

Figure 3: Model performance on context-dependent datasets. The chart compares three metrics: *Non-debiased* (standard score on original, ambiguous sentences), *Decontextualized* (score on sentences decontextualized by GPT-5), and *Debiased* (a stricter score designed to filter out "lucky guesses"). The comparison highlights the reliance of baseline models on explicit decontextualization and the effectiveness of the ContextCheck method.

Context_MeetingBank) to their stricter *Debiased* score (61.7%), indicating heavy reliance on explicit rewriting. By contrast, ContextCheck achieves the strongest results on both evaluation datasets. In particular, ContextCheck-Direct yields nearly identical outcomes across all three settings (93.5% vs. 92.2%), demonstrating that conditioning on preceding context is an effective and robust alternative to decontextualization. For more concrete context-related examples, please see Appendix Section I.1.

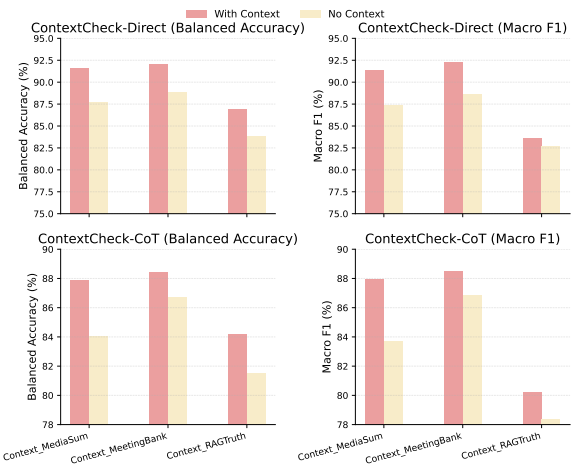

Figure 4: Ablation study on the impact of context availability. The figure compares performance on context-dependent datasets with and without preceding sentences provided as context. The results demonstrate that context is critical for improving verification accuracy.

(b): *How critical is preceding context for verification accuracy?* Fig. 4 quantifies the effect of including preceding sentences during verification. To isolate this effect, we construct a *no-context* setting for both Direct and CoT variants of our model, where preceding sentences are removed at both training and inference time. In other words, each target sentence is verified in isolation, without access to earlier sentences in the same response. Across all three datasets, removing context leads to a consistent and meaningful drop in accuracy. On Context_MeetingBank, the Macro F1 of ContextCheck-Direct decreases from 92.24% to 88.6%. On Context_MediaSum, ContextCheck-CoT falls from 87.98% to 83.7%. These results show that context is often required to resolve ambiguous references, maintain discourse continuity, and interpret claims correctly. Without this information, models are more likely to misclassify. Both Direct and CoT variants of our verifier gain substantially from context, confirming its central role in robust verification.

(c): *Does ContextCheck introduce an efficiency bottleneck?* We compare ContextCheck against a modular baseline that first resolves coreferences with FastCoref (Otmazgin et al., 2023) (90.5M parameters), then verifies each resolved sentence with ClearCheck-8B. Results on Context_MediaSum (623 samples, single H100 GPU) are shown in Table 3. ContextCheck-Direct uses the same N LLM calls and comparable wall time (317s vs. 340s), while achieving substantially higher accuracy. ContextCheck-CoT requires more wall time, but this overhead comes from generating structured reasoning traces (Disambiguation $\rightarrow$ Planning $\rightarrow$ Execution), a decoding cost shared by any CoT-based verifier (e.g., Reasoning-CV-8B, ClearCheck-8B) and unrelated to the N-pass design. Full LLM-based decontextualization would require $2N$ calls (N for rewriting + N for verification), making it at least twice as expensive.

Method	LLM Calls	Wall Time	BAcc	Macro F1
FastCoref + ClearCheck-8B	N	340s	76.5	72.4
ContextCheck-Direct	N	317s	91.6	91.4
ContextCheck-CoT	N	480s	87.9	88.0

Table 3: Cost and accuracy comparison of ContextCheck against a modular coreference-then-verify baseline on Context_MediaSum.

Due to space limits, additional analyses are deferred to Section E, where we examine the impact of training data refinement and compare sentence-level versus passage-level verification granularity.

6 Conclusion

We presented ContextCheck, a framework for sentence-level faithfulness verification that conditions on preceding sentences to resolve ambiguity in generated responses. In addition, we constructed new datasets specifically designed for context-dependent verification. By fine-tuning a Llama-3.1-8B-Instruct verifier on this curated corpus, ContextCheck achieves state-of-the-art performance on context-dependent benchmarks while maintaining strong results on single-sentence tasks. We also conducted a fine-grained error analysis based on 14 hallucination types offering insights into the strengths and weaknesses of different verifiers. Our findings show that context-aware, sentence-level verification improves both accuracy and interpretability, providing a practical approach to faithfulness verification.

Limitations

Our training relies solely on SFT. While ContextCheck-Direct achieves strong performance, the gains of ContextCheck-CoT over its direct counterpart remain limited in certain settings, suggesting that further work could explore stronger post-training methods such as reinforcement learning from human feedback (RLHF) or direct preference optimization (DPO) to improve the model's chain-of-thought reasoning ability.

Impact Statement

This work develops methods for faithfulness verification of LLM outputs to improve their reliability. A potential risk is that incorrect verifier judgments may create a false sense of reliability, especially in high-stakes domains, so further domain-specific validation is required before deployment.

Acknowledgements

Y. Yin and M. Zhou acknowledge the support of NSF-IIS 2212418. We thank anonymous reviewers of ARR for their thoughtful comments.

References

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Jian Chen, Peilin Zhou, Yining Hua, Yingxin Loh, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. 2024. Fintextqa: A dataset for long-form financial question answering. *ACL*.

Manuel Cossio. 2025. A comprehensive taxonomy of hallucinations in large language models. *arXiv preprint arXiv:2508.01781*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *ICML*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yuzhe Gu, Ziwei Ji, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. Anah-v2: Scaling analytical hallucination annotation of large language models. *NeurIPS*, 37:60012–60039.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *NeurIPS*, 28.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? *NAACL*.

Yebowen Hu, Tim Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. Meetingbank: A benchmark dataset for meeting summarization. *ACL*.

Alon Jacovi, Moran Ambar, Eyal Ben-David, Uri Shaham, Amir Feder, Mor Geva, Dror Marcus, and Avi Caciularu. 2024a. Coverbench: A challenging benchmark for complex claim verification. *arXiv preprint arXiv:2408.03325*.

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

Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, and 1 others. 2025. The facts grounding leaderboard: Benchmarking llms' ability to ground responses to long-form input. *arXiv preprint arXiv:2501.03200*.

Ziwei Ji, Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2024. Anah: Analytical annotation of hallucinations in large language models. *ACL*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. *ACL*.

Mahnaz Koupaee, Jake W Vincent, Saab Mansour, Igor Shalyminov, Han He, Hwanjun Song, Raphael Shu, Jianfeng He, Yi Nian, Amy Wing-mei Wong, and 1 others. 2025. Faithful, unfaithful or ambiguous? multi-agent debate with initial stance for summary evaluation. *NAACL*.

Ádám Kovács and Gábor Recski. 2025. Lettucedetect: A hallucination detection framework for rag applications. *arXiv preprint arXiv:2502.17125*.

Deren Lei, Yaxi Li, Siyao Li, Mengya Hu, Rui Xu, Ken Archer, Mingyu Wang, Emily Ching, and Alex Deng. 2025. Factcg: Enhancing fact checkers with graph-based multi-hop data. *NAACL*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Hongjun Liu, Yilun Zhao, Arman Cohan, and Chen Zhao. 2025a. Sucea: Reasoning-intensive retrieval for adversarial fact-checking through claim decomposition and editing. *arXiv preprint arXiv:2506.04583*.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *EMNLP*.

Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. 2025b. Verifact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. *EMNLP*.

Yining Lu, Noah Ziems, Hy Dang, and Meng Jiang. 2025. Optimizing decomposition for optimal claim verification. *ACL*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. Expertqa: Expert-curated questions and attributed answers. *NAACL*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *EMNLP*.

Kushan Mitra, Dan Zhang, Sajjadur Rahman, and Estevam Hruschka. 2025. Factlens: Benchmarking fine-grained fact verification. *ACL*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ACL*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *ACL*.

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. *ACL*.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically informed multi expert scorers for coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2752–2760, Dubrovnik, Croatia. Association for Computational Linguistics.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. Medhallu: A comprehensive benchmark for detecting medical hallucinations in large language models. *Preprint*, arXiv:2502.14302.

Elisei Rykov, Kseniia Petrushina, Maksim Savkin, Valerii Olisov, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. 2025. When models lie, we learn: Multilingual span-level hallucination detection with psiloqa. *ACL*.

Wooseok Seo, Seungju Han, Jaehun Jung, Benjamin Newman, Seungwon Lim, Seungbeen Lee, Ximing Lu, Yejin Choi, and Youngjae Yu. 2025. Verifying the verifiers: Unveiling pitfalls and potentials in fact verifiers. *COLM*.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *EMNLP*.

Liyan Tang, Tanya Goyal, Alexander R Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin F Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. *ACL*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024a. Minicheck: Efficient fact-checking of llms on grounding documents. *EMNLP*.

Liyan Tang, Igor Shalyminov, Amy Wing-mei Wong, Jon Burnsky, Jake W Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, and 1 others. 2024b. Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization. *NAACL*.

Vectara. 2024. Hhem v2: A new and improved factual consistency scoring model. Vectara Blog.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *EMNLP*.

Song Wang, Xun Wang, Jie Mei, Yujia Xie, Si-Qing Chen, and Wayne Xiong. 2025. Developing a reliable, fast, general-purpose hallucination detection and mitigation service. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 971–978, Albuquerque, New Mexico. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, and 1 others. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. *EMNLP*.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024a. A closer look at claim decomposition. *ACL*.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024b. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. *arXiv preprint arXiv:2412.13175*.

Stefan Webb and Michael Schuler. 2025. Introducing halloumi: A state-of-the-art claim-verification model. Oumi's Blog.

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:NeurIPS.

Shaokun Zhang, Ming Yin, Jieyu Zhang, Jiale Liu, Zhiguang Han, Jingyang Zhang, Beibin Li, Chi Wang, Huazheng Wang, Yiran Chen, and 1 others. 2025. Which agent causes task failures and when? on automated failure attribution of llm multi-agent systems. *arXiv preprint arXiv:2505.00212*.

Zhi Zheng and Wee Sun Lee. 2025. Reasoning-cv: Fine-tuning powerful reasoning llms for knowledge-assisted claim verification. *arXiv preprint arXiv:2505.12348*.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. Mediasum: A large-scale media interview dataset for dialogue summarization. *NAACL*.

Ziyu Zhuang. 2025. Pass-fc: Progressive and adaptive search scheme for fact checking of comprehensive claims. *arXiv preprint arXiv:2504.09866*.

A More Related Work

A.1 Faithfulness Verification Data

Several datasets have been proposed for faithfulness verification, either as evaluation benchmarks or as training resources. AGGREFACT (Tang et al., 2023) evaluates factual consistency in summarization including CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018)). WiCE (Kamoi et al., 2023) evaluates textual entailment based on Wikipedia claims. REVEAL (Jacovi et al., 2024b) is a Chain-of-Thought reasoning dataset in open-domain QA, and ClaimVerify (Liu et al., 2023) evaluates groundedness in responses from generative search engines. FactCheck (Wang et al., 2024) constructs a document-level factuality benchmark, while ExpertQA (Malaviya et al., 2024) targets long-form QA with verified attributions. LFQA (Chen et al., 2024) assesses LLM-generated sentences for factual support by the documents that sourced through either human curation, model retrieval, or random selection. TofuEval (Tang et al., 2024b) focus on dialogue summarization (MediaS and MeetB). RAGTruth (Niu et al., 2024) analyzes word-level hallucinations in various domains and tasks within retrieval-augmented generation frameworks. CoverBench (Jacovi et al., 2024a) is a benchmark for complex claim verification, spanning multiple domains (finance, law, biomedical), task types (multi-hop, numerical, causal), long inputs, and standardized formats, including multiple table representations. HoVer (Jiang et al., 2020) is a fact verification dataset built on Wikipedia, where each claim requires multi-hop verification over evidence sentences that may span multiple articles. It provides both an open-domain setting, which includes document retrieval, and a closed-form setting, where the relevant documents are given. SciFact (Wadden et al., 2020) consists of expert-written scientific claims, each paired with research paper abstracts annotated with verdict labels (Supported, Refuted, or No-Info) and evidence rationales identifying the supporting or refuting sentences. Several large-scale benchmarks have also been constructed specifically to train hallucination detectors. ANAH (Ji et al., 2024) and ANAH-v2 (Gu et al., 2024) provide bilingual, sentence-level annotations with hallucination types, corrections, and retrieved fragments, scaled further through iterative self-training. Similarly, PsiloQA (Rykov et al., 2025) introduces a multilingual benchmark for span-level detection to pinpoint specific erroneous phrases. More recently, ClearCheck (Seo et al., 2025) refines existing benchmarks by correcting annotation errors and filtering ambiguous cases, yielding CLEARFACTS and GRAYFACTS subsets, and further augments training with ANLI data (Nie et al., 2019) and a newly synthesized multi-hop fact verification dataset to improve reasoning-intensive verification. To the best of our knowledge, despite these efforts in granularity, reasoning complexity, and annotation quality, existing benchmarks still treat each sentence as an isolated claim, overlooking the need for context-aware verification that models discourse dependencies such as pronoun references and anaphora.

A.2 Open-Domain Factuality Verification

The dominant paradigm for open-domain factuality verification is to decompose and verify (Min et al., 2023; Wang et al., 2024; Wei et al., 2024; Song et al., 2024; Liu et al., 2025b). This approach segments a complex response into atomic sub-claims, each of which can be individually checked against external knowledge sources. Decomposing claims sharpens the focus for retrieval and can improve the controllability of the verification process (Mitra et al., 2025). However, this strategy introduces a critical trade-off. Empirical studies show that factuality scores vary significantly with the granularity of decomposition, suggesting that the optimal claim size is verifier-dependent (Wanner et al., 2024a; Hu et al., 2025; Lu et al., 2025). Furthermore, excessive decomposition can strip away essential context, introducing ambiguity and new sources of error by obscuring entity references. Thus, decomposition creates an inherent tension between simplifying claims for verification and preserving their contextual integrity.

A.3 Decontextualization in Verification

In decompose-then-verify pipelines, decomposition often strips away essential context, leaving atomic claims ambiguous and difficult to verify. To address this, existing approaches (Wanner et al., 2024b; Liu et al., 2025a; Zhuang, 2025) employ decontextualization, rewriting claims into self-contained forms by resolving pronouns or clarifying entities. This typically requires large LLMs such as GPT-4 to identify ambiguous spans and then rewrite the original text, a process that is both costly and error-prone. While such methods treat context as something to be engineered away, our ContextCheck instead integrates disambiguation

directly into the verification process, allowing the model to reason with context.

A.4 Multi-Agent Debate for Factuality

Another relevant line of work uses multi-agent debate (Du et al., 2023; Koupaee et al., 2025), where several LLM agents critique and refine one another's reasoning to improve factual accuracy. While effective for evaluation, these frameworks are prohibitively expensive for inference, as each decision requires multiple calls to large models. Inspired by this paradigm, we adapt its core principle for a different purpose: generating high-quality training data. We leverage multi-agent debate during our data curation phase to produce reliable annotations. This allows us to distill the benefits of multi-agent reasoning into a single, compact verifier that remains highly efficient at inference time.

B Evaluation Benchmark Statistics

To provide a clear statistical overview of the datasets used in our evaluation, Table 4 presents a detailed breakdown of each evaluation benchmark. The table enumerates the total number of samples and specifies the counts of instances labeled as containing a hallucination versus those that are non-hallucinated.

Dataset	Hallucination	Not Hallucination	Total Samples
CNN	57	501	558
XSum	273	285	558
WICE	247	111	358
REVEAL	1310	400	1710
ClaimVerify	299	789	1088
FactCheck	1190	376	1566
ExpertQA	731	2971	3702
LFQA	790	1121	1911
MediaS	172	561	733
MeetB	150	627	777
RAGTruth	1535	16113	17648
CoverBench	99	139	238
Hover	133	148	281
SciFact	47	42	89
Context_MediaSum	391	232	623
Context_MeetingBank	120	113	233
Context_RAGTruth	126	528	654

Table 4: Statistics for the evaluation benchmarks, showing the distribution of samples with and without hallucinations.

C Additional Experimental Setup

Training Data Details. The training corpus is drawn from three sources. First, we constructed about 6,000 context-dependent dialogue samples by generating topic-focused summaries from 6,154 MediaSum and 1,846 MeetingBank documents. Each summary was segmented into sentences, filtered for contextual dependency, and annotated for factuality through majority voting among Claude-Sonnet-4.0, GPT-4.1, and GPT-4o. CoT trajectories from GPT-4.1 were retained for CoT-based training. Second, we curated 8,000 examples from the RAGTruth corpus (Niu et al., 2024). We applied the same context-dependency filter and preserved only responses with at least one context-dependent sentence. Ground-truth hallucination labels were kept, and GPT-4.1 was prompted to generate corresponding chain-of-thought explanations. Third, we refined 82,000 single-sentence cases from the ClearCheck dataset (Seo et al., 2025). A multi-agent debate pipeline was used: two instances of Claude-Sonnet-4.0 served as agents with different roles. Agent A was given the original ClearCheck label and produced a justification, while Agent B generated a new label independently without seeing the original. Disagreements were then adjudicated by GPT-5, yielding high-quality labels for self-contained claims and ensuring the reliability of the single-sentence training corpus.

Human Validation of Context-dependent Evaluation Dataset Labels. To further assess the reliability of our newly constructed evaluation datasets, we conducted a human validation study on Context_MediaSum and Context_MeetingBank. These datasets were derived from held-out documents and labeled through consensus among four frontier LLMs (GPT-5, GPT-4.1, GPT-4o, and Claude-Sonnet-4.0), providing a high level of initial reliability. We validated this pipeline with a human annotation study involving three annotators (two PhD students and one industry engineer). We evaluated a total of **120 samples**, comprising 60 from Context_MediaSum and 60 from Context_MeetingBank, with each set balanced between 30 hallucinated and 30 non-hallucinated cases. All annotators followed the standard instructions provided in Appendix K.6.

The results demonstrate exceptional human consistency. Pairwise agreement scores were consistently high (A-B = 0.975, A-C = 0.975, B-C = 0.967). Corresponding pairwise Cohen's κ values were κ(A-B) = 0.950, κ(A-C) = 0.950, and κ(B-C) = 0.933, all falling within the "almost perfect agreement" range. Using majority voting as the adjudicated human label, the LLM-based consensus annotation achieved an overall accuracy of **98%** (118/120). Residual disagreements, both among annotators and between humans and LLMs, were

primarily attributed to minor differences in strictness regarding borderline hallucinations rather than fundamental inconsistencies in task understanding. These results confirm that our LLM consensus labels are highly reliable and serve as trustworthy benchmarks.

Implementation Details. We fine-tuned Llama-3.1-8B-Instruct using the LLaMA-Factory framework. Full-parameter training was performed with DeepSpeed ZeRO Stage 3 for memory optimization. Key hyperparameters include a learning rate of 1.0×10^{-6} with a linear scheduler, a warmup ratio of 0.03, and 2 epochs of training. We used an effective batch size of 4 (per-device batch size of 1 with 4 gradient accumulation steps) and a maximum sequence length of 28,000 tokens. Training was conducted in BF16 precision with gradient checkpointing enabled to reduce memory usage. All experiments were run on 8 NVIDIA H100 GPUs and completed in approximately 4 hours.

D Evaluation Metrics

This section provides detailed definitions for the evaluation metrics and procedures used in our experiments.

D.1 Label Mapping for Binary Evaluation

Our evaluation framework is centered on a binary classification of faithfulness: determining whether a given statement constitutes a hallucination. While our annotation and some models produce three-class outputs (*Attributable*, *Not Attributable*, *Contradicted*), we unify these for metric calculation to ensure consistent comparison across models. Specifically, we group *Not Attributable* and *Contradicted* into a single *hallucination* class, mapped to `label=0`, and map *Attributable* to the *non-hallucination* class `label=1`.

D.2 Core Metrics

To ensure a fair evaluation on datasets with imbalanced class distributions, we use Macro F1 and Balanced Accuracy as our primary performance metrics. This approach is consistent with the evaluation of our baselines.

Macro F1 Score. This metric addresses class imbalance by calculating the F1 score for each class independently before taking their unweighted average, ensuring that performance on the minority class is weighted equally. For a given class c, the F1 score is the harmonic mean of precision (P_c) and recall (R_c):

$$\text{F1}_c = 2 \times \frac{P_c \times R_c}{P_c + R_c} \tag{1}$$

The Macro F1 score is then the arithmetic mean of the F1 scores for all classes C:

$$\text{Macro F1} = \frac{1}{|C|} \sum_{c \in C} \text{F1}_c \tag{2}$$

Balanced Accuracy (BAcc). This metric is the average of recall obtained on each class, preventing inflated scores on imbalanced datasets. It is defined as the average of the true positive rate (recall of the positive class) and the true negative rate (recall of the negative class). The formula is as follows:

$$\text{BAcc} = \frac{1}{2} \left(\frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} \right) \tag{3}$$

where TP, TN, FP, and FN represent the counts of true positives, true negatives, false positives, and false negatives, respectively.

D.3 Bootstrapping for a Stabilized Metric

It is important to note that our standard evaluation pipeline utilizes greedy decoding (temperature $= 0$) with a fixed random seed. Consequently, the inference process is fully **deterministic**: repeated evaluations on the fixed test set yield identical predictions, rendering standard variance analysis across generation runs inapplicable. To rigorous assess the stability of our metrics against data distribution variance rather than generation noise, we employ a bootstrapping procedure. This allows us to derive a robust point estimate that generalizes better than a single score calculated on a static split.

Our implementation follows these steps:

1. **Paired Resampling:** We treat the set of (ground truth, model prediction) pairs as our original sample population.

2. **Iteration:** We perform 1,000 bootstrap iterations. In each, we draw a new sample of the same size as the original population by resampling with replacement.

3. **Metric Calculation:** For each of the 1,000 bootstrap samples, we calculate the metric of interest (Macro F1 or Balanced Accuracy), which yields a distribution of 1,000 metric scores.

4. **Stabilized Point Estimate:** We compute a 95% confidence interval from this distribution. The final metric score reported in our paper is the midpoint of this interval, serving as a stabilized estimate of the model's true performance.

D.4 Scoring Settings: Non-debiased, Decontextualized, and Debiased

We report results under three complementary settings to disentangle ambiguity from genuine reasoning.

Non-debiased. Models are evaluated directly on the original (potentially ambiguous) sentences. Predictions are compared to gold labels to compute the confusion matrix and downstream metrics (Macro F1, BAcc).

Decontextualized. For each ambiguous sentence, we also generate a rewritten version using GPT-5, in which all referents and entities are made explicit (i.e., a decontextualized form). Models are then evaluated on these rewritten sentences, which approximate an upper bound of performance since ambiguity has been removed. The original gold labels are retained, and we manually verified on a sample of cases that the rewritten versions remain semantically equivalent to the originals, serving only to clarify references without altering factual content.

Debiased. We pair each ambiguous sentence with its GPT-5–rewritten counterpart and enforce a strict correctness rule. Let $y \in \{0, 1\}$ denote the gold label, and let $\hat{y}^{(a)}$ and $\hat{y}^{(r)}$ be the model's binary predictions on the *ambiguous* (base) and *rewritten* sentences, respectively. An instance is counted as correct only if both predictions match the gold label. Operationally, we construct a debiased prediction $\hat{y}^{(d)}$ as:

$$\hat{y}^{(d)} = \begin{cases} 1 - y, & \text{if } \hat{y}^{(a)} = y \text{ and } \hat{y}^{(r)} \neq y \\ \hat{y}^{(a)}, & \text{otherwise.} \end{cases} \quad (4)$$

This rule flips a base prediction only in cases where the model appears correct on the ambiguous version but fails on the rewritten one, thereby penalizing "lucky guesses." We then compute Macro F1 and Balanced Accuracy from $\hat{y}^{(d)}$ versus y, with confidence intervals estimated by paired bootstrapping (1,000 resamples).

Table 5: Statistical Significance: **ContextCheck-Direct** vs. **MiniCheck-7B**.

Dataset	Relative Improv.	p-value	Sig ($p < .05$)
CNN	+0.74	0.078	False
XSum	−2.01	0.394	False
WiCE	−1.40	0.305	False
REVEAL	−0.07	0.005	**True**
ClaimVerify	+3.37	0.016	**True**
FactCheck	−0.53	0.021	**True**
ExpertQA	+0.73	0.024	**True**
LFQA	+2.01	0.074	False
MediaS	−0.75	0.000	**True**
MeetB	+2.76	0.000	**True**
RAGTruth	+5.02	0.000	**True**
Coverbench	+3.95	0.180	False
Hover	−0.11	0.541	False
SciFact	−1.59	0.599	False
Context_MediaSum	+10.05	0.000	**True**
Context_MeetingBank	+17.99	0.000	**True**
Context_RAGTruth	+4.81	0.000	**True**
Average	**+2.65**	**3.81e-4**	**True**

Table 6: Statistical Significance: **ContextCheck-CoT** vs. **MiniCheck-7B**.

Dataset	Relative Improv.	p-value	Sig ($p < .05$)
CNN	+2.78	0.076	False
XSum	−1.13	0.504	False
WiCE	−2.15	0.097	False
REVEAL	+1.43	0.002	**True**
ClaimVerify	−3.89	0.787	False
FactCheck	+0.14	0.639	False
ExpertQA	+0.24	0.004	**True**
LFQA	+0.37	0.744	False
MediaS	+1.93	0.000	**True**
MeetB	+4.08	0.000	**True**
RAGTruth	+8.85	0.000	**True**
Coverbench	+1.04	0.000	**True**
Hover	+0.19	0.171	False
SciFact	+1.55	0.653	False
Context_MediaSum	+6.34	0.000	**True**
Context_MeetingBank	+14.41	0.000	**True**
Context_RAGTruth	+2.05	0.000	**True**
Average	**+2.25**	**0.018**	**True**

D.5 Statistical Significance Analysis

To rigorously validate that the observed performance gains are statistically meaningful and not merely artifacts of data variance, we conducted **paired bootstrapping significance tests** with 1,000 resamples across all evaluation datasets. Following the evaluation protocol established by MiniCheck (Tang et al., 2024a), this procedure computes the p-value for the performance difference between our model and the baselines. We utilize **Balanced Accuracy** as the primary metric for this analysis.

We compared ContextCheck against the two strongest baselines: MiniCheck-7B and ClearCheck-8B. Detailed statistical results are provided in Table 5 through Table 8. In these tables, the column **Significant** ($p < 0.05$) confirms

Table 7: Statistical Significance: **ContextCheck-Direct** vs. **ClearCheck-8B**.

Dataset	Relative Improv.	p-value	Sig ($p < .05$)
CNN	−1.30	0.092	False
XSum	+1.08	0.085	False
WiCE	−1.31	0.168	False
REVEAL	+0.97	0.787	False
ClaimVerify	+2.53	0.718	False
FactCheck	−3.92	0.408	False
ExpertQA	−0.13	0.000	**True**
LFQA	+2.34	0.046	**True**
MediaS	+3.53	0.016	**True**
MeetB	+4.15	0.000	**True**
RAGTruth	+6.80	0.000	**True**
Coverbench	−0.45	0.000	**True**
Hover	−0.26	0.000	**True**
SciFact	−0.36	0.474	False
Context_MediaSum	+16.97	0.000	**True**
Context_MeetingBank	+26.10	0.000	**True**
Context_RAGTruth	+7.64	0.000	**True**
Average	**+3.79**	**0.034**	**True**

Table 8: Statistical Significance: **ContextCheck-CoT** vs. **ClearCheck-8B**.

Dataset	Relative Improv.	p-value	Sig ($p < .05$)
CNN	+0.74	0.304	False
XSum	+1.96	0.207	False
WiCE	−3.80	0.042	**True**
REVEAL	+2.08	0.000	**True**
ClaimVerify	−0.65	0.385	False
FactCheck	−1.84	0.807	False
ExpertQA	−0.62	1.000	False
LFQA	+0.47	0.044	**True**
MediaS	+4.65	0.000	**True**
MeetB	+3.32	0.000	**True**
RAGTruth	+5.86	0.000	**True**
Coverbench	+4.45	0.728	False
Hover	+1.28	0.707	False
SciFact	−0.92	0.702	False
Context_MediaSum	+13.26	0.000	**True**
Context_MeetingBank	+22.52	0.000	**True**
Context_RAGTruth	+4.88	0.000	**True**
Average	**+3.65**	**0.044**	**True**

whether a statistically reliable difference exists between the two models, while the direction of that difference (i.e., whether ContextCheck is superior) is indicated exclusively by the **Relative Improvement** column.

The analysis yields two key findings. First, on standard single-sentence benchmarks, ContextCheck remains highly competitive with both baselines; on several datasets, we outperform them with significant p-values, and in instances where baselines exhibit slightly higher raw scores, the differences are typically not statistically significant, indicating comparable performance. Second, on context-dependent benchmarks (Context_MediaSum, Context_MeetingBank, and Context_RAGTruth), ContextCheck achieves substantial improvements over both MiniCheck-7B and ClearCheck-8B. Crucially, these performance gains are statistically significant across all three datasets, confirming the robustness of our method in handling context-sensitive verification.

E Ablation Analysis

E.1 Impact of Refined Training Data

Dataset	ClearCheck-8B	Context-CoT (refined)
CNN	70.31 / 69.34	69.74 / 69.03
XSum	74.60 / 74.53	**76.35 / 76.15**
WiCE	**83.11 / 82.04**	80.16 / 78.73
REVEAL	86.91 / 79.85	**88.78 / 82.59**
ClaimVerify	76.40 / **77.55**	**76.63** / 77.12
FactCheck	**80.90 / 79.12**	80.14 / 78.74
ExpertQA	**60.00** / 52.60	59.88 / **53.02**
LFQA	86.33 / 86.88	**86.85 / 87.02**
MediaS	71.62 / 75.31	**73.22 / 76.16**
MeetB	76.95 / 79.02	**78.60 / 79.86**
RAGTruth	**79.98** / 71.74	79.14 / **70.17**
CoverBench	66.88 / 65.28	**71.89 / 70.70**
Hover	51.58 / 37.49	**52.67 / 38.33**
SciFact	88.08 / 87.98	**89.06 / 89.20**
Average	75.26 / 72.77	**75.94 / 73.34**

Table 9: Balanced Accuracy/Macro F1 results of ClearCheck-8B vs. Context-CoT trained on refined ClearCheck data. Context-CoT shows consistent improvements, confirming the benefits of higher-quality labels with CoT supervision.

To isolate the contribution of data quality, we compare the baseline ClearCheck-8B verifier against our ContextCheck-CoT model when trained specifically on the refined ClearCheck dataset. As detailed in Section C, this refined dataset was curated using our multi-agent debate pipeline, leveraging consensus among multiple frontier LLMs to generate high-fidelity labels and CoT rationales.

Table 9 reports the Balanced Accuracy and Macro F1 across 14 benchmarks. ContextCheck-CoT trained on the refined ClearCheck data achieves superior average performance, demonstrating that higher-quality labels combined with explicit CoT supervision lead to more reliable gains across diverse benchmarks compared to the original baseline.

E.2 Impact of Verification Granularity: Sentence vs. Passage

How does fine-grained sentence-level verification compare to passage-level evaluation? Unlike other single-sentence benchmarks, the CNN

Dataset	CNN	CNN_split
HallOumi-8B	64.13 / 64.72	-
ContextCheck-Direct	64.25 / 67.70	69.01 / 68.51

Table 10: Comparison of passage-level verification (CNN) and sentence-level verification with aggregation (CNN_split) on the CNN dataset. Results show that splitting passages into sentences and aggregating predictions produces more accurate judgments than verifying the passage as a whole.

dataset presents claims as three-sentence passages, which prior works have typically treated as atomic units. To evaluate the benefit of fine-grained verification, we conducted an ablation study using ContextCheck-Direct. We compared verifying the passage as a monolithic whole against a *split-and-aggregate* strategy. In the latter, each passage is decomposed into individual sentences and verified contextually. The passage-level label is then derived via logical aggregation: a passage is *Attributable* only if all constituent sentences are *Attributable*; if any sentence is *Contradicted*, the passage is labeled *Contradicted*; otherwise, it is *Not Attributable*.

As shown in Table 10, this split-and-aggregate approach (CNN_split) significantly outperforms treating the passage as a single unit (CNN). At the passage level, ContextCheck-Direct performs on par with HallOumi-8B (64.25 vs. 64.13), but improves substantially to 69.01 when utilizing the split-and-aggregate method. This finding confirms that decomposing complex claims into sentence-level units allows for more precise error detection. To ensure a fair comparison, all CNN results reported in Table 1 utilize this split-and-aggregate protocol.

F Two-Level Hallucination Taxonomy

To systematically analyze hallucination patterns and guide our model's training, we developed a two-level error taxonomy designed for two distinct purposes: granular error analysis and creating a tractable learning objective.

The first level provides a fine-grained classification of 14 error subtypes (R1–R14), detailed in Table 11. Notably, this includes a specific category for *external knowledge enrichment* (R12), a prevalent error type in key benchmarks like REVEAL and FactCheck (see Fig. 6); the detailed prompt for identifying this error is provided in K.6. This prompt design makes a critical distinction between forbidden "Enrichment" knowledge (adding new facts) and permitted "Decoding" knowledge (using common knowledge for interpretation), to avoid penalizing basic reasoning while strictly controlling factual grounding. Recognizing that training a model on all 14 subtypes is a significant challenge, we abstract them into a higher-level set of seven core attributability principles (C1–C7), presented in Table 12. These seven principles are embedded in our system prompt to guide the verifier's reasoning. The relationship between these two levels is shown in Table 13, which maps each fine-grained error to a core principle.

G Hallucination Type Analysis

To better understand the challenges posed by different benchmarks, we analyzed the distribution of hallucination error types across our evaluation datasets, as detailed in Fig. 5 and Fig. 6.

Overall Distribution. Across all datasets combined, the most frequent error is *Unsupported specification/concretization* (R5), accounting for 30.4% of all hallucinations. This indicates that models frequently invent specific details not present in the source material. Other common errors include *Direct contradiction* (R1, 12.7%), *External knowledge enrichment* (R12, 11.9%), *Invented mechanism* (R7, 9.5%), and *Entity misidentification* (R8, 9.0%).

Dataset-Specific Characteristics. The error distributions vary significantly across benchmarks, revealing their distinct evaluation focuses beyond the most common errors. Each dataset exhibits a unique "error signature" that highlights different facets of faithfulness. For instance, some benchmarks are highly specialized: SciFact almost exclusively tests for *Direct contradiction* (R1, 83.0%), while CoverBench, which includes complex reasoning tasks involving structured data like tables, is unique in its focus on *Quantitative/measurement errors* (R2, 34.0%). Similarly, Hover is heavily skewed towards detecting *External knowledge enrichment* (R12, 48.5%). In contrast, benchmarks for retrieval and attribution show a more complex mix. Both REVEAL and FactCheck are balanced between R12 and R5 errors, but are distinguished by their secondary characteristics: REVEAL has a high rate of *Entity misidentification* (R8, 16.9%), whereas FactCheck notably contains many instances of *Invented mechanism* (R7, 12.4%). Other datasets present different challenges; XSum, for example, is characterized by its high

ID	Error Type	Definition
R1	Direct contradiction/opposition	Direct factual, logical, or semantic contradiction against the document.
R2	Quantitative/measurement errors	Wrong numbers, units, dimensions, or conversions.
R3	Temporal/spatial/sequential inconsistencies	Wrong time, place, sequence, or order.
R4	Scope expansion/over-generalization	Extending claims beyond the stated scope or population.
R5	Unsupported specification/concretization	Adding unsupported specific details when the document only provides general information.
R6	Logical relationship errors	Incorrect causal, conditional, or comparative relations.
R7	Invented mechanism/process/explanation	Adding mechanisms, processes, or motivations not described in the document.
R8	Entity/attribute/source misidentification	Misidentifying entities, attributes, or sources.
R9	Normative/prescriptive claims added	Adding judgments about what should/ought to be done.
R10	Modality/intensity/frequency alteration	Changing certainty, intensity, or frequency levels.
R11	Subjective value judgment added	Adding evaluative opinions not present in the document.
R12	External knowledge enrichment	Adding external verifiable facts not in the document.
R13	Context/qualifier/emphasis alteration	Removing or changing important qualifiers, conditions, or emphasis.
R14	Invalid logical inference/composition	Drawing unsupported conclusions or improper aggregation.

Table 11: Hallucination error types (R1–R14).

ID	Judging Criterion	Rule Summary
C1	Scope & Quantification	The claim must match the scope and quantifiers in the document; no over-generalization beyond the stated population or context.
C2	Causality & Logic	Causal or logical links must be explicitly stated in the document; correlation is insufficient to imply causation.
C3	Certainty & Modality	The certainty, intensity, and frequency must match the document; no upgrading possibility into certainty.
C4	Context & Qualifiers	All qualifiers, conditions, and contextual constraints in the document must be preserved.
C5	Subjective & Normative Claims	Subjective judgments, normative prescriptions, or invented explanations are only attributable if explicitly present in the document.
C6	Temporal & Spatial Consistency	Timeline, sequence, and location must remain consistent with the document.
C7	Entity Integrity	All entities, attributes, and measurements must exactly match the document description.

Table 12: Seven high-level judging criteria (C1–C7).

Criterion	Error Types	Mapped IDs	Explanation
C1 Scope & Quantification	Scope / Specification	R4, R5	Over-generalization or unsupported specification.
C2 Causality & Logic	Logical errors / Invalid inference	R6, R14	Incorrect causal/comparative relations or invalid inference/aggregation.
C3 Certainty & Modality	Modality alteration	R10	Altered certainty, intensity, or frequency.
C4 Context & Qualifiers	Qualifier alteration	R13	Loss or change of qualifiers, conditions, or emphasis.
C5 Subjective & Normative Claims	Mechanisms / Normativity / Subjectivity	R7, R9, R11	Added mechanisms, prescriptions, or subjective judgments.
C6 Temporal & Spatial Consistency	Temporal/spatial inconsistencies	R3	Time, location, or sequence inconsistencies.
C7 Entity Integrity	Contradiction / Quantitative / Entity / Enrichment	R1, R2, R8, R12	Direct contradiction, numeric error, entity/source misidentification, or external enrichment.

Table 13: Mapping between judging criteria (C1–C7) and hallucination error types (R1–R14).

proportion of *Entity misidentification* (R8, 25.6%) following the lead of R5, while ClaimVerify's signature includes a significant presence of *Context/qualifier alteration* (R13) as its second-most common error.

Analysis of Context-Dependent Datasets. A key contribution of our work is the curation of context-dependent datasets. When comparing these datasets to their corresponding versions within the LLM-AggreFact benchmark, we observe a significant shift towards more complex and logical error types. While the original MediaS, MeetB, and RAGTruth datasets are heavily dominated by direct errors like *Unsupported specification* (R5), our context-rich versions present a more diverse and challenging error profile. For instance, there is a notable increase in errors requiring deeper reasoning, such as *Invalid logical inference/composition* (R14) in Context-MediaSum and a significant rise in *Invented mechanism/process* (R7) across all three curated datasets. Our Context-MeetingBank particularly stands out, shifting from an R5-dominance of 46.1% in the original to a profile led by more nuanced errors like *Invented mechanism* (R7), *Subjective value judgments* (R11), and *Qualifier alterations* (R13). Overall, this analysis confirms that our newly curated datasets introduce a more complex set of hallucinations, pushing models beyond simple fact-checking to more sophisticated contextual and logical reasoning.

H Fine-Grained Performance Analysis

To gain a more granular understanding of different models' capabilities, we evaluate their performance on each of the 14 error subtypes defined in our taxonomy. The results, presented in Fig. 7, reveal the distinct strengths of our two ContextCheck models.

ContextCheck-Direct's Broad Superiority. The analysis reveals that our ContextCheck-Direct model is the top performer across a wide majority of error categories, demonstrating remarkable robustness. It not only excels at identifying direct factual errors like *Direct contradiction* (R1) and *Unsupported specification* (R5), but surprisingly, it also outperforms the CoT variant on several complex logical categories, including *Invented mechanism* (R7) and *Invalid logical inference* (R14).

CoT's Specialization and Interpretability. While ContextCheck-CoT shows lower accuracy on several benchmarks compared to the Direct model, it offers a crucial advantage: interpretability. The explicit chain-of-thought provides a clear, step-by-step reasoning process that allows for easier error analysis and helps build trust in the verification outcome. Furthermore, the CoT model demonstrates a clear specialization in tasks that benefit from this detailed procedural verification. Its most significant advantage is on *Quantitative/measurement errors* (R2), where its chain-of-thought explanations is well-suited for the complex mathematical calculations found in benchmarks like CoverBench.

Identifying Universal Challenges. The fine-grained analysis also highlights a universal challenge for all current verifiers: *Context/qualifier/emphasis alteration* (R13). As shown in panel (m) of Fig. 7, all models struggle significantly with this category, with the highest accuracy barely reaching 45%. This indicates that detecting subtle changes in scope, conditions, or emphasis remains a difficult open problem for faithfulness verification.

Summary of Analysis. In summary, this per-category breakdown validates the robustness of both ContextCheck models. It highlights ContextCheck-Direct as a powerful and broadly effective verifier, while establishing ContextCheck-CoT as a valuable alternative that trades some accuracy for crucial interpretability and excels in tasks requiring detailed procedural or quantitative verification.

I Case Analysis

I.1 Debiased Analysis

The following case (see J.1), drawn from the Context_MediaSum dataset and evaluated with the ClearCheck verifier, illustrates how disambiguation reveals hidden weaknesses in verification models and motivates the use of our Debiased score. Without disambiguation, the model produces the correct label—but only by relying on shallow heuristics, essentially a *lucky guess*. Once disambiguation is introduced, the brittleness of this reasoning is exposed: the model reverses its judgment, showing that it did not substantively engage with the claim–document relationship.

Pre-disambiguation. Given the original ambiguous target sentence, the model returned `[Not Attributable]`. While the label is correct, the

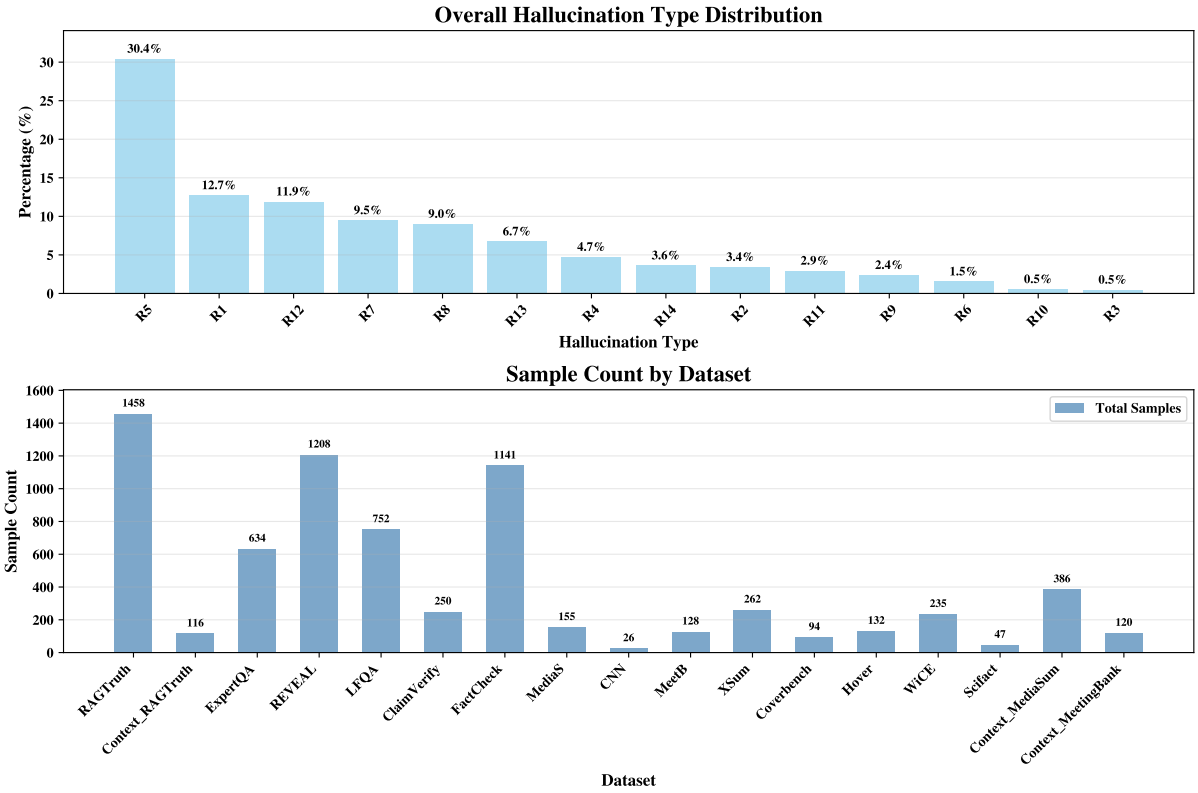

Figure 5: Statistical overview of hallucination distributions and dataset sizes. (Top) The overall percentage distribution of the 14 fine-grained hallucination error types across all combined datasets. (Bottom) The total number of samples for each individual evaluation benchmark.

path to it was superficial. The model relied on a brittle heuristic—spotting that the transcript *does not explicitly* (i) state that state-level actions "reflected urgency" and (ii) identify shortages *specifically* in New York and New Jersey—rather than verifying claim–evidence alignment. It even exhibited a minor extraction flaw by implicitly reading a *nationwide* shortage as if it were "across both states." Thus, the correctness here is accidental (heuristic match) rather than evidence of genuine understanding.

Post-disambiguation. After GPT-5 rewrote the sentence to make the referents explicit ("the measures by New York and New Jersey"; "healthcare workers across New York and New Jersey"), the model flipped to `[Attributable]`. This exposes two reasoning errors. First, a *scope narrowing* error: it collapsed a transcript-level claim about *nationwide* shortages into a *state-specific* assertion (NY/NJ). Second, an *invented causal attribution*: it inferred that the shortages themselves meant the measures "reflected urgency," despite no such attribution in the transcript.

Implication. This case shows how disambiguation functions as a stress test: removing referential ambiguity strips away heuristic shortcuts and reveals whether the model maintains claim–evidence fidelity under stricter conditions. Conventional accuracy would (mis)credit the pre-disambiguation output as a success, masking the shallow reasoning and extraction imprecision. By contrasting predictions across the ambiguous and disambiguated forms, the *Debiased Score* penalizes such accidental correctness and delivers a sharper estimate of true reasoning robustness.

I.2 Benefit of Disambiguation

The following case (see J.2), drawn from the Context_MeetingBank dataset and evaluated with the Reasoning-CV verifier, illustrates how disambiguation strengthens verification by *eliminating ambiguity that would otherwise mislead the model into spurious errors*. In the ambiguous form, the model misclassified the target sentence as `[Refute]`, overlooking broader contextual evidence. After disambiguation, the claim was rewritten with explicit referents, enabling the model to correctly align the claim with the grounding evidence and recover the

right label.

Pre-disambiguation. The original ambiguous target sentence—"This initiative is part of a broader effort to create a vibrant community space, which includes plans for additional housing and amenities in the surrounding area."—was misclassified as `[Refute]`. Because the phrase "this initiative" was vague, the model failed to connect it to the Lincoln Park project mentioned in the transcript. Instead, it judged the sentence as an overgeneralization, treating the broader claim as unsupported even though the transcript explicitly described expanded housing units, a teachers' village, and additional amenities. This reflects an error of *under-attribution*: the model dismissed valid evidence due to referential ambiguity.

Post-disambiguation. After GPT-5 rewrote the target sentence with explicit referents—"The Lincoln Park Landscaping Project will enhance the park using grant funding up to $981,280 along with impact fees, and is part of a broader effort that includes added housing and amenities in the surrounding area."—the model corrected its label to `[Support]`. The clarified form explicitly anchored the claim to "Lincoln Park Landscaping Project" and the exact funding figure, both of which appear verbatim in the transcript. This eliminated the ambiguity around "this initiative" and guided the verifier to integrate housing and amenity expansions into the correct reasoning chain.

Implication. This case demonstrates the *benefit of disambiguation*: rewriting the target sentence with explicit referents removed the vagueness around "this initiative" and directly tied it to the Lincoln Park Landscaping Project described in the transcript. This clarification allowed the model to correctly integrate the evidence of housing, the teachers' village, and other amenities, leading to the right attribution. In other words, disambiguation did not change the underlying facts but rephrased the claim so that the model could properly ground it in context.

I.3 ContextCheck Can Perform Self-Disambiguation

The following case (see J.3), drawn from our `Context_RAGTruth` dataset and verified with the ContextCheck-CoT model, illustrates how the verifier can perform self-disambiguation without external rewriting. The target sentence—"The five-year-younger player has achieved five sacks this season, compared to Emmanuel Ogbah's 2.5."—contains an underspecified referent: the phrase "the five-year-younger player" does not explicitly identify which entity is being compared to Ogbah. ContextCheck, however, exploited the *preceding response context* to correctly resolve the reference to *Chase Young*, who had been introduced earlier as a potential Dolphins trade target. It then decomposed the claim into verifiable components and aligned each with the transcript: (i) Young is five years younger than Ogbah, (ii) Young has five sacks this season, and (iii) Ogbah has 2.5 sacks. All components were directly supported. On this basis, the model produced the correct judgment of `[Attributable]`. This case highlights ContextCheck's capacity for *self-disambiguation*: by integrating response-level context into its reasoning, the verifier can resolve vague descriptive references, anchor them to the appropriate entities, and validate factual claims without requiring any externally rewritten inputs.

I.4 ContextCheck Excels in Numerical Reasoning Verification

The following case (see J.4), drawn from our `CoverBench` dataset, showcases ContextCheck-CoT's strong performance in verifying claims that require multi-step numerical reasoning. The target sentence—"The net change in cash during 2016 was -312."—cannot be validated by surface-level extraction alone. Although the grounding document provides disaggregated cash flows from operating, investing, and financing activities, the net change is not explicitly stated and must be computed as $262 + (-472) + (-102) = -312$.

Here, ContextCheck-CoT excelled: it correctly located the relevant table entries, carried out the arithmetic step by step, and confirmed that the computed result precisely matched the claim, yielding the correct decision of `[Attributable]`. This example illustrates that ContextCheck-CoT not only reads financial tables but also reasons through intermediate calculations in a structured and faithful manner.

By contrast, all baseline models failed. MiniCheck-7B misclassified with low confidence, Reasoning-CV-8B and HallOumi-8B incorrectly treated year-end balances as net changes, and ClearCheck-Direct mistakenly subtracted 2016 balances from 2017, producing a spurious "+6.019B" result. These consistent errors underscore a com-

mon limitation: the baselines rely on shallow extraction heuristics and cannot reliably handle derived numerical quantities.

Overall, this case demonstrates that ContextCheck-CoT's structured chain-of-thought provides a decisive advantage in *faithful numerical verification*. It integrates heterogeneous evidence, performs intermediate computations correctly, and avoids the systematic pitfalls that trap baseline verifiers when claims involve quantities that must be inferred rather than read off directly.

J Use of AI Assistants

We acknowledge the use of LLMs to assist with language polishing and writing refinement. The scientific ideas, experimental designs, and data analyses are the original work of the authors, who take full responsibility for the content of this paper.

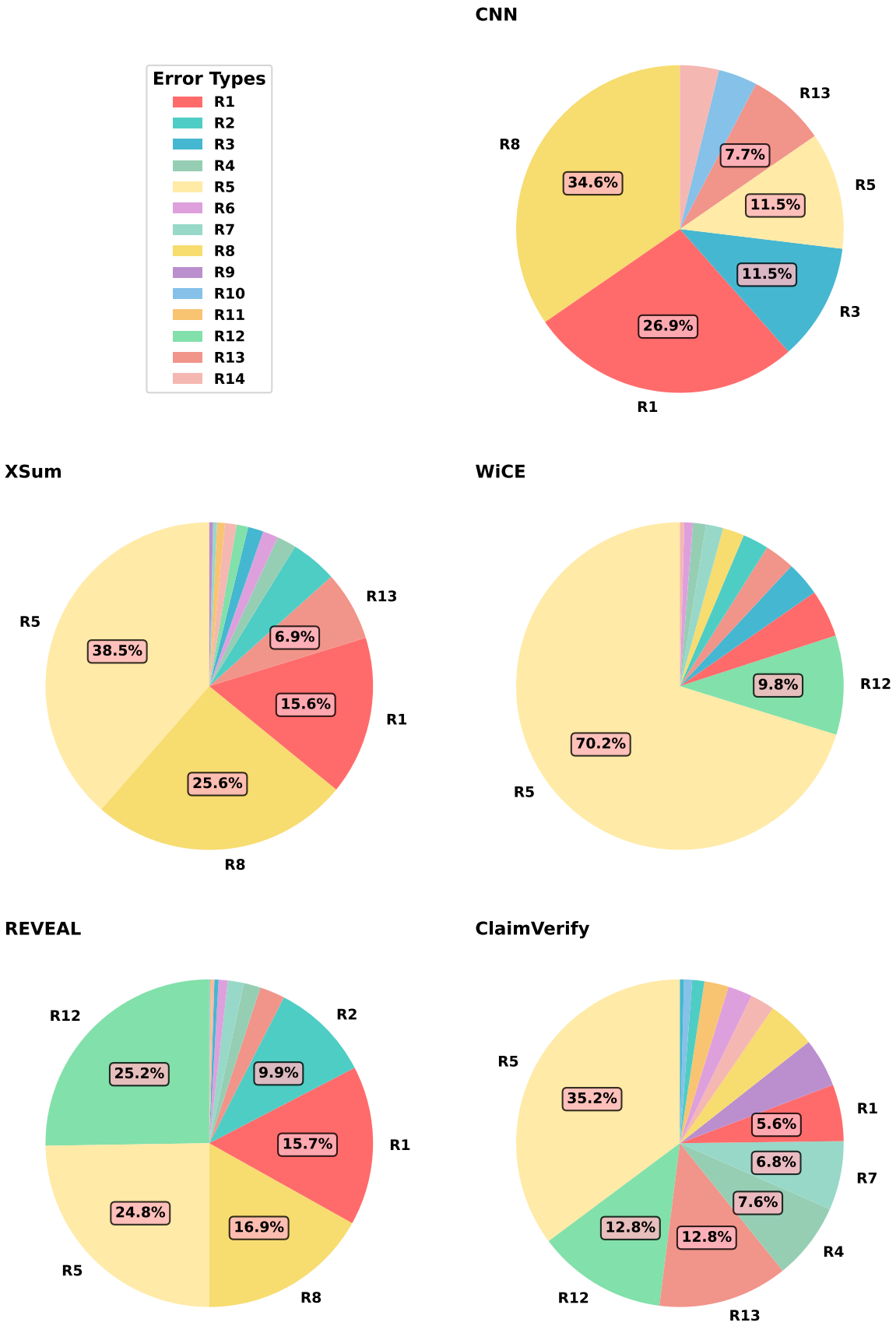

Figure 6: Per-dataset distribution of hallucination error types (continued).

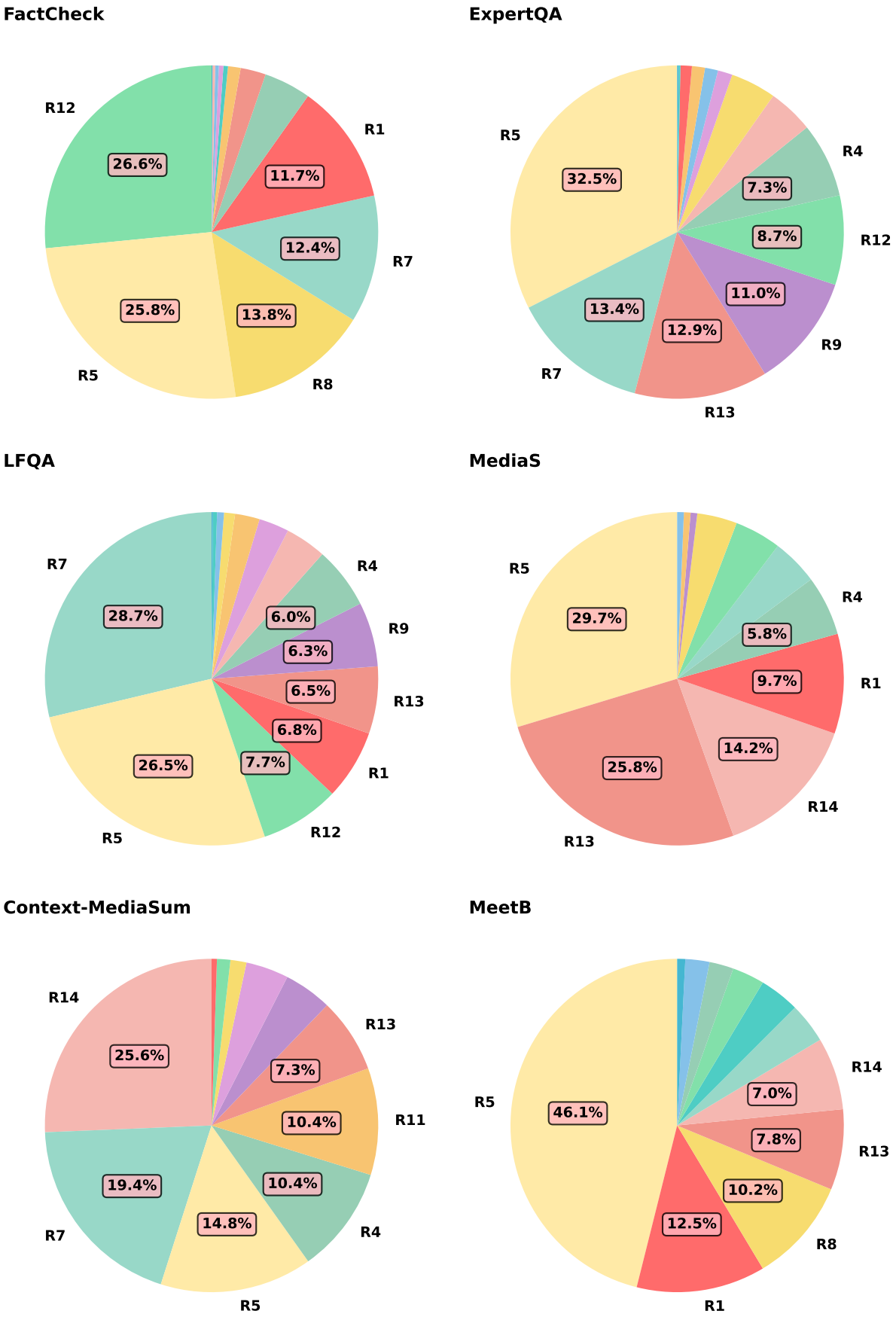

Figure 6: Per-dataset distribution of hallucination error types (continued).

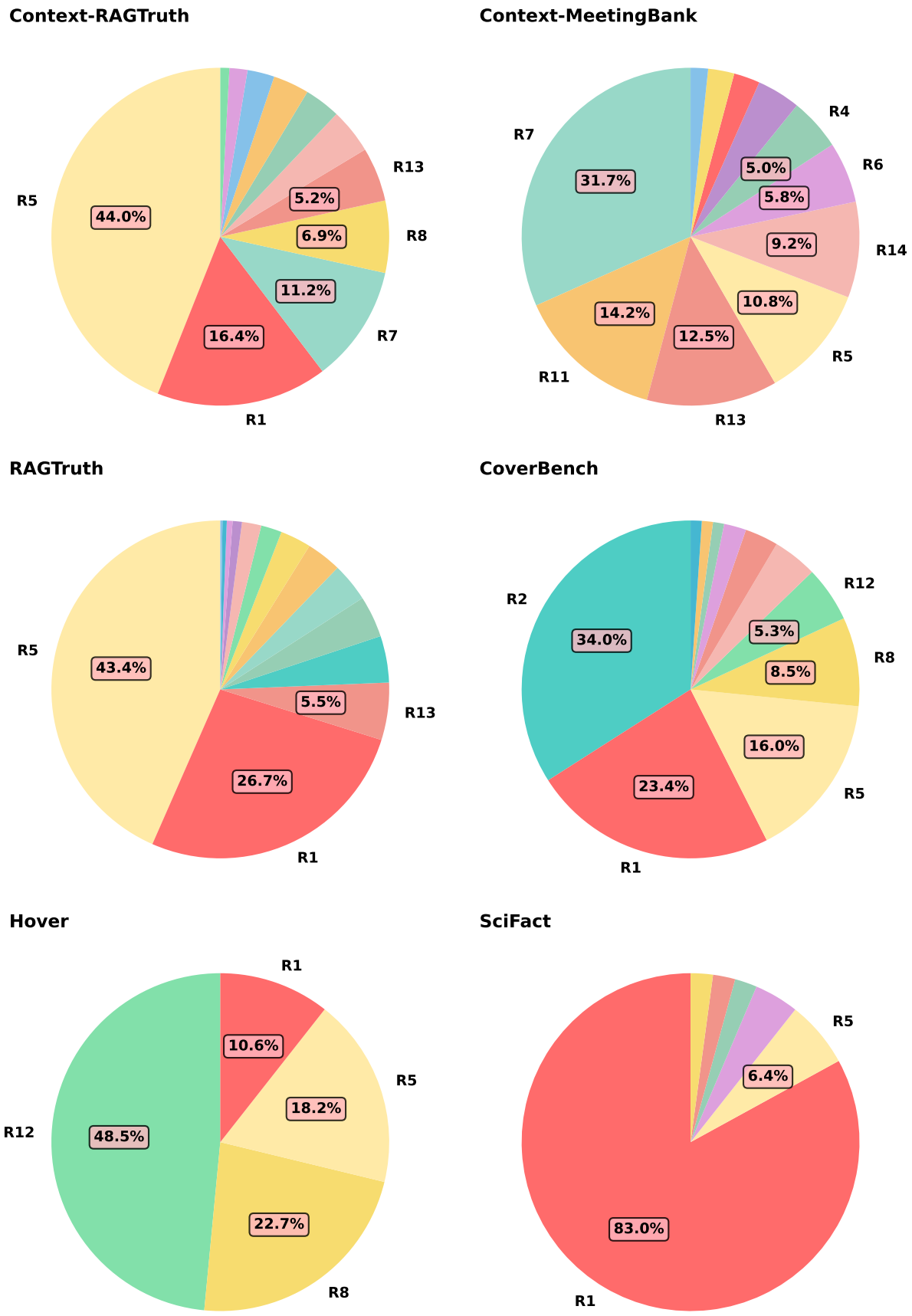

Figure 6: Per-dataset distribution of hallucination error types.

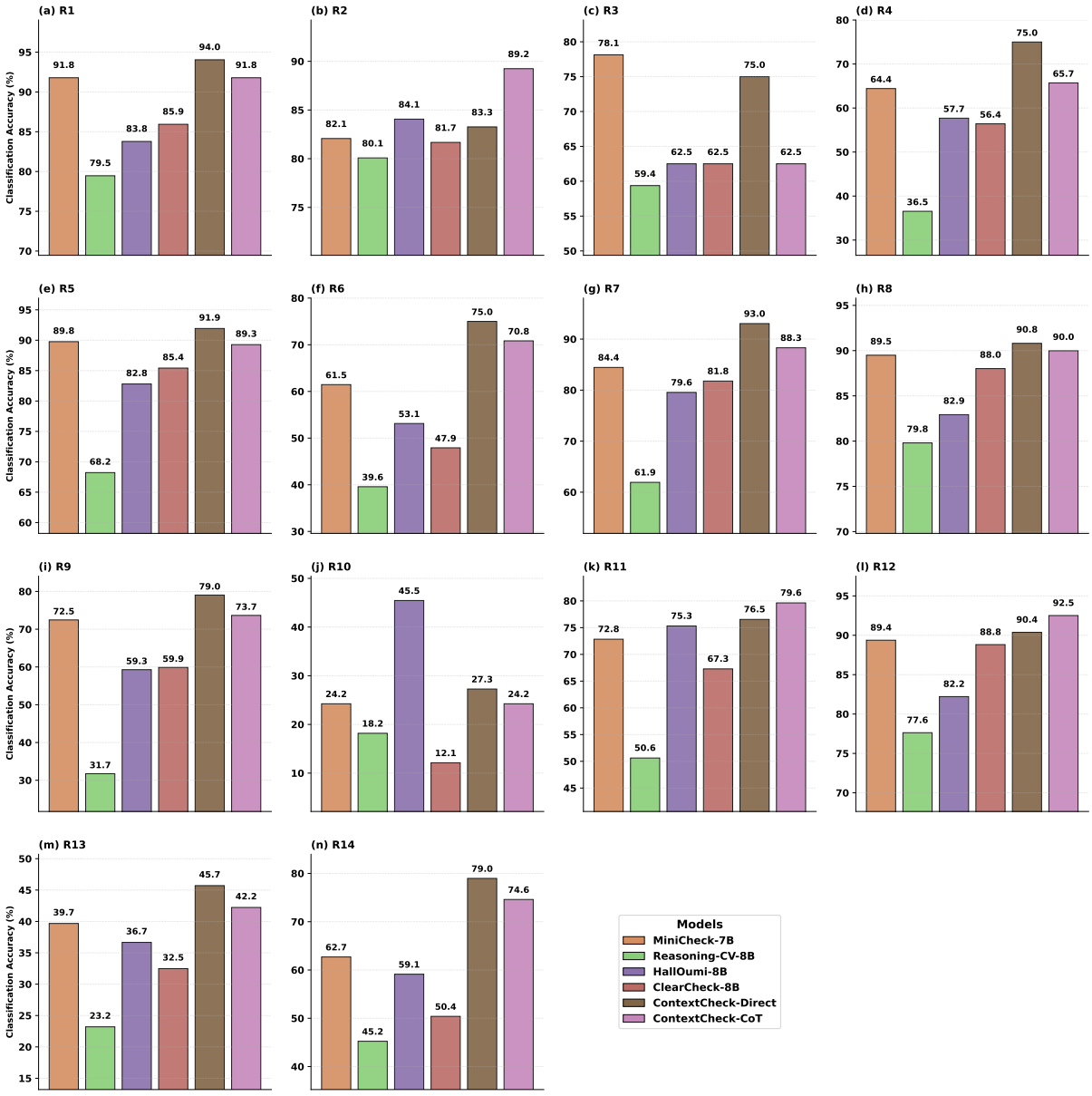

Figure 7: Fine-grained model performance across the 14 hallucination types. Each panel reports the per-hallucination-type recall of evaluated verifiers on a specific category (R1–R14), *i.e.*, the proportion of instances within that category that were correctly classified. This breakdown provides a granular comparison of model strengths and weaknesses, highlighting the consistently strong performance of the ContextCheck models.

J.1: Case Study: Debiased Score Example from Context_MediaSum

Grounding Document (Transcript Excerpt):
In New Jersey, the governor issued a statewide stay-at-home order and mandated the closure of all nonessential retail businesses. Across the river, New York Governor Andrew Cuomo announced that the Javits Center would be converted into a 1,000-bed temporary field hospital, one of four such sites planned statewide. Meanwhile, healthcare workers nationwide reported severe shortages of protective equipment, with many being asked to reuse items designed for single use. Nurses and doctors described working in "complete chaos," lacking adequate staff, masks, and ICU beds. Some hospitals instructed workers to reuse protective gear all day or save it for the next, despite the risks of cross-contamination. Frontline workers expressed fear and frustration, warning that these shortages left both patients and staff vulnerable as the crisis intensified.

Context:
- New York and New Jersey implemented significant measures to combat the coronavirus outbreak, with New Jersey's governor issuing a comprehensive stay-at-home order and mandating the closure of all non-essential retail businesses starting at 9 PM.
- In a parallel response, New York Governor Andrew Cuomo announced the conversion of the Javits Convention Center into a temporary field hospital capable of housing 1,000 beds, marking one of four emergency field hospital sites planned across the state to alleviate strain on the healthcare system.

Target Sentence (v1, pre-disambiguation):
These state-level responses reflected the growing urgency of the crisis, though healthcare workers across both states reported severe shortages of essential medical supplies and protective equipment, with many forced to reuse single-use items.

Target Sentence (v2, disambiguated):
The measures by New York and New Jersey reflected the growing urgency of the crisis, and healthcare workers across New York and New Jersey reported severe shortages of medical supplies and PPE, with many forced to reuse single-use items.

Ground Truth:
Not Attributable

Pre-disambiguation Model Output:

```
[Extraction] According to the document, healthcare workers across both states reported severe
shortages of essential medical supplies and protective equipment.
[Extraction] The document also mentions that some healthcare workers are being asked to reuse
single-use items.
[Extraction] However, the document does not explicitly state that these state-level responses
reflected the growing urgency of the crisis.
[Inference] Although the document does mention the growing urgency of the crisis, it is not
directly related to the state-level responses.
[Conclusion] So, the statement is [Not Attributable].
```

Post-disambiguation Model Output:

```
[Extraction] According to the document, New York Governor Andrew Cuomo said that the Javits
Center will become a field hospital for New Yorkers.
[Extraction] The document also mentions that New Jersey's governor issued a statewide
stay-at-home order and told all nonessential retail businesses to close.
[Extraction] Additionally, the document states that healthcare workers across the country
reported severe shortages of medical supplies and PPE, with many forced to reuse single-use
items.
[Inference] This implies that the measures by New York and New Jersey reflected the growing
urgency of the crisis.
[Conclusion] So, the statement is [Attributable].
```

J.2: Case Study: Benefit of Disambiguation from Context_MeetingBank

Grounding Document (Transcript Excerpt):
Report from Public Works, Parks, Recreation and Marine. Recommendation to execute a Third Amendment to contract with Plenary Properties Long Beach to modify the conditions of conveyance of the Mid-Block site, extend the project long-stop date, and include additions to Lincoln Park. Enter into an agreement with the Harbor Department to accept and expend grant funding in an amount not to exceed $981,280 for the Lincoln Park Landscaping Project. Staff noted that the redevelopment plan envisioned increasing housing on the Mid-Block site from roughly 300 to 580 units (with 10% affordable), creating a 100-unit teachers' village (half affordable), and adding community benefits such as waived deposits, free Long Beach transit, and access to community space. The project also includes enhancements to Lincoln

Park, funded partly through impact fees, and aims to create a more vibrant community space with added housing and amenities.

Context (preceding sentence):

- The Lincoln Park Landscaping Project aims to enhance the park through the acceptance and expenditure of grant funding, not exceeding $981,280, alongside impact fees.

Target Sentence (v1, pre-disambiguation):
This initiative is part of a broader effort to create a vibrant community space, which includes plans for additional housing and amenities in the surrounding area.

Target Sentence (v2, disambiguated):
The Lincoln Park Landscaping Project will enhance the park using grant funding up to $981,280 along with impact fees, and is part of a broader effort that includes added housing and amenities in the surrounding area.

Ground Truth:
Attributable

Pre-disambiguation Model Output (Incorrect):

```
## Reasoning Step 1: The context discusses a motion related to a Third Amendment to a contract
with Plenary Properties to modify conditions for the conveyance of the Mid-Block site, which
involves extending the project's long-stop date and including additions to Lincoln Park.
## Reasoning Step 2: The context mentions plans to increase the amount of housing on the
Mid-Block site from approximately 300 units to 580 units, with 10% of these units being
affordable and a grocer being brought in to downtown.
## Reasoning Step 3: The context also mentions the creation of a 100-unit teachers' village,
with half of these units being affordable, and additional benefits such as no deposit payments,
free Long Beach transit, and access to community space.
## Reasoning Step 4: The context indicates that the project includes enhancements to Lincoln
Park, which suggests an effort to create a more vibrant community space.
## Reasoning Step 5: The context does not explicitly state that the initiative is part of
a broader effort to create a vibrant community space, but it does imply that the project
includes various amenities and developments that could contribute to a more vibrant community.
## Reasoning Step 6: The mention of a teachers' village and additional housing suggests an
effort to enhance the community by providing more living spaces and amenities.
## Reasoning Step 7: The context does not provide explicit information about other amenities
in the surrounding area, but the enhancements to Lincoln Park and the increased housing could
be part of a broader effort to create a more vibrant community space.
## Reasoning Step 8: The claim is partially supported by the context, as there are plans for
additional housing and amenities, which could contribute to a vibrant community space, but
the broader effort is not explicitly detailed.
## Reasoning Step 9: The context does not provide sufficient evidence to fully support the
claim that the initiative is part of a broader effort to create a vibrant community space,
as it lacks explicit details on other amenities and spaces outside of Lincoln Park and the
Mid-Block site.
## Reasoning Step 10: The claim is partially supported by the context, as there are elements
that align with the idea of creating a vibrant community space, but the broader effort is not
fully detailed.
{refute}
```

Post-disambiguation Model Output (Correct):

```
## Reasoning Step 1: The claim states that the Lincoln Park Landscaping Project will enhance the
park using grant funding up to $981,280. The context mentions that there is a recommendation
to accept and expend grant funding in the amount not to exceed $981,280 for the Lincoln Park
Landscaping Project. This supports the first part of the claim.
## Reasoning Step 2: The claim also mentions that the project will use impact fees. The
context states that some of the dollars for the project are mostly getting from impact fees,
which supports this part of the claim.
## Reasoning Step 3: The claim suggests that the Lincoln Park Landscaping Project is part of a
broader effort that includes added housing and amenities in the surrounding area. The context
discusses the development of the Mid-Block site, which includes plans for increased housing, a
teachers' village, and enhancements to Lincoln Park, indicating that the landscaping project
is part of a broader effort to develop the area.
## Reasoning Step 4: The context mentions the addition of 580 units of housing, a teachers'
village, and other amenities, which aligns with the claim that the project is part of a
broader effort to enhance the area with added housing and amenities.
## Reasoning Step 5: The context does not explicitly state that the Lincoln Park Landscaping
Project is directly linked to the specific amount of $981,280 for the entire project, but it
does mention that this amount is for the Lincoln Park Landscaping Project, suggesting it is
a significant component of the broader effort.
## Reasoning Step 6: The context does not provide explicit details on the exact amount of
impact fees used for the Lincoln Park Landscaping Project, but it does mention that impact
fees are part of the funding sources, which supports the claim's implication of additional
funding sources beyond the grant.
## Reasoning Step 7: The context does not contradict any part of the claim regarding the use
of grant funding, impact fees, and the broader effort to enhance the area with housing and
amenities.
Based on the context provided, the claim is supported by the information available. The
Lincoln Park Landscaping Project is indeed part of a broader effort to enhance the area with
added housing and amenities, and it is funded in part by grant funding up to $981,280 and
impact fees.
{support}
```

J.3: Case Study: ContextCheck Self-Disambiguation from Context_RAGTruth

Grounding Document (Transcript Excerpt):
Dolphins Trade Target: Chase Young
The Miami Dolphins are coming off a controversial loss to the Philadelphia Eagles on Sunday night which saw two early-season Super Bowl contenders go head-to-head on national television.
The Eagles defense (and the refs) stood tall in taking down the Dolphins, who saw a much improved play from their defense, but as the trade deadline approaches all NFL journalists, sports talk podcasts, and shows airing on national TV and every sports streaming platform are focused on which players will be switching teams.
The Dolphins will have an influx of talent coming back to action in the coming weeks, such as Xavien Howard, Jalen Ramsey, Connor Williams, Terron Armstead, and De'Von Achane.
However, there are still some needs along the offensive line, and in areas of the defense, some of which can be addressed at the trade deadline.
Chase Young
Probably not the first name that comes to mind when thinking of needs for the Dolphins, but the struggles of the defensive line in getting consistent pressure with their front four has been a talking point all season.
CBS Sports' Josina Anderson reports that "several identifiable teams" are prepared to pull off a trade for Young or Montez Sweat from the Commanders, and while it may seem unlikely that Young will land in Miami, it's certainly something that should be, at the very least, explored.
First, Emmanuel Ogbah has been ineffective this season, collecting just 2.5 sacks on the season, a far cry from the 9 sacks he had in both the 2020 and 2021 seasons (prior to injury).
To be fair, the likes of Christian Wilkins have not played to their expected level, and the line as a whole has struggled.
Trading for Young, who is five years younger and has 5 sacks this season, would be a massive upgrade for the Dolphins.
It must be mentioned that trading for Young would lead to an immediate need to sign him to a long-term extension.
Miami is certainly going to have some serious questions and decisions to make regarding their cap situation in the future, but if they were willing to sign Indianapolis Colts running back Jonathan Taylor to a lucrative contract, they certainly could do so with Young, and perhaps it may be a bit cheaper.
Bringing Young to a rush that includes Wilkins, Zach Sieler, Jaelen Phillips, and Bradley Chubb would certainly boost the line's potency as they prepare for a playoff push.
Miami has struggled generating sacks this season, so adding Young would go a long way into boosting that number, and helping other areas of the defense as well.

Context:

- The Miami Dolphins are reportedly considering a trade for Chase Young or Montez Sweat from the Commanders.

- CBS Sports reports that several teams are prepared to make a move for either player.
- Despite the Dolphins' defensive line struggling with consistent pressure this season, a trade for Young could prove beneficial.

Target Sentence:
The five-year-younger player has achieved five sacks this season, compared to Emmanuel Ogbah's 2.5.

Ground Truth:
Attributable

Model Output (ContextCheck, with Disambiguation):

```
[Disambiguation]
Need-Rewrite: Yes
Action: "The five-year-younger player" was resolved to "Chase Young" using the preceding
context and GD.
Claim: Chase Young is five years younger than Emmanuel Ogbah and has achieved five sacks this
season, compared to Emmanuel Ogbah's 2.5.
[Analysis]
[Planning]
- Step 1: Extract information about Chase Young's age relative to Emmanuel Ogbah.
- Step 2: Extract information about Chase Young's sacks this season.
- Step 3: Extract information about Emmanuel Ogbah's sacks this season.
- Step 4: Compare the claim to the document for accuracy.
[Execution]
- Executing Step 1:
[Extraction] The GD states: "Trading for Young, who is five years younger and has 5 sacks
this season, would be a massive upgrade for the Dolphins."
[Inference] This directly supports that Young is five years younger than Ogbah.
- Executing Step 2:
[Extraction] The GD states Young "has 5 sacks this season."
- Executing Step 3:
[Extraction] The GD states: "Emmanuel Ogbah has been ineffective this season, collecting just
2.5 sacks on the season, a far cry from the 9 sacks he had in both the 2020 and 2021 seasons
(prior to injury)."
- Executing Step 4:
[Inference] The claim accurately reflects the information in the document: Young is five
years younger and has five sacks, while Ogbah has 2.5 sacks.
[Decision]
Attributable
```

J.4: Case Study: CoverBench Numerical Reasoning

Grounding Document:

> 36 | BHGE 2017 Form 10-K, Liquidity and Capital Resources. Our objective in financing our business is to maintain sufficient liquidity, adequate financial resources, and financial flexibility in order to fund the requirements of our business.
>
> At December 31, 2017, we had cash and equivalents of $7.0 billion compared to $981 million of cash and equivalents at December 31, 2016. Cash and equivalents includes $997 million of cash held on behalf of GE at December 31, 2017. At December 31, 2017, approximately $3.2 billion of our cash and equivalents was held by foreign subsidiaries compared to approximately $878 million at December 31, 2016.
>
> A substantial portion of the cash held by foreign subsidiaries at December 31, 2017 has been reinvested in active non-U.S. business operations. At December 31, 2017, our intent is, among other things, to use this cash to fund the operations of our foreign subsidiaries, and we have not changed our indefinite reinvestment decision as a result of U.S. tax reform but will reassess this during the course of 2018. If we decide at a later date to repatriate those funds to the U.S., we may be required to provide taxes on certain of those funds, however, due to the enactment of U.S. tax reform, repatriations of foreign earnings will generally be free of U.S. federal tax but may incur other taxes such as withholding or state taxes.
>
> On July 3, 2017, in connection with the transactions, BHGE LLC entered into a new five-year $3 billion committed unsecured revolving credit facility (2017 Credit Agreement) with commercial banks maturing in July 2022. As of December 31, 2017, there were no borrowings under the 2017 Credit Agreement. On November 3, 2017, BHGE LLC entered into a commercial paper program under which it may issue from time to time up to $3 billion in commercial paper with maturities of no more than 397 days. At December 31, 2017, there were no borrowings outstanding under the commercial paper program. The maximum combined borrowing at any time under both the 2017 Credit Agreement and the commercial paper program is $3 billion.

On November 6, 2017, we announced that our board of directors authorized BHGE LLC to repurchase up to $3 billion of its common units from the company and GE. The proceeds of such repurchase that are distributed to the company will be used to repurchase Class A shares of the company on the open market or in privately negotiated transactions. On December 15, 2017, we filed a shelf registration statement on Form S-3 with the SEC to give us the ability to sell up to $3 billion in debt securities in amounts to be determined at the time of an offering. Any such offering, if it does occur, may happen in one or more transactions. The specific terms of any securities to be sold will be described in supplemental filings with the SEC. The registration statement will expire in 2020.

During the year ended December 31, 2017, we used cash to fund a variety of activities including certain working capital needs and restructuring costs, capital expenditures, business acquisitions, the payment of dividends and share repurchases. We believe that cash on hand, cash flows generated from operations and the available credit facility will provide sufficient liquidity to manage our global cash needs.

Cash flows provided by (used in) each type of activity were as follows for the years ended December 31. Our largest source of operating cash is payments from customers, of which the largest component is collecting cash related to product or services sales including advance payments or progress collections for work to be performed. The primary use of operating cash is to pay our suppliers, employees, tax authorities and others for a wide range of material and services.

Cash Flows (in millions):

	2017	2016	2015
Operating activities	-799	262	1277
Investing activities	-4130	-472	-466
Financing activities	10919	-102	-515

Target Sentence:
The net change in cash during 2016 was -312.

Ground Truth:
Attributable

Model Outputs:
ClearCheck-CoT

```
[Disambiguation]
Need-Rewrite: No
Action: None
Claim: The net change in cash during 2016 was -312.
[Analysis]
[Planning]
- Step 1: Extract the cash flow information for 2016 from the document.
- Step 2: Calculate the net change in cash for 2016 using the provided cash flow data.
- Step 3: Compare the calculated net change with the claimed amount of -312.
[Execution]
- Executing Step 1: [Extraction] The document provides a cash flow table showing cash flows for
operating, investing, and financing activities for the years ended December 31, 2017, 2016,
and 2015. For 2016, the cash flows are: - Operating activities: $262 - Investing activities:
-$472 - Financing activities: -$102
- Executing Step 2: [Inference] To find the net change in cash for 2016, I need to sum the
cash flows from each activity: $262 (operating) + (-$472) (investing) + (-$102) (financing)
= $262 - $472 - $102 = -$312.
- Executing Step 3: [Inference] The calculated net change of -$312 exactly matches the claim
of "-312."
[Decision] Attributable
```

MiniCheck-7B

```
Raw probability: 0.2939. Hallucination.
```

Reasoning-CV-8B

```
## Reasoning Step 1: The claim states that the net change in cash during 2016 was -312. To
verify this, we need to examine the cash and equivalents figures for 2016 and 2017.
## Reasoning Step 2: According to the context, at December 31, 2016, the company had cash and
equivalents of $981 million.
## Reasoning Step 3: At December 31, 2017, the company had cash and equivalents of $7.0
billion.
## Reasoning Step 4: To find the net change in cash, we subtract the 2016 amount from the
2017 amount: $7,000,000,000 - $981,000,000 = $6,019,000,000.
## Reasoning Step 5: The claim states a net change of -312, but the actual net change from
2016 to 2017 is $6,019,000,000, which is significantly different from -312.
## Reasoning Step 6: The claim is incorrect as the actual net change in cash from 2016 to
2017 is not -312, but rather a positive $6,019,000,000.
{refute}
```

HallOumi-8B

```
<|r1|><The net change in cash during 2016 was -312.><|subclaims|><The net change in cash is
calculated by subtracting the cash at the end of the year from the cash at the beginning of
the year.><The cash at the end of 2016 was $981 million.><The cash at the beginning of 2016
was $7.0 billion.><end||subclaims><|cite|><|s1|><end||cite><|unsupported|><|explain|><The
document mentions that the cash and equivalents at December 31, 2016, was $981 million
and at December 31, 2017, was $7.0 billion. However, it does not provide the exact amount of
cash at the beginning of 2016. The net change in cash can be calculated as $7.0 billion -
$981 million = $6.019 billion, which is not equal to -312.><end||explain><end||r>
```

ClearCheck

```
[Extraction] According to the document, the cash and equivalents at December 31, 2016, was
$981 million.
[Extraction] The cash and equivalents at December 31, 2017, was $7.0 billion.
[Inference] The net change in cash during 2016 can be calculated by subtracting the cash and
equivalents at December 31, 2016, from the cash and equivalents at December 31, 2017.
[Inference] The net change in cash during 2016 is $7.0 billion - $981 million = $6.019 billion.
[Conclusion] So, the statement is [Contradictory].
```

ContextCheck-Direct

```
Not Attributable
```

K Prompt Templates

K.1 Prompt Templates for Data Annotation

This section provides the complete set of prompts used for our data curation and annotation pipeline, as illustrated in Fig. 1. These prompts are designed to execute the key stages of our methodology: (1) generating challenging, context-dependent claims from source documents using topic-focused summarization (K.4); (2) analyzing these claims to filter for context-dependency (K.5); (3) annotating each claim for faithfulness with a structured chain-of-thought using a detailed set of judging principles (K.6); and (4) ensuring data quality through multi-agent adjudication (K.7) and fine-grained error classification (K.8).

K.1: Prompt Template for Document Topic Categorization

You are an expert document analyst. Your task is to identify and categorize topics in the given document.

Task:
Please analyze the following document and identify exactly 3 topics discussed, categorizing each as either **Main Topic** or **Marginal Topic**.

Definitions:

- **Main Topic**: The primary focus or central subject matter that the document extensively discusses. These are the core themes that the document is fundamentally about.
- **Marginal Topic**: Information in the document that is not the main focus but is still part of the context. These topics are typically less prominent, provide additional context, background information, or support the main topics, but are not the primary focus.

Instructions:

1. Identify exactly 3 topics total from the document.
2. Ensure exactly 2 are **Main Topic** and 1 is **Marginal Topic**.
3. For each topic, provide:
   - A clear, specific topic name (2–8 words).
   - A brief description (1–2 sentences).
   - The category (`Main Topic` or `Marginal Topic`).

Document:
{document}

Required Response Format (JSON):

```
{
    "topics": [
        {
            "name": "Topic Name",
            "description": "Brief description of the topic",
            "category": "Main Topic" | "Marginal Topic"
        }
    ],
    "document_summary": "One sentence summary of the overall document"
}
```

Respond only with the JSON, no additional text.

K.2: Prompt Template: Summary (Short)

Please generate a concise summary of the following document focusing on the specified topics.

Requirements:

- Write exactly 1–3 sentences (50–200 words).
- Focus on the following topics: {topics}.
- IMPORTANT: Aim for strong coherence between sentences — use pronouns, transitional phrases, and logical connections.
- Sentences should build upon the previous one with clear relationships (causal, temporal, or referential).
- Ensure the summary flows naturally with evident inter-sentence dependencies.
- Do NOT include filler words, acknowledgments, or meaningless phrases (e.g., "Sure, I understand", "Certainly", "Of course").
- Start directly with substantive content and focus on factual information only.

Document:
{document}

Topics to focus on:
{topic_list}

Summary:

K.3: Prompt Template: Summary (Medium)

Please generate a comprehensive summary of the following document focusing on the specified topics.

Requirements:

- Write exactly 4–5 sentences (250–400 words).
- Focus on the following topics: {topics}.
- IMPORTANT: Aim for strong coherence between sentences — use pronouns, transitional phrases, and logical connections.
- Sentences should reference or build upon previous content with clear relationships.
- Create evident dependencies between sentences for stronger coherence.
- Do NOT include filler words, acknowledgments, or meaningless phrases (e.g., "Sure, I understand", "Certainly", "Of course").
- Start directly with substantive content and focus on factual information only.

Document:
{document}

Topics to focus on:
{topic_list}

Summary:

K.4: Prompt Template: Summary (Long)

Please generate a detailed summary of the following document focusing on the specified topics.

Requirements:

- Write exactly 6–7 sentences (450–600 words).
- Focus on the following topics: {topics}.
- IMPORTANT: Aim for strong coherence between sentences — use pronouns, transitional phrases, and logical connections.
- Try to create a unified narrative with clear inter-sentence relationships and dependencies.
- Aim to have sentences logically follow from the previous ones with evident connections.
- Develop topics progressively through coherent reasoning chains.
- Do NOT include filler words, acknowledgments, or meaningless phrases (e.g., "Sure, I understand", "Certainly", "Of course", "Let me summarize").
- Start directly with substantive content and focus on factual information only.

Document:
{document}

Topics to focus on:
{topic_list}

Summary:

K.5: Prompt Template: Context Dependency Analyzer

You are an expert context dependency analyzer for hallucination detection.

Task:
Decide whether each sentence in a response can be fully inferred from the grounding document alone, or whether it requires previous sentences in the response (context) to be properly understood and verified. Judge only context dependency, not factual correctness.

Key Definitions:

- Context = only the earlier sentences in the response.
- Grounding document = the fact source. It is not context.
- If a sentence can be fully understood and verified against the grounding document without earlier response sentences → Not context-dependent.
- If a sentence contains references that cannot be uniquely resolved from the grounding document alone → Context-dependent.

Rules for Context Dependency:
1. Uses pronouns/references (he, she, it, this, these, such, that achievement, former/latter, previous one, the results) that cannot be uniquely resolved from the grounding document.
2. Refers to entities/events introduced only in earlier response sentences.
3. Makes comparisons/summaries that require earlier sentences.
4. Uses ellipsis or implicit carryover from earlier content.
5. If any part of the sentence needs earlier response sentences, mark the whole sentence as context-dependent.

A sentence is not context-dependent if:

- All entities and claims are self-contained.
- Pronouns can be resolved unambiguously using only the grounding document.
- Do not confuse "story continuation" with context dependency. A sentence may flow narratively, but if it can still be checked independently, it is not context-dependent.
- Example: GD: "In 2019, Council formed the 14th Street GID. Resolution 1268 approved the 2021 budget." Resp: "[s2] Resolution 1268 formally established the GID." → This claim is contradicted by GD, but it is still self-contained and does not need earlier response sentences.

Output Format:
Analysis: For each sentence, explain why it is / isn't context-dependent. Result: Either `No` (if no sentences depend on context) or a comma-separated list like `s1,s2` (no spaces).

Quick Checklist:

- Ambiguous pronoun in GD? → Context-dependent
- Refers to entity/event introduced only in earlier response? → Context-dependent
- Uses "this/these/previous/former/latter/such/that achievement/results"? → usually Context-dependent
- Otherwise self-contained? → Not context-dependent

Example
Grounding Document:
Dr. James Wilson is a cardiologist at City Hospital. Dr. Robert Chen is a neurologist at the same hospital. Wilson published a groundbreaking study on heart disease treatment. Chen has been researching brain disorders for 15 years.
Response:
• [s1] Dr. Wilson published an important medical study recently.
• [s2] He has made significant contributions to cardiac medicine.
• [s3] Dr. Chen focuses on neurological research.
Analysis:
• [s1] Not context-dependent — "Dr. Wilson" is clearly identified and the study is mentioned in the grounding document.
• [s2] Context-dependent — "He" is ambiguous because both Dr. Wilson and Dr. Chen are mentioned in the grounding document. We need the previous sentence [s1] to know that "He" refers to Dr. Wilson.
• [s3] Not context-dependent — "Dr. Chen" is clearly identified and his research focus is mentioned in the grounding document.
Result:
s2

Additional Example
Grounding Document:
The forensic team found a single blood trace at the crime scene. The trace contained DNA evidence that matched the suspect's profile. Laboratory analysis confirmed the trace was fresh, deposited within 24 hours of discovery.
Response:
• [s1] Investigators discovered important physical evidence at the scene.
• [s2] The trace provided crucial DNA information for identification.
• [s3] Laboratory tests confirmed its recent origin.
Analysis:
• [s1] Not context-dependent — "physical evidence" can be inferred from the grounding document mentioning the blood trace.
• [s2] Not context-dependent — "The trace" can be uniquely resolved from the grounding document (only one trace mentioned), and DNA information is explicitly stated.
• [s3] Not context-dependent — "its recent origin" refers back to the trace's freshness mentioned in the grounding document.
Result:
No

Now, begin the task.

Grounding Document: {GROUNDING_DOCUMENT}

Response: {SENTENCE_BULLETS}

K.6: Prompt Template: Hallucination Label Annotation

You are an expert hallucination detector.

Instructions:

1. Your task is to determine if a Statement is attributable to the Document. You must base your decision ONLY on the information within the Document.

2. When multiple Statements are provided, evaluate each one independently. The judgment for one Statement must not influence the judgment for another.

3. Your answer must be one of three labels:

   - `[Attributable]`: All key information in the Statement is supported by or directly inferable from the Document, with no conflicting information.
   - `[Not Attributable]`: Any key part of the Statement is not mentioned or supported by the Document.
   - `[Contradicted]`: Any key part of the Statement is clearly contradicted by the Document.

4. Statements are identified by `<SOS>...<EOS>` markers. Content before `<SOS>` is context for disambiguation only (e.g., pronouns, tense, entity resolution, comparative references, time/condition anchoring, etc.) and NEVER factual evidence. Disambiguation must be performed using only the context before `<SOS>`. If resolving ambiguity (such as pronouns or references) requires information from the Grounding Document, do not resolve it in the `[Disambiguation]` section. Instead, address it in the `[Analysis]` section, where you may use the Grounding Document for inference.

5. **External Knowledge Guidelines:**

   - **Permitted ("Decoding" Knowledge):** You may use basic, universally known knowledge that helps interpret the meaning of the text without adding new factual content. This includes knowledge necessary for understanding language itself (synonyms, abbreviations, basic definitions) and performing simple operations on information explicitly present in the text (basic arithmetic, unit understanding, temporal/spatial relationships). Examples: 'largest city' for 'most populous', 'WWII' for World War II, calculating percentages from given numbers.
   - **Forbidden ("Enrichment" Knowledge):** A claim is `[Not Attributable]` if it adds a new, specific, verifiable fact not present in the text, even if that fact is true. Examples: Adding a first name ('Jennifer') to a surname ('Lynch'), or a chemical formula ('Fe_2O_3') to a chemical name ('iron(III) oxide').

6. Before your final decision, you must show your reasoning in an `[Analysis]` block. First, create a step-by-step `[Planning]` section that breaks down the verification process. Then, follow your plan in an `[Execution]` section, using interleaved `[Extraction]` and `[Inference]` tags for each step, as shown in the examples.

Key Judging Principles:

1. **Scope & Quantification:** Claims must match the Document's scope and quantifiers ('all', 'some', 'most'). No overgeneralization beyond stated populations or contexts, and no unsupported specification of concrete details when only general information is provided. Common violations: adding specific numbers ("many" $\rightarrow$ "157"), exact locations, precise times, or detailed procedures when the Document uses vague terms.

2. **Causality & Logic:** Causal links ('A caused B'), comparative relationships ('better than'), and conditional statements ('if-then') require explicit statement in the Document. Correlation is insufficient for causation. Avoid improperly combining separate statements into unsupported conclusions.

3. **Certainty & Modality:** Statement certainty, intensity, and frequency must match the Document's. Don't upgrade possibilities ('may') to facts ('is'), or change intensity ('slightly' $\rightarrow$ 'dramatically') or frequency ('rarely' $\rightarrow$ 'often').

4. **Context & Qualifiers:** Preserve all important qualifiers, conditions, limitations, and contextual constraints from the Document. Dropping qualifiers like 'some', 'in mice', 'under laboratory conditions' fundamentally alters meaning and scope.

5. **Subjective & Normative Claims:** Only attributable if the Document itself makes that specific judgment. Do not add evaluative opinions ('better/elegant'), normative prescriptions ('should/must'), or explanatory mechanisms/motivations not described in the Document. This includes invented reasons, purposes, step-by-step processes, or "because" explanations not explicitly stated.

6. **Temporal & Spatial Consistency:** Timeline, sequence, location, and order must align with the Document. Events must occur in the correct temporal sequence and spatial context as described.

7. **Entity Integrity:** All entities, attributes, measurements, and their properties must be precisely as described in the Document. Mismatches are `[Contradicted]`; absences are `[Not Attributable]`. Claims requiring external knowledge beyond basic language understanding are `[Not Attributable]`.

Format and Examples:

You must follow this format and reasoning style strictly. Each Example below is a separate, independent task.

Example 1
Grounding Document (GD):
A preliminary report from our security team suggests the device may be vulnerable to the new cyber-attack vector. Further testing is required.
Target Sentences:
- <SOS>The device is definitively vulnerable to the new cyber-attack.<EOS>
Answer:

```
=====SENTENCE_1=====

[Disambiguation]
Need-Rewrite: No
Action: None
Claim: The device is definitively vulnerable to the new cyber-attack.

[Analysis]
[Planning]
- Step 1: Extract the document's statement on the device's vulnerability.
- Step 2: Compare the certainty level of the document's statement with the claim's
  certainty ("definitively vulnerable").
[Execution]
- Executing Step 1:
  [Extraction] The GD states the device "may be vulnerable" according to a
  "preliminary report."
- Executing Step 2:
  [Inference] The Statement's claim of being "definitively vulnerable" is a
  statement of certainty. This directly conflicts with the Document's language of
  possibility ("may be"). This violates Key Judging Principles #3 (Certainty &
  Modality).

[Decision]
Contradicted
```

Example 2
Grounding Document (GD):
The AlphaGo program defeated Lee Sedol 4-1 in a five-game match. The program's success was attributed to its deep neural networks. Its successor, AlphaZero, was even more powerful.
Target Sentences:
- Lee Sedol lost the match against AlphaGo. <SOS>Its successor was more powerful than it was.<EOS>
- The program used advanced algorithms for this victory. <SOS>This approach was based on traditional rule-based systems.<EOS>
Answer:

```
=====SENTENCE_1=====

[Disambiguation]
Need-Rewrite: Yes
Action: "Its" was resolved to "AlphaGo's" and "it" was resolved to "AlphaGo"
using the preceding sentence and GD context.
Claim: AlphaGo's successor was more powerful than AlphaGo was.

[Analysis]
[Planning]
- Step 1: Identify AlphaGo's successor from the document.
- Step 2: Verify the document's comparison of power between the successor and AlphaGo.
[Execution]
- Executing Step 1:
  [Extraction] The GD states, "Its successor, AlphaZero, was even more powerful."
- Executing Step 2:
  [Inference] The document directly states that the successor (AlphaZero) was
  "even more powerful," which supports the rewritten claim.

[Decision]
Attributable

=====SENTENCE_2=====
```

```
[Disambiguation]
Need-Rewrite: Yes
Action: "This approach" was resolved to "the advanced algorithms approach" using
the preceding sentence context.
Claim: The advanced algorithms approach was based on traditional rule-based systems.

[Analysis]
[Planning]
- Step 1: Identify what approach is being referenced in the document.
- Step 2: Check what the document states about the technical foundation of this approach.
- Step 3: Compare the document's description with the claim about rule-based systems.
[Execution]
- Executing Step 1:
  [Extraction] The GD states "The program's success was attributed to its deep
  neural networks."
  [Inference] The preceding sentence refers to "advanced algorithms for this
  victory," which corresponds to the approach mentioned in the document.
- Executing Step 2:
  [Extraction] The document explicitly states that the program's success was
  attributed to "deep neural networks."
  [Inference] Deep neural networks are a machine learning approach, not traditional
  rule-based systems.
- Executing Step 3:
  [Inference] The claim states the approach "was based on traditional rule-based
  systems," while the document clearly states it was based on "deep neural
  networks." These are fundamentally different and contradictory technical
  approaches. This violates Key Judging Principles #7 (Entity Integrity) - the
  technical foundation is directly contradicted.

[Decision]
Contradicted
```

Example 3
Grounding Document (GD):
Microsoft announced a new AI assistant called Copilot. The assistant integrates with Office applications. Early tests showed 30% productivity improvements. The company plans to release it next quarter.
Target Sentences:
- Microsoft unveiled its latest AI assistant technology. <SOS>This revolutionary tool will transform workplace productivity forever.<EOS>
Answer:

```
=====SENTENCE_1=====

[Disambiguation]
Need-Rewrite: Yes
Action: "This revolutionary tool" was resolved to "Microsoft's new AI assistant tool"
using the preceding sentence context.
Claim: Microsoft's new AI assistant tool will transform workplace productivity forever.

[Analysis]
[Planning]
- Step 1: Identify the specific AI assistant technology mentioned in the document.
- Step 2: Examine the document for any claims about transforming workplace productivity.
- Step 3: Assess whether "revolutionary" and "forever" characterizations are supported.
[Execution]
- Executing Step 1:
  [Extraction] The GD identifies the AI assistant as "Copilot" and describes it as
  Microsoft's "new AI assistant."
  [Inference] The preceding sentence refers to "latest AI assistant technology,"
  which corresponds to Copilot.
- Executing Step 2:
  [Extraction] The document mentions "30% productivity improvements" from early
  tests, but contains no claims about "transforming workplace productivity."
  [Inference] The document provides specific, limited test results but does not make
  broad transformational claims.
- Executing Step 3:
  [Inference] The claim introduces evaluative language ("revolutionary") and temporal
  absolutes ("forever") not present in the document. The document describes factual
  test results, not revolutionary transformation. The claim also projects permanent,
```

```
  universal change ("forever") which is not supported by limited early tests. This
  violates Key Judging Principles #5 (Subjective & Normative Claims) and #3
  (Certainty & Modality).

[Decision]
Not Attributable
```

—

Now, begin the task.

Grounding Document (GD):
{GROUNDING_DOCUMENT}

Target Sentences:
{SENTENCE_BULLETS}

K.7: Prompt Template: Adjudicator Agent

You are a senior fact-checking expert and adjudicator. Your task is to analyze a reference document and a related claim, then render a final verdict on two conflicting judgments from different models.

Reference Document (Grounding Evidence):
{EVIDENCE}

Claim to Verify:
{CLAIM}

Two Conflicting Judgments:

- **Judgment 1 (Model A's Label):** {MODEL_A_LABEL}
- **Judgment 2 (Model B's Label):** {MODEL_B_LABEL}

Model Reasoning:
Model A's Reasoning:
{MODEL_A_REASONING}

Model B's Reasoning:
{MODEL_B_REASONING}

Your Task and Judging Standard:
Carefully analyze the Reference Document and the Claim. Based on the following standard of "reasonable attribution", determine which of the two judgments is most accurate and logical.

Judging Criteria:

1. **Reasonable Inference and Integration are PERMITTED:** Logical inference, information integration, and paraphrasing based on the source text are allowed.
2. **"Decoding" Knowledge is PERMITTED:** Basic, universally known external knowledge is allowed only for interpreting the text (e.g., synonyms, abbreviations, simple definitions).
3. **"Enrichment" Knowledge is FORBIDDEN:** The claim cannot add new, specific, verifiable facts not present in the source text, even if true.
4. **Context is CRUCIAL:** Use context to determine precise meaning and internal structure.
5. **Distinguish Contradiction from Nuance:** Small wording variations do not automatically invalidate a claim.
6. **Timelines and Causality:** Ensure sequence and causality align with or can be reasonably inferred from the text.

Required Response Format:

- **Most Reasonable Judgment:** `Model A's Label` | `Model B's Label`
- **Justification:** Provide a detailed, step-by-step explanation referencing criteria and specific text; also explain why the other judgment is less accurate.

K.8: Prompt Template: Hallucination Type Classifier

You are an expert hallucination type classifier.

Instructions:

1. Your task is to classify the type of hallucination present in a Statement when compared to the provided Document. The Statement is **ALREADY CONFIRMED** to contain hallucination — however, if after thorough comparison you find the Statement is actually fully supported by the Document, you may conclude that **no hallucination is present**. In such cases, explicitly state that no hallucination exists instead of forcing a category assignment.
2. When multiple Statements are provided, evaluate each one independently. The classification for one Statement must not influence the classification for another.
3. You must identify one or more hallucination types from the list below. **CLASSIFICATION STRATEGY**: Prioritize identifying the single MOST RELEVANT hallucination type that best captures the primary issue. If necessary, you may assign a second type, but no more than two types total. Focus on the dominant characteristic of the hallucination rather than listing all possible applicable categories. **IMPORTANT**: Always prioritize assigning the hallucination to existing categories R1–R14 first. Only use R15 as a last resort when the hallucination truly cannot be reasonably classified into any of the specific categories R1–R14, even with flexible interpretation.
4. **External Knowledge Guidelines:**

- **Permitted ("Decoding" Knowledge):** You may use basic, universally known knowledge that helps interpret the *meaning* of the text without adding new factual content. This includes knowledge necessary for understanding language itself (synonyms, abbreviations, basic definitions) and performing simple operations on information explicitly present in the text (basic arithmetic, unit understanding, temporal/spatial relationships). Examples: 'largest city' for 'most populous', 'WWII' for World War II, calculating percentages from given numbers.
- **Forbidden ("Enrichment" Knowledge):** Claims that add new, specific, verifiable facts not present in the text represent hallucinations, even if those facts are true. Examples: Adding a first name ('Jennifer') to a surname ('Lynch'), or a chemical formula (Fe_2O_3) to a chemical name ('iron(III) oxide').

5. Before your final decision, you must show your reasoning in an `[Analysis]` block. First, create a step-by-step `[Planning]` section that breaks down the classification process. Then, follow your plan in an `[Execution]` section, using interleaved `[Extraction]` and `[Inference]` tags for each step. In your final `[Classification]`, identify the primary (most relevant) hallucination type first, and only add a secondary type if it significantly contributes to understanding the hallucination. If you determine there is **no hallucination**, explicitly write: `No hallucination present`.

6. **Special attention:** Always check for subtle hallucination patterns such as:

   - Dropping or altering qualifiers (e.g., "some" → omitted; "in mice" → omitted → becomes "in humans").
   - Expanding scope/generalizing beyond the document's population.
   - Inferring causation, purpose, or mechanism not stated.
   - Improperly merging two separate statements into one conclusion.

   These errors often fall under **R13 (Context/qualifier/emphasis alteration)** or **R14 (Invalid logical inference/composition)**. Be vigilant for them.

7. **Important distinction – R6 vs R14:** These types are easily confused:

   - **R6 (Logical relationship errors):** Wrong relationship *type* between entities both mentioned in document (e.g., Doc: "A and B both increased" → Claim: "A caused B" – wrong causal relationship).
   - **R14 (Invalid logical inference/composition):** Improper *combination* of separate document information (e.g., Doc: "A happened. B happened." → Claim: "A and B are related" – wrong aggregation).

   Quick test: Specific entities with wrong relationships → R6. Separate facts wrongly combined → R14.

Hallucination Types:

- **R1. Direct contradiction/opposition** — Statement directly conflicts with, contradicts, negates, or opposes document facts. This includes factual contradictions, logical oppositions, and semantic negations (e.g., Doc: "There are 9." Claim: "There are 14."; Doc: "The project succeeded." Claim: "The project failed."; Doc: "X is possible." Claim: "X is impossible."; Doc: "Increased." Claim: "Decreased.").
- **R2. Quantitative/measurement errors** — Incorrect numerical values, units, dimensions, conversions, or other quantitative information (e.g., "0.15%" vs "1.5%"; wrong km↔mile conversion; "5 minutes" vs "5 hours"; "3 people" vs "30 people").
- **R3. Temporal/spatial/sequential inconsistencies** — Inconsistencies in time, space, location, sequence, or order. Timeline, spatial relationships, and sequential arrangements must be consistent with the document (e.g., Doc: "Ended in 2021." Claim: "Currently ongoing."; Doc: "Located in Paris." Claim: "Located in London."; Doc: "First A, then B." Claim: "First B, then A.").
- **R4. Scope expansion/over-generalization** — Extending claims beyond the document's stated scope, population, or domain. Claims must match the document's scope, quantifiers ('all', 'some', 'most'), and applicable contexts. Includes generalizing from specific cases or limited contexts (e.g., lab-only results → "all real-world scenarios"; "Some users reported..." → "Users generally experience..."; "In this study..." → "Research shows...").
- **R5. Unsupported specification/concretization** — Adding specific concrete details when the document only provides general or abstract information (e.g., "Invest in public transport." → "Added 12 subway lines"; "Many participants" → "157 participants"; "Improved performance" → "Increased by 23%").
- **R6. Logical relationship errors** — Incorrectly inferring causal, conditional, or comparative relationships not explicitly stated in the document. This includes correlation→causation, reversing cause-effect, adding unsupported if-then relationships, or creating comparisons not made in the document (e.g., Doc: "A and B both increased." Claim: "A caused B to increase"; Doc: "X works." Claim: "X works better than Y").
- **R7. Invented mechanism/process/explanation/motivation** — Adding explanatory mechanisms, causal processes, step-by-step procedures, intentions, motivations, or purposes not described in the document (e.g., Doc: "Sales increased." → Claim: "Sales increased due to improved customer service training"; Doc: "The CEO resigned." → Claim: "The CEO resigned to pursue other opportunities"; Doc: "The policy was implemented." → Claim: "The policy was implemented to reduce costs").

- **R8. Entity/attribute/source misidentification** — Wrong identification of objects, people, places, categories, sources, or their attributes. All entities, their properties, and attributions must be precisely as described in the document (e.g., Doc: "CEO John Smith" → Claim: "CFO John Smith"; Doc: "Red car" → Claim: "Blue car"; Doc: "University study" → Claim: "Government study"; Doc: "According to Dr. A" → Claim: "According to Dr. B").
- **R9. Normative/prescriptive claims added** — Adding normative judgments about what "should/must/ought to" be done, what would be "better/appropriate/necessary," or evaluative critiques about what was "wrong/insufficient/inadequate" — whether referring to past actions, current situations, or future recommendations — with no basis in the document.
- **R10. Modality/intensity/frequency alteration** — Incorrect changes to certainty levels, intensity, frequency, or degree. Statement modality must match the document's. Don't upgrade possibilities to facts or change intensity/frequency (e.g., "may" → "will"; "rarely" → "often"; "slightly increased" → "dramatically increased"; "some" → "most").
- **R11. Subjective value judgment added** — Adding purely evaluative opinions about quality, aesthetics, or desirability ("better/elegant/comfortable/beautiful") not expressed in document. Differs from R9 which focuses on normative claims about what should be done. Only attributable if the document itself makes that specific judgment.
- **R12. External knowledge enrichment — CRITICAL**: Claims that require "Enrichment" Knowledge (see External Knowledge Guidelines) — adding new, specific, verifiable facts not present in the document text, even if those facts are true in reality. This includes solving complex riddles that require outside knowledge, adding geographic/historical details not stated in the document, or introducing technical specifications not mentioned. Note: This differs from permitted "Decoding" Knowledge which only helps interpret existing text meaning.
- **R13. Context/qualifier/emphasis alteration** — Removing or changing important qualifiers, conditions, limitations, contextual constraints, or emphasis that alter the meaning or focus (e.g., "Only under low temperature, X works" → "X works"; "According to one study, Y is true" → "Y is true"; Doc emphasizes A over B → Claim emphasizes B over A; Doc: neutral tone → Claim: critical tone).
- **R14. Invalid logical inference/composition** — Drawing specific conclusions that don't logically follow from the given premises, or incorrectly combining separate pieces of information from the document in ways not supported by the text. This includes unsupported "therefore/so" conclusions and improper aggregation of distinct statements (e.g., Doc mentions A and B separately → Claim: "A and B work together"; Doc: "X happened. Y happened." → Claim: "X and Y are related").
- **R15. Other hallucination types — LAST RESORT ONLY**: Use only if none of R1–R14 apply.

Format and Examples:
You must follow this format and reasoning style strictly. Each Example below is a separate, independent task.

CLASSIFICATION OUTPUT REQUIREMENTS:

- Identify the PRIMARY hallucination type that best captures the main issue.
- Only add a SECONDARY type if it provides significant additional insight.
- Maximum of TWO types per classification.
- List the most relevant type first.
- If no hallucination is present, explicitly state: `No hallucination present`.

Example 1
Grounding Document (GD):
A preliminary report from our security team suggests the device may be vulnerable to the new cyber-attack vector. Further testing is required.
Target Sentence:
The device is definitively vulnerable to the new cyber-attack.
Answer:
Claim: The device is definitively vulnerable to the new cyber-attack.
[Analysis]
[Planning]

1. Extract the document's statement on the device's vulnerability and certainty level.
2. Compare the certainty level with the claim's certainty ("definitively vulnerable").
3. Identify the specific type of hallucination.

[Execution]
Executing Step 1: `[Extraction]` The GD states the device "may be vulnerable" according to a "preliminary report."
Executing Step 2: `[Inference]` The Statement claims the device is "definitively vulnerable" which expresses certainty, while the document expresses uncertainty with "may be vulnerable."

Executing Step 3: [Inference] This represents a change from uncertain ("may be") to certain ("definitively"), which is a modality change.
[Classification]
R10. Modality change

Example 2
Grounding Document (GD):

Feature	AlphaPhone 10	BetaMax Pro	Gamma Z
Battery Life	Up to 22 hours	Up to 28 hours	Up to 24 hours
Storage	128GB, 256GB	256GB, 512GB	128GB, 256GB
Price (USD)	$799	$1099	$749

Target Sentence:
The BetaMax Pro has the longest battery life and is the only model that offers a 1TB storage option.
Answer:
Claim: The BetaMax Pro has the longest battery life and is the only model that offers a 1TB storage option.
[Analysis]
[Planning]

1. Verify the battery life claim against the document.

2. Check the storage options claim against the document.

3. Identify hallucination types for any incorrect parts.

[Execution]
Executing Step 1: [Extraction] Battery life data shows AlphaPhone 10: 22 hours, BetaMax Pro: 28 hours, Gamma Z: 24 hours. [Inference] BetaMax Pro does have the longest battery life (28 hours), so this part is accurate.
Executing Step 2: [Extraction] Storage options for BetaMax Pro are listed as "256GB, 512GB" in the document. [Inference] The claim states BetaMax Pro offers "1TB storage option," but the document shows only 256GB and 512GB options.
Executing Step 3: [Inference] The 1TB storage claim directly contradicts the documented storage options, representing a direct factual contradiction.
[Classification]
R1. Direct contradiction

Example 3
Grounding Document (GD):
Studies show that people who exercise regularly tend to have lower rates of depression. The correlation was observed across multiple age groups.
Target Sentence:
Regular exercise prevents depression by releasing endorphins that directly combat depressive symptoms.
Answer:
Claim: Regular exercise prevents depression by releasing endorphins that directly combat depressive symptoms.
[Analysis]
[Planning]

1. Extract what the document states about exercise and depression.

2. Compare the document's correlation statement with the claim's causation statement.

3. Analyze the mechanistic explanation provided in the claim.

4. Identify hallucination types.

[Execution]
Executing Step 1: [Extraction] The document states there is a correlation between regular exercise and lower depression rates.
Executing Step 2: [Inference] The claim states exercise "prevents depression," implying causation, while the document only mentions correlation.
Executing Step 3: [Inference] The claim provides a specific mechanism ("releasing endorphins that directly combat depressive symptoms") that is not mentioned in the document.
Executing Step 4: [Inference] This represents both a correlation-to-causation leap and an invented mechanism. The primary issue is the logical relationship error (correlation→causation), while the secondary issue is the added mechanistic explanation.
[Classification]
R6. Logical relationship errors; R7. Invented mechanism/process

Example 4

Grounding Document (GD):
The study was conducted in laboratory conditions using mice. Results showed a 15% improvement in cognitive performance after treatment.
Target Sentence:
The treatment significantly improves cognitive performance in humans and should be the first-line therapy for cognitive decline.
Answer:
Claim: The treatment significantly improves cognitive performance in humans and should be the first-line therapy for cognitive decline.
[Analysis]
[Planning]

1. Extract the scope and subjects of the original study.

2. Compare the study scope with the claim's scope.

3. Analyze the prescription/recommendation in the claim.

4. Identify hallucination types.

[Execution]
Executing Step 1: `[Extraction]` The study was conducted on mice in laboratory conditions, showing 15% cognitive improvement.
Executing Step 2: `[Inference]` The claim extends results from mice to humans, which represents scope upscaling from a specific laboratory setting with mice to general human application.
Executing Step 3: `[Inference]` The claim adds a prescription that the treatment "should be the first-line therapy," which is not stated or implied in the document.
Executing Step 4: `[Inference]` This represents both over-generalization and addition of unsupported prescription. The primary issue is scope expansion from mice studies to human application, while the secondary issue is the normative claim about therapy choice.
[Classification]
R4. Scope expansion/over-generalization; R9. Normative/prescriptive claims added

Example 5
Grounding Document (GD):
Rates of interreligious marriage vary widely: In the United States, it is just under 50%, in the United Kingdom, around 53%; in France; around 30%, and in Australia and Mexico, as low as 10%.
Target Sentence:
The rate of interreligious marriage in the country from which the sandwich named after the successor of the National Rail originates is 53%.
Answer:
Claim: The rate of interreligious marriage in the country from which the sandwich named after the successor of the National Rail originates is 53%.
[Analysis]
[Planning]

1. Attempt to identify the country being referred to using only the provided document.

2. Analyze whether external knowledge is required to solve this reference.

3. Identify the hallucination type.

[Execution]
Executing Step 1: `[Extraction]` The document provides marriage rates for several countries but doesn't explain the complex reference to "the sandwich named after the successor of the National Rail."
Executing Step 2: `[Inference]` To solve this riddle and identify the country (the United Kingdom), one would need extensive external knowledge about British history, nobility, railway systems, and sandwich names. This information is not present in the document.
Executing Step 3: `[Inference]` According to the External Knowledge Guidelines, this represents forbidden "Enrichment" Knowledge – the claim adds verifiable facts (the identity of the country through the riddle solution) not present in the document text. While the 53% rate is indeed correct for the UK, identifying the UK requires external knowledge enrichment.
[Classification]
R12. External knowledge enrichment

Example 6
Grounding Document (GD):
The research team published their findings in a peer-reviewed journal. The study involved 100 participants and took place over 6 months.
Target Sentence:
The research team should have included a control group to make their study more credible.

Answer:
Claim: The research team should have included a control group to make their study more credible.
[Analysis]
[Planning]

1. Extract what the document states about the study methodology.

2. Analyze what the claim adds that's not in the document.

3. Systematically check if this fits any existing categories R1–R14.

4. Only if none apply, consider R15.

[Execution]
Executing Step 1: `[Extraction]` The document mentions the study was published, had 100 participants, and lasted 6 months, but provides no details about study design or methodology.
Executing Step 2: `[Inference]` The claim introduces a methodological critique about needing a control group, which is not mentioned or implied in the document.
Executing Step 3: `[Inference]` Checking existing categories: Could this be R9 (normative/prescriptive claims)? Yes! The claim states what the researchers "should have" done, which is a normative judgment about past actions being insufficient or inadequate. This fits the updated R9 definition that includes evaluative critiques about what was "wrong/insufficient/inadequate."
Executing Step 4: `[Inference]` R9 is the appropriate classification as it captures normative claims about what should have been done.
[Classification]
R9. Normative/prescriptive claims added

Example 7
Grounding Document (GD):
The new medication has shown positive results in some clinical trials with middle-aged patients, but further research is needed before drawing broad conclusions about its overall effectiveness.
Target Sentence:
The new medication has shown positive results in clinical trials.
Answer:
Claim: The new medication has shown positive results in clinical trials.
[Analysis]
[Planning]

1. Extract the exact scope and qualifiers from the document regarding the medication's effectiveness.

2. Compare the target sentence to see if it preserves those qualifiers.

3. Identify whether the removal of qualifiers significantly changes the meaning.

4. Classify according to the most relevant hallucination type.

[Execution]
Executing Step 1: `[Extraction]` The GD says the medication showed positive results only in some clinical trials with middle-aged patients, and emphasized that further research is needed.
Executing Step 2: `[Inference]` The claim says the medication has shown positive results in clinical trials — dropping both "some" and the restriction to "middle-aged patients."
Executing Step 3: `[Inference]` This eliminates critical qualifiers and context, making the statement broader and stronger than supported.
Executing Step 4: `[Inference]` This is not a direct contradiction or modality change, but specifically a qualifier/context alteration.
[Classification]
R13. Context/qualifier/emphasis alteration

Example 8
Grounding Document (GD):
The trial showed promising results in **some elderly patients**, but further testing is required.
Target Sentence:
The trial showed promising results in patients.
Answer:
Claim: The trial showed promising results in patients.
[Analysis]
[Planning]

1. Extract the specific scope and qualifiers from the document regarding trial results.

2. Compare the document's scope with the claim's scope.

3. Identify what qualifiers or restrictions have been removed.

4. Classify the type of hallucination.

[Execution]
Executing Step 1: `[Extraction]` The GD states the trial showed promising results in "some elderly patients" with the qualifier that "further testing is required."
Executing Step 2: `[Inference]` The claim states results were shown "in patients" without any age restriction or quantity qualifier.
Executing Step 3: `[Inference]` The claim drops both the qualifier "some" and the age restriction "elderly," broadening the scope beyond what the document supports.
Executing Step 4: `[Inference]` This represents removal of important qualifiers and contextual constraints that alter the meaning.
[Classification]
R13. Context/qualifier/emphasis alteration

Example 9
Grounding Document (GD):
The company released a new phone. In the same quarter, profits increased by 10%.
Target Sentence:
The company's new phone release caused profits to increase by 10%.
Answer:
Claim: The company's new phone release caused profits to increase by 10%.
[Analysis]
[Planning]

1. Extract the facts presented in the document.

2. Analyze the logical relationship implied in the claim.

3. Determine if the causal relationship is supported by the document.

4. Classify the type of hallucination.

[Execution]
Executing Step 1: `[Extraction]` The GD presents two separate facts: "The company released a new phone" and "In the same quarter, profits increased by 10%."
Executing Step 2: `[Inference]` The claim states that the phone release "caused" the profit increase, establishing a direct causal relationship.
Executing Step 3: `[Inference]` The document only mentions temporal co-occurrence ("in the same quarter") but provides no evidence of causation. The claim incorrectly merges two separate statements into a causal conclusion.
Executing Step 4: `[Inference]` This represents improper composition of distinct information without logical support from the document.
[Classification]
R14. Invalid logical inference/composition

Now, begin the task.

Grounding Document (GD):
{GROUNDING_DOCUMENT}

Target Sentence:
{SENTENCE}

K.2 Prompt Templates for SFT Training

This section provides the complete prompt templates used for the four SFT training configurations: `no_context_direct`, `context_direct`, `no_context_cot`, and `context_cot`. The goal of these templates is to capture different supervision signals for hallucination detection under varied reasoning and context settings:

- **No-Context, Direct Output** trains the model to provide immediate labels without access to preceding sentences or reasoning steps. (K.9)
- **Context-Aware, Direct Output** adds local response context for disambiguation but still requires concise, label-only answers. (K.10)
- **No-Context, Chain-of-Thought (CoT)** encourages explicit structured reasoning for a single statement, while ignoring any preceding response sentences. (K.11)

- **Context-Aware, Chain-of-Thought (CoT)** is the most comprehensive setting, combining both structured reasoning and the ability to resolve ambiguous references using prior response context. (K.12)

Together, these configurations allow us to study how context usage and reasoning style influence the model's labeling behavior and interpretability.

K.9: Prompt Template: No-Context, Direct Output (`no_context_direct`)

This setting instructs the model to provide a direct label for a single sentence without considering any preceding response context and without generating a chain of thought.

You are an expert hallucination detector. Compare the Statement ONLY against the reference document.

LABEL DEFINITIONS:

- `[Attributable]`: All key information in the Statement is supported by or directly inferable from the Document, with no conflicting information.
- `[Not Attributable]`: Any key part of the Statement is not mentioned or supported by the Document.
- `[Contradicted]`: Any key part of the Statement is clearly contradicted by the Document.

Analyze the Statement and respond with exactly one of the three labels: `Attributable`, `Not Attributable`, or `Contradicted`.

Document:
{DOCUMENT}

Statement:
{STATEMENT}

K.10: Prompt Template: Context-Aware, Direct Output (`context_direct`)

This setting instructs the model to provide a direct label for a target sentence, allowing it to use the preceding sentences (context) for disambiguation. It does not require a chain of thought.

You are an expert hallucination detector. Compare the Statement ONLY against the reference document.

LABEL DEFINITIONS:

- `[Attributable]`: All key information in the Statement is supported by or directly inferable from the Document, with no conflicting information.
- `[Not Attributable]`: Any key part of the Statement is not mentioned or supported by the Document.
- `[Contradicted]`: Any key part of the Statement is clearly contradicted by the Document.

CONTEXT HANDLING:

- Statements are identified by `<SOS>...<EOS>` markers. Content before `<SOS>` is optional context for disambiguation only and NEVER factual evidence. Disambiguation must be performed using only the context before `<SOS>`.

Analyze the Statement and respond with exactly one of the three labels: `Attributable`, `Not Attributable`, or `Contradicted`.

Document:
{DOCUMENT}

Statement:
{PRECEDING_CONTEXT} `<SOS>` {TARGET_SENTENCE} `<EOS>`

K.11: Prompt Template: No-Context, Chain-of-Thought (`no_context_cot`)

This setting instructs the model to perform a detailed, structured reasoning process (CoT) for a single sentence without considering any preceding response context.

You are an expert hallucination detector. Your task is to determine if a Statement is attributable to the Document. You must base your decision ONLY on the information within the Document.

LABEL DEFINITIONS:

- `[Attributable]`: All key information in the Statement is supported by or directly inferable from the Document, with

no conflicting information.

- [Not Attributable]: Any key part of the Statement is not mentioned or supported by the Document.
- [Contradicted]: Any key part of the Statement is clearly contradicted by the Document.

KEY JUDGING PRINCIPLES:

1. **Scope & Quantification:** Claims must match the Document's scope and quantifiers ('all', 'some', 'most'). No overgeneralization or unsupported specification of details.
2. **Causality & Logic:** Causal links, comparisons, and conditionals require explicit statement in the Document. Correlation is insufficient for causation.
3. **Certainty & Modality:** Statement certainty, intensity, and frequency must match the Document's. Don't upgrade possibilities to facts, or change intensity/frequency.
4. **Context & Qualifiers:** Preserve all important qualifiers, conditions, and contextual constraints from the Document.
5. **Subjective & Normative Claims:** Only attributable if the Document itself makes that specific judgment. Do not add evaluative opinions, normative prescriptions, or invented explanations.
6. **Temporal & Spatial Consistency:** Timeline, sequence, location, and order must align with the Document.
7. **Entity Integrity:** All entities, attributes, measurements, and their properties must be precisely as described in the Document. Mismatches are [Contradicted]; absences are [Not Attributable].

CONTEXT HANDLING:

- Statements are identified by <SOS>...<EOS> markers. Content before <SOS> is optional context for disambiguation only and NEVER factual evidence. Disambiguation must be performed using only the context before <SOS>. If resolving ambiguity requires information from the Document, address it in the [Analysis] section.

OUTPUT FORMAT:

```
[Disambiguation]
Need-Rewrite: <Yes or No>
Action: <Describe any disambiguation action taken, or "None">
Statement: <The original or rewritten statement being judged>

[Analysis]
[Planning]
- Step 1: <Your first verification step>
- Step 2: <Your second verification step>
...
[Execution]
- Executing Step 1:
  [Extraction] <Relevant information from the Document>
  [Inference] <Comparison and conclusion for this step>
- Executing Step 2:
  [Extraction] <...>
  [Inference] <...>

[Decision]
<One of the three labels: Attributable, Not Attributable, or Contradicted>
```

Provide your analysis following the structured format above.

Document:
{DOCUMENT}

Statement:
<SOS> {STATEMENT} <EOS>

K.12: Prompt Template: Context-Aware, Chain-of-Thought (context_cot)

This is the most comprehensive setting, instructing the model to perform a detailed reasoning process (CoT) for a target sentence while using the preceding sentences for context-aware disambiguation.

You are an expert hallucination detector. Your task is to determine if a Statement is attributable to the Document. You

must base your decision ONLY on the information within the Document.

LABEL DEFINITIONS:

- [Attributable]: All key information in the Statement is supported by or directly inferable from the Document, with no conflicting information.
- [Not Attributable]: Any key part of the Statement is not mentioned or supported by the Document.
- [Contradicted]: Any key part of the Statement is clearly contradicted by the Document.

KEY JUDGING PRINCIPLES:

1. **Scope & Quantification:** Claims must match the Document's scope and quantifiers ('all', 'some', 'most'). No overgeneralization or unsupported specification of details.
2. **Causality & Logic:** Causal links, comparisons, and conditionals require explicit statement in the Document. Correlation is insufficient for causation.
3. **Certainty & Modality:** Statement certainty, intensity, and frequency must match the Document's. Don't upgrade possibilities to facts, or change intensity/frequency.
4. **Context & Qualifiers:** Preserve all important qualifiers, conditions, and contextual constraints from the Document.
5. **Subjective & Normative Claims:** Only attributable if the Document itself makes that specific judgment. Do not add evaluative opinions, normative prescriptions, or invented explanations.
6. **Temporal & Spatial Consistency:** Timeline, sequence, location, and order must align with the Document.
7. **Entity Integrity:** All entities, attributes, measurements, and their properties must be precisely as described in the Document. Mismatches are [Contradicted]; absences are [Not Attributable].

CONTEXT HANDLING:

- Statements are identified by <SOS>...<EOS> markers. Content before <SOS> is optional context for disambiguation only and NEVER factual evidence. Disambiguation must be performed using only the context before <SOS>. If resolving ambiguity requires information from the Document, address it in the [Analysis] section.

OUTPUT FORMAT:

```
[Disambiguation]
Need-Rewrite: <Yes or No>
Action: <Describe any disambiguation action taken, or "None">
Statement: <The original or rewritten statement being judged>

[Analysis]
[Planning]
- Step 1: <Your first verification step>
- Step 2: <Your second verification step>
...
[Execution]
- Executing Step 1:
  [Extraction] <Relevant information from the Document>
  [Inference] <Comparison and conclusion for this step>
- Executing Step 2:
  [Extraction] <...>
  [Inference] <...>

[Decision]
<One of the three labels: Attributable, Not Attributable, or Contradicted>
```

Provide your analysis following the structured format above.

Document:
{DOCUMENT}

Statement:
{PRECEDING_CONTEXT} <SOS> {TARGET_SENTENCE} <EOS>