Rethinking Evaluation for LLM Hallucination Detection: A Desiderata, A New RAG-based Benchmark, New Insights

Wenbo Chen[1], Veena Padmanabhan[1], Tootiya Giyahchi[1], Elaine Wong[1], Leman Akoglu[1,2]

[1]Amazon [2]Carnegie Mellon University
{wbchen, veenapad, tootiya, elawong}@amazon.com, lakoglu@andrew.cmu.edu

Abstract

Hallucination, broadly referring to unfaithful, fabricated, or inconsistent content generated by LLMs, has wide-ranging implications. Therefore, a large body of effort has been devoted to detecting LLM hallucinations, as well as designing benchmark datasets for evaluating these detectors. In this work, we first establish a **desiderata** of properties for hallucination detection benchmarks (HDBs) to exhibit for effective evaluation. A critical look at existing HDBs through the lens of our desiderata reveals that *none of them exhibits all the properties*. We identify two largest gaps: **(1) RAG-based** grounded benchmarks with long context are severely lacking (partly because length impedes human annotation); and **(2)** Existing benchmarks do not make available realistic **label noise** for stress-testing detectors although real-world use-cases often grapple with label noise due to human or automated/weak annotation. To close these gaps, we build and open-source **a new RAG-based HDB** called TRIVIA+ that underwent a rigorous human annotation process. Notably, our benchmark exhibits all desirable properties including **(1)** TRIVIA+ contains samples with the longest context in the literature; and **(2)** we design and share four sets of noisy labels with different, both sample-dependent and sample-independent, noise schemes. Finally, we perform **experiments on RAG-based HDBs**, including our TRIVIA+, using popular SOTA detectors that reveal **new insights**: **(i)** ample room remains for current detectors to reach the performance ceiling on RAG-based HDBs, **(ii)** the basic LLM-as-a-Judge baseline performs competitively, and **(iii)** label noise hinders detection performance. We expect that our findings, along with our proposed benchmark[1], will motivate and foster needed research on hallucination detection for RAG-based tasks.

[1]We release the benchmark dataset with expert annotations at github.com/amazon-science/hallucination-benchmark-trivialplus.

1 Introduction

Modern large language models (LLMs) have propelled advances in various real-world domains including e-commerce (Jiang et al., 2024), medicine (Thirunavukarasu et al., 2023), law (tho; Greenstein), to name a few. While LLM-driven AI technologies offer real-world impact, the fact that they *hallucinate* remains a big obstacle in the safe and secure usage of generative AI tools (Natalie Sherman; Magesh et al., 2024; Hong et al., 2024).

Various efforts have been devoted to preventing LLM hallucinations. Mitigation strategies include retrieval-augmented generation (RAG) (Li et al., 2024), reasoning (Dhuliawala et al., 2024), self-reflection (Ji et al., 2023), self-refinement (Madaan et al., 2023), among others (Zhang et al., 2023). Nevertheless, hallucinations persist making both proactive mitigation and post hoc detection essential (Luo et al., 2024). As a result, the field has seen a surge of interest on hallucination detection approaches (Huang et al., 2025; Luo et al., 2024; Zhang et al., 2023) and a large body of hallucination detection benchmark (HDB) datasets have also been developed (see Table 1).

Despite the long list of HDBs, we point to a large gap: **only a handful of RAG-based HDBs exists in the literature**. These HDBs are especially hard for humans to annotate, as they (1) exhibit considerably *long context* settings, and (2) associate with *knowledge-intensive* tasks (Lewis et al., 2020). In this work, we build and open-source a new RAG-based HDB called TRIVIA+; representative of knowledge-intensive LLM tasks and thus grounded on often long context—rendering hallucination detection even more challenging. Furthermore, we identify characteristics that a HDB ideally exhibits. While admittedly not exhaustive, the list offers a systematic lens to assess existing HDBs. Lastly, experiments on RAG-based HDBs report new findings. The following summarizes the

contributions of this work.

- **Desiderata for HDBs (Hallucination Detection Benchmarks):** We compose a list of desirable properties for a HDB to exhibit, through the lens of which we scrutinize the existing HDBs—revealing that none of them satisfies all the criteria in our desiderata. (See Table 1.)
- **A New Benchmark:** We introduce TRIVIA+, a new RAG-based HDB that exhibits seven key properties in our desiderata. Namely, it consists of **(i)** *organic*, naturally-occurring hallucinations; **(ii)** *human-verified* labels, with each sample annotated up to 6 times; **(iii)** *long-context* (i.e. high-demand, hard-to-label) LLM tasks; **(iv)** training *labels with realistic noise* (for supervised detectors); **(v)** *extrinsic and intrinsic* hallucinations; from **(vi)** *multiple* LLMs; and **(vii)** *multiple* domains.
- **Empirical Findings:** Experiments on RAG-based HDBs find that **(1)** hallucination prevails in grounded tasks and current detectors leave ample room to ceiling performance, **(2)** while supervised and fine-tuned models can be effective, the simple LLM-as-a-Judge performs competitively, in contrast to reports in past literature (Niu et al., 2024), and **(3)** sample-dependent (as opposed to random) label noise hinders detection significantly.
- **Open Problems:** Our findings highlight the scarcity of **(a)** HDBs with long context—and effective detectors on RAG-based generations, **(b)** HDBs with realistic label noise—and detectors leveraging robust learning; and **(c)** HDBs that span from multiple domains—and detectors that can generalize across domains.

2 A Desiderata for Hallucination Detection Benchmarking

We start with the question: *What properties should a HDB exhibit?* We identify seven desirable properties that we discuss and motivate in this section. While our list may not be exhaustive, it provides a solid foundation on which we scrutinize existing HDBs in the next section.

D1. Organic (i.e. real, naturally-occurring) Hallucinations: An HDB should ideally contain content organically generated by an LLM, i.e. on its **natural responses**. In contrast, hallucinations injected directly (manual) or indirectly (via prompting an LLM to hallucinate) are considered **non**-organic. In general, we find that despite looking similar to a human eye, non-organic hallucinations are easier to detect and serve as deceptive benchmarks (see Figure 1).

D2. Human-Verified, Reliable Test Labels: To keep a fair and accurate record of progress on a task, it is paramount for benchmark datasets to be equipped with trustworthy labels for evaluation. In the ideal scenario, labels provided by *many*, *expert human* annotators with *perfect agreement* would be considered **gold-grade** ground-truth. Label collection is particularly challenging due to (i) the presence of *long-context* samples (e.g. RAG) (ii) the *knowledge-intensive* task of reviewing and absorbing complex-content and (iii) the presence of *subtle* hallucinations (especially in long-form responses) akin to a "needle" in the "haystack" (i.e. long context), possibly eluding even the most vigilant annotators. Nevertheless, due diligence should be undertaken to reach as trustworthy and gold-grade labels as possible—with full transparency on the process, regarding the count, expertise, and dis/agreement of the annotators. We note that while human-verified labels are the gold standard, they are costly and time-consuming to obtain, especially for long-context RAG tasks. LLM-as-a-judge labels (e.g. (Ji et al., 2024)) offer a practical but noisy alternative, further motivating D4.

D3. Long Context (high-demand yet hard-to-label) Tasks: Among the many tasks LLMs are employed on, a HDB should consider tasks (i) that are practically-useful and *popular* (i.e. in high demand), while (ii) potential hallucinations are *hard* for humans (even experts) to quickly identify. Then, such a benchmark would contribute to keeping a healthy record of most effective detectors on such high-demand, high-value tasks. RAG-based long context tasks fall exactly under this category with little emphasis in the HDB literature.

D4. Realistic Training Labels: While several hallucination detectors are unsupervised due to the laborious label gathering involved, some utilize labels. Labels are obtained through various means, including (i) employing an LLM-as-a-Judge (or other unsupervised detectors) or (ii) manual labeling, i.e. human annotation. Depending on the count and expertise of the labelers, one may end up with **silver-grade** labels. In either case, the labels would be **noisy**. Further, the labels would exhibit systematic, *sample-dependent* noise, as opposed to random noise. Then, it would be beneficial for

Table 1: Comparison of HDBs w.r.t. desirable desiderata (Sec. 2). Proposed TRIVIA+ exhibits all seven properties; with **(i)** organic generations, **(ii)** human-verified labels (for evaluation), **(iii)** four sets of realistic label noise (for training), on **(iv)** RAG-based long-context tasks (knowledge-intensive, hard-to-annotate), with **(v)** both extrinsic and intrinsic hallucinations, from **(vi)** multiple modern LLMs and **(vii)** multiple domains. **None of the existing HDBs fully meets the desiderata.** RAG-based HDBs are few, and are highlighted in red. Those marked with $^{\times}$ do not make LLM generations available. Symbol ? denotes that it is unclear whether the dataset spans multiple domains.

#	Dataset	Hal.s: Organic	Test-Labels: Human	Train-Labels: Realistic noise	Context: Long/RAG	Type: Faith.	LLMs: Multiple	Domains: Multiple
1	HADES (Liu et al., 2022)		✓					?
2	Frank (Marfurt and Henderson, 2022)	✓				✓		
3	TLHD-CNNDM (Marfurt and Henderson, 2022)	✓	✓			✓		
4	SummaC (Laban et al., 2022)	✓	✓			✓		✓
5	WikiBio+ (Manakul et al., 2023)	✓	✓					
6	HaluEval (Li et al., 2023)				✓	✓		✓
7	HILT (Rawte et al., 2023)	✓$^{\times}$	✓				✓	?
8	PHD (Yang et al., 2023)	✓	✓					✓
9	DelucionQA (Sadat et al., 2023)	✓$^{\times}$	✓		✓	✓		
10	RAGTruth (Niu et al., 2024)	✓	✓		✓	✓	✓	✓
11	TofuEval (Tang et al., 2024)	✓	✓			✓	✓	?
12	FAVABench (Mishra et al., 2024)	✓	✓				✓	✓
13	SHROOM (Mickus et al., 2024)	✓	✓	*		✓	✓	✓
14	DiaHalu (Chen et al., 2024a)	✓	✓			✓	✓	✓
15	ERBench (Oh et al., 2024)	✓	auto				✓	✓
16	Dolly (AC) (Hu et al., 2024)	✓	✓			✓	✓	✓
17	Dolly (NC) (Hu et al., 2024)	✓	✓		✓	✓	✓	✓
18	ANAH (Ji et al., 2024)	✓	✓			✓	✓	✓
19	HaluEval-Wild (Zhu et al., 2025)	✓$^{\times}$						✓
20	FACTS (Jacovi et al., 2025)	✓$^{\times}$			✓	✓	✓	✓
21	Mu-SHROOM (Vázquez et al., 2025)	✓	✓	*			✓	✓
22	FaithEval (Ming et al., 2025)		✓			✓	✓	✓
23	FaithBench (Bao et al., 2025)	✓	✓			✓	✓	✓
24	TRIVIA+ (this paper)	✓	✓	✓	✓	✓	✓	✓

an HDB to alleviate the label-collection burden on detection teams by incorporating labeled training data for downstream detectors, while allowing the training data to exhibit *varying degrees of label quality* in line with the aforementioned realistic settings. While (semi-/supervised) detection teams may split the gold-grade or "clean" evaluation data into train/test for their purposes, we remark that a *noisy* training data would better reflect reality.

D5. Comprehensive Hallucination Types: LLM hallucinations are defined in various ways across the literature, broadly referring to unfaithful, fabricated, inconsistent, or irrational content. We adopt the terminology: **intrinsic** and **extrinsic** hallucinations. Former refers to outputs that are **inconsistent** with the provided reference context. Latter refer to outputs that are **unverifiable** by the context. Others have termed both types in our definition as **faithfulness** hallucination, while used the term **factuality** hallucination for content that is factually incorrect or fabricated (Huang et al., 2025). In comparison, we categorize factual errors that contradict the source context as intrinsic, while those with no basis in external knowledge as extrinsic hallucination since by induction they also have no basis in the reference context. The differences in terminology aside, a HDB ideally represents different types of hallucinations.

D6. LLM Diversity: It is important to derive organic, natural responses from a number of different LLMs, to keep an unbiased record of progress in solving the detection task. Furthermore, a benchmark is as relevant as the LLMs it has employed, therefore it is desirable to consider modern and/or popular LLMs. Admittedly, this is a moving target/goal given the fast-pace of evolution today's LLMs go through, but a relevant one nevertheless.

D7. Multiple Domains: Hallucination detection is relevant in many scenarios as LLMs are employed in diverse domains. As such, detectors that can generalize across domains become critical. HDBs that provide multi-domain generations stress-test and promote domain generalization.

3 Existing Benchmarks – A Critical Look

Our desiderata provides us with an analytical lens through which we can take a critical look at the existing hallucination detection benchmarks and

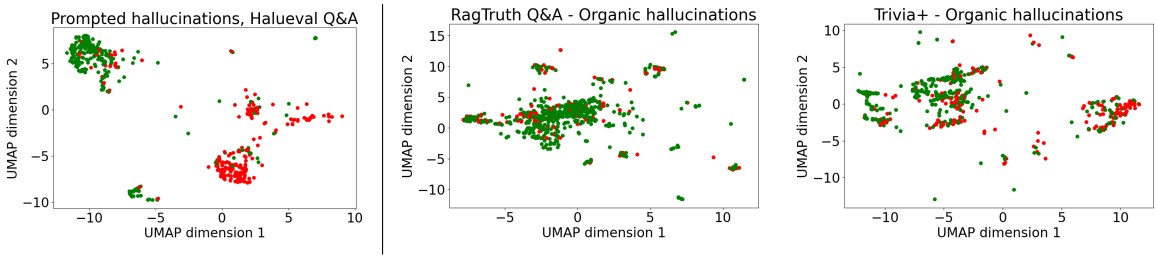

Figure 1: (best in color) Supervised test split UMAP embeddings generated by fitting on the train split using `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli` (Laurer et al., 2022). **(left) prompted** positive samples (i.e. hallucinations) in red in HaluEval stand out away from negatives (green); whereas **organic** hallucinations in RAGTruth **(center)** and TRIVIA+ **(right)** blend in with and resemble the negatives. This visual separability is supported quantitatively in Table 4: SFT achieves 99.6% F1 on non-organic HaluEval but only 66–69% F1 on organic benchmarks.

highlight their strengths and weaknesses. This section elaborates on desirable benchmark properties, as organized into three: Core Properties (**D1**–**D2**), Largest Gaps in literature (**D3**–**D4**), and Diversity Considerations (**D5**–**D7**). Table 1 provides a summary, illustrating the gaps in the literature.

3.1 Core Properties: D1. Organic Generations & D2. Verified Test Labels

The core properties in our desiderata are: **Organic** (text) and **Human-verified** (labels). These are arguably the two *must-have* criteria that any hallucination benchmark should exhibit. Organic content represent naturally-generated (vs. contrived) hallucinations, while human-verified labels offer trustworthy evaluation results and a fair record of progress on hallucination detection.

Table 1 shows that not all benchmarks readily satisfy these core criteria. *Why do we have benchmarks with non-organic hallucinations?* When recruiting human-annotators is prohibitive, the typical practice is to go backwards and inject *known* hallucinations: (1) directly; by manual perturbation[2] (Liu et al., 2022) or (2) indirectly; by dictating (i.e. prompting) an LLM to hallucinate (Li et al., 2023). However, known injected hallucinations offer only the *illusion* of control on label quality, since LLMs may fail to follow instruction when prompted to create hallucinations—yielding false-positive label noise, while the "negative" samples may contain (organic) hallucinations—yielding false-negative label noise. Further, prompted LLM hallucinations differ from organically-generated ones, as Figure 1 illustrates, and thus may have limited representation of *natural* hallucinations.

3.2 Largest Gaps: D3. RAG-based Tasks & D4. Realistic Train Labels

While diligently human-verified labels are essential for valid evaluation, they are costly to obtain. As a result, various supervised approaches in the literature often resort to data augmentation with pseudo or proxy labels for training, including several top teams on the SemEval 2024 SHROOM challenge[3] (Chen et al., 2024c; Rykov et al., 2024; Borra et al., 2024). Most common practice is to employ an LLM as a "judge", sometimes cross-checking consistency between multiple LLMs (Sun et al., 2024; Jacovi et al., 2025). However, the degree of noise these training datasets exhibit are often unknown and the vulnerability of detectors to noisy training labels has not been systematically studied.

Alarmingly, Zhu *et al.* show that modern models like BERT are (1) quite vulnerable to *sample-dependent* label noise from weak supervision, despite being robust against *random* noise; and that (2) noise-handling methods do not always improve its performance and may even deteriorate it (Zhu et al., 2022). Their study focused on text classification, while there exists no study on susceptibility of hallucination detectors to label noise from weak supervision. Further, existing work on robust text classification with language models do not consider hallucination detection (Agro and Aldarmaki, 2023; Qi et al., 2023; Chong et al., 2022).

These motivate the necessity of benchmarks with realistic, *sample-dependent* noisy training labels. The current literature, however, falls short in offering a common ground to stress-test existing detectors and thereby fostering research on LNL (learn-

[2]Given LLM-generated text, few manually-selected words are masked and re-sampled (varying temperature) from the LLM to yield possibly hallucinated output. As such, generation is intervened rather than natural.

[3]https://helsinki-nlp.github.io/shroom/2024

ing with noisy labels) in this problem context. In fact, this is the largest gap we find in the literature as Table 1 highlights.[4]

The second major gap is the shortage of available benchmarks that involve RAG-based LLM generations, which limits their ability to reflect how modern LLMs are typically used—often heavily augmented with RAG for accessing relevant, up-to-date information to create accurate, contextually rich, and grounded responses. The few that are RAG based have short context, that is up to 3× smaller than TRIVIA+ on average (see Table 2)—falling short in representing the advances in efficient, long-context LLMs.

3.3 Diversity: D5. Hallucination Types, D6. LLMs, D7. Domains

Ideally a benchmark should be representative of hallucinations of various nature, various modern LLMs in popularity, as well as diverse domains they are used in. Most earlier "closed-book" benchmarks focus on factuality of LLM responses, with approaches such as FActScore (Min et al., 2023) evaluating factual precision at a fine-grained, atomic claim level. On the other hand, faithfulness becomes relevant for context-driven tasks, such as summarization, paraphrasing, RAG-based QA. Unfaithful content can be inconsistent with (intrinsic) or unverifiable from (extrinsic) the given context, which *subsumes* factual errors in QA tasks assuming accurate context—thus involving various types of hallucinations.

We see in Table 1 that while the majority of existing benchmarks consider faithfulness, include multiple LLM generations[5] from various domains, fewer than half satisfy all 3 criteria simultaneously.

4 Proposed Benchmark: TRIVIA+

We introduce TRIVIA+, a novel dataset contributing to the need for RAG-based HDBs with natural hallucinations and human-validated labels. We provide a detailed description of the contents and the label annotation process as follows.

[4]In the table, ∗ depicts that SemEval benchmarks do not originally provide noisy labels, yet, individual labels from 5 annotators are shared, which can be used to apply noising strategies (Chong et al., 2022) to their final training labels.

[5]In the table, ✓× depicts HDBs which do not make available the LLM generations as well as the labels, including FACTS, HaluEval-Wild and DelucionQA.

4.1 Description

TRIVIA+ contains generations from three less-represented but popular LLMs: A commercially available SOTA LLM, Gemma-7b, and Mixtral 8x7b. For domain diversity, we source prompts from multiple established datasets, specifically TRIVIAQA (Joshi et al., 2017), NaturalQuestions (Kwiatkowski et al., 2019), MS-MARCO(Bajaj et al., 2016), CovidQA(Möller et al., 2020) and DROP(Dua et al., 2019). Each prompt contains reference material, pulled directly from the source dataset, and a question (or query). Our dataset construction methodology employs a strategic filtering approach: We first query the commercially available SOTA LLM[6], the largest of the three, with questions from these datasets and use ROUGE Score (Lin, 2004) comparisons between ground truth answers and generated responses to identify low-similarity cases (with a filter for similarity scores < 0.1). Our hypothesis is that low ROUGE overlap between the generated answer and the ground-truth answer is likely to correlate with a higher hallucination rate, as low-overlap answers are more likely to deviate from the expected content. This filtering serves as a resource-efficient strategy to enrich the proportion of true hallucinations given limited annotator resources, without altering LLM outputs—thereby preserving their organic nature. Using the same context-question pairs, we then prompt the two other LLMs, creating annotation triplets comprising context, question, and answers from all three models. Each triplet receives multiple labels from annotators to ensure label reliability, with specifics of the annotation process described below.

4.2 Human Annotation

Labels & Definitions: We instructed annotators to provide labels at a sentence level, with each sentence receiving one of four labels: 'Supported', 'Contradicted', 'Not Mentioned' and 'Supplementary', closely aligning with the definition used by FACTS (Jacovi et al., 2025) . The definition aligns with faithfulness: 'Not Mentioned' and 'Contradicted' labels correspond to 'unfaithful', and 'Supported' and 'Supplementary' correspond to 'faithful'. More details are included in Apdx. A.5.

[6]Legal constraints prevent us from naming the specific LLM other than few tech-specs; $>$ 150B parameters, 2024 release. Labeling methodology is model-agnostic and findings should generalize.

Table 2: Stats for RAG-based HDBs. TRIVIA+ exhibits the longest context samples with higher domain diversity.

	context length (#char.s)			#samples	%hal.s	#LLMs	domains
Benchmark	median	mean	max				
HaluEval (QA)	321	344	1557	20K	50	1	Multi-hop QA (HotPotQA-based)
RAGTruth (QA)	1.2K	1.3K	2.8K	989	29.1	6	Web searches (MS-MARCO-based)
Dolly (NC)	2.97K	3.1K	5.99K	100	44.5	7	Web searches (MS-MARCO-based)
TRIVIA+	2.8K	9.3K	94K	3224	35	3	Paragraph reasoning, Web Searches, Medical docs, Wikipedia

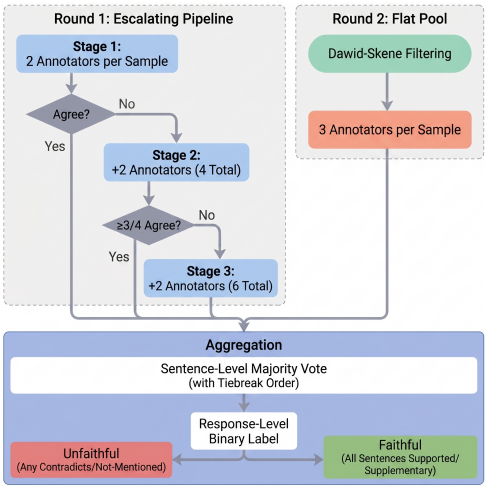

Figure 2: Overview of the annotation pipeline. Round 1 uses an escalating strategy: each sentence starts with 2 annotators and escalates to 4 or 6 upon disagreement. Round 2 employs a flat pool of 3 annotators per sample after filtering low-quality workers via the Dawid-Skene model. Sentence-level labels are aggregated by majority vote, then mapped to a binary response-level label.

Annotators: Two rounds of training were performed with a pool of 18 of annotators fluent in English. The team's annotations were audited by the authors to provide detailed feedback in cases where systematic error was observed. At the end of two rounds of training, we measured average accuracy when two annotators agreed on an annotation at 89% but individual accuracy was about 77%. We conjecture that our dataset was particularly challenging for human annotators due to long-context examples. We gathered multiple votes per sample from annotators to boost accuracy through aggregation (described below).

4.2.1 Multi-Vote Annotation for Evaluation Test Labels

As illustrated in Figure 2, the annotation task was divided into two rounds and was performed at the sentence level. In the first round, we utilized a multi-stage pipeline. Two annotators labeled each sample. When disagreement was observed, two additional annotators provided labels. If there still was no clear majority in voting (three out of four labels being consistent), two additional labels were gathered. As such, each sample received up to six annotations. For the second round, we removed several low performing annotators from our annotator pool based on the Dawid–Skene model (Dawid and Skene, 1979), and for the remaining two thirds of the data, three annotators labeled each sample.

Labels were collected at sentence level but aggregated to the answer (i.e. response) level, keeping the strictest label during aggregation (i.e. 'Contradicted' the strictest, then 'Not Mentioned', both mapping to 'unfaithful' binary labels). This strictest-label aggregation follows standard practice in faithfulness evaluation (Jacovi et al., 2025; Niu et al., 2024). We will additionally release sentence-level labels to enable researchers to explore alternative aggregation strategies (e.g. weighted, majority-vote). A worked example of this aggregation process is provided in Appendix A.6.

4.2.2 Noising Strategies for Training Labels

Besides gold-grade/clean labels as described above, TRIVIA+ includes four different sets of noisy labels. These are derived from two different sources: **1)** noise from *weak supervision*, and **2)** noise from *annotator judgement*. Noisy labels can be used to study robust learning under label noise, and aim to mimic the typical scenarios wherein semi-/supervised detectors are likely exposed to some form of noisy labeled data during training.

1) Noise from Weak-Supervision: For training supervised detectors, pseudo or proxy labels can be obtained from any unsupervised detector, often an LLM employed as a "judge" (Gekhman et al., 2023; Mickus et al., 2024; Zhu et al., 2025). Following common practice, we instruct a commercially available SOTA LLM[7] to evaluate whether the LLM responses are coherent with the provided context. While Zheng et al. (2023) showed strong LLM "judges" like GPT-4 can match crowdsourced human preferences well, with over 80% agreement

[7]Our prompt is given in Appendix A.1.

on dialog, we note that these techniques do not perform well on all datasets, such as TRIVIA+ and Dolly NC (see Table 4), and the labels derived are noisy; concretely on TRIVIA+, this approach only achieves 74.9% accuracy.

2) Noise from Annotators: We employ the two noising methods by Chong et al. (2022) to mimic noise driven by human error; namely dissenting worker (DW), and dissenting label (DL), and additionally simulate random label flips (RF). DW selects one annotator at random, and applies all of their labels which disagree with the final labels, simulating gaps in annotator training. We extend this to allow choosing multiple annotators to reach a desired noise level. DL replaces final labels with disagreeing labels at random, simulating imperfect quality control. RF randomly flips a fraction of the final labels. We simulate 15% label noise in each setting. Note that both DW and DL incur realistic, sample-dependent noise, whereas RF yields sample-independent noise due to random flips, which we include for comparison.

5 Experiments

Our work focuses on hallucinations in RAG-based long context LLM generations. As such, we design experiments primarily to study the detection performance on those HDBs in the literature as well as our proposed TRIVIA+ using popular state-of-the-art methods. Different from all HDB literature, we also investigate the effect of noise in both training and test data. We use 'clean' vs. noisy to refer to labels that are human-verified vs. not, respectively.

Datasets: RAG-based benchmarks are quite limited in the literature as Table 1 underscores. Among the existing five (highlighted in red), FACTS (Jacovi et al., 2025) does not make their LLM generations available. DelucionQA (Sadat et al., 2023) is narrow in scope, targeting QA from a car's manual as context. Thus, we use the remaining three: **HaluEval** (Li et al., 2023) captures (for RAG-based QA) prompted hallucinations, utilizing `gpt-3.5-turbo` to generate incorrect answers, **RAGTruthQA** (Niu et al., 2024) uses reference–question pairs from `MS-MARCO` (Bajaj et al., 2016), a reading comprehension dataset, and human-annotated natural hallucinations from `gpt-3.5-turbo-0613`, `gpt-4-0613`, `Llama-2-7B-chat`, `Llama-2-13B-chat`, `Llama-2-70B-chat` and `Mistral-7B-Instruct`. Lastly, **Dolly(NC)** (Hu et al., 2024) provides a small annotated sample derived from `GPT4`, `GPT-3.5-Turbo`, `InstructGPT`, `Falcon (Falcon-40B-Instruct)`, `Alpaca (Alpaca-7B)`, `LLaMA2(70B-Chat)` and `Claude 2`. Table 2 provides basic statistics, while we refer to the original papers for details. Notably, TRIVIA+ contains samples with significantly longer context than the RAG-based HDBs in the literature.

Detection Methods: For hallucination detection, we employ both un/supervised approaches that are widely used in practice. Unsupervised detectors include **SelfCheckGPT** (we use a temperature setting of 1.0 and the NLI variant, as recommended by the authors (Manakul et al., 2023), which we equip with `GPT-4-mini` as well as `Claude-Sonnet-3.5-v2` separately, using three generations to determine consistency); and **LLM-as-a-Judge** (zero-shot with fixed, carefully hand-engineered prompt, see Apdx. A.1). Those that leverage labels include few-shot **FS**, which augments the LLM-as-a-Judge prompt with three random misclassified examples from the validation set; prompt-optimized **PO**, which is obtained by inserting the FS prompt into Anthropic's prompt optimization tool (Anthropic, 2025); and **SFT**, which performs supervised instruction fine-tuning of `Mistral-7B-Instruct-v0.2` using LoRA (Hu et al., 2022) (r=16). Appendix A.2 provides details on model configurations. We use train/test splits consistent with original datasets, and further split the former into train/val at 80/20 at random using reference/context to split to ensure no data overlap between splits.

5.1 Detection Results on RAG-based HDBs

Table 4 presents all detection performance results of 5 different detectors on 4 grounded HDBs including TRIVIA+. We report Precision, Recall and F1 as LLM-as-a-Judge, FS, and PO provide binary labels. For SelfCheckGPT and SFT we select a threshold that maximizes the F1 score. Apdx. Table 7 reports additional metrics, ROC-AUC, PR-AUC and Accuracy, with similar results.

First, we observe a **stark difference between performances on the non-organic HaluEval versus on the other HDBs with organic hallucinations**. Specifically, F1 scores on HaluEval are consistently above 0.8 across all detectors, and as high as 0.996 for SFT. It is expected that HaluEval benefits greatly from supervised fine-tuning, provided that its non-organic hallucinations are quite

separable as illustrated earlier in Figure 1.

In contrast, detection performances on organic RAG-based HDBs are considerably lower. F1 score remains strictly below 0.7 or lower across all detectors, including SFT. While SFT offers a boost particularly on RAGTruth, the gap between unsupervised and supervised detectors remain underwhelming as compared to those on HaluEval. In fact, we find that **LLM-as-a-Judge stands out as a simple yet competitive approach**, often rivaling those that leverage labels. This is in contrast to results reported in earlier work (Niu et al., 2024), where LLM-as-a-Judge approaches underperformed supervised techniques; the change may be attributed to the recent rapid performance improvement in LLMs as well as carefully engineering our prompt (included in Apdx. A.1).

In short, these results reveal a **significant gap between current detector performances and the optimally achievable or practically desirable performance on organic RAG-based HDBs**. Recent unsupervised methods for RAG-based hallucination detection, such as ReDeEP (Zhang et al., 2024) and LUMINA (Xu et al., 2025), also tackle this problem. We evaluate ReDeEP on TRIVIA+ and RAGTruth with both LLaMA-2-7B and 13B backbones (more details in Appendix A.8). F1 drops from 0.630/0.612 on RAGTruth to 0.535/0.531 on TRIVIA+ for 7B/13B respectively, further confirming that TRIVIA+ poses a greater challenge. Notably, 14–22% of TRIVIA+ responses cannot be processed due to long contexts, exposing a structural limitation of attention-based detection on long-context benchmarks.

Furthermore, stratifying detection performance on TRIVIA+ by context length (Table 3) reveals that **all detectors degrade sharply on long contexts** (>5K characters), with F1 drops of 0.09–0.23 compared to short contexts. This long-context regime is unique to TRIVIA+. RAGTruth and Dolly cannot test it due to their shorter contexts (Table 2).

5.2 Effect of Label Noise

In this section, we study the effect of label noise on evaluation (where test labels have noise) as well as on training (where train labels have noise). We use four types of noise: three sample-dependent—Weak Supervision (WS) from an LLM, as well as Dissenting Worker (DW) and Dissenting Label (DL) from human annotation—and one sample-independent scheme, Random Flipping (RF), for comparison. All noise levels are at 15%.

Table 3: Detector F1 on TRIVIA+ stratified by context length (in characters). All detectors degrade on long contexts (>5K).

Method	Short (<1K)	Med. (1K–5K)	Long (>5K)
SFT	0.725	0.702	0.504
SC-GPT (C)	0.739	0.732	0.508
SC-GPT (G)	0.700	0.632	0.506
LLM-aaJ	0.712	0.722	0.621
FS	0.711	0.732	0.594
PO	0.701	0.725	0.535

Table 4: **Detection results leave ample room to ceiling performances** on RAG-QA HDB datasets by SC-GPT (G): SelfCheckGPT with GPT-4-mini, SC-GPT (C): SelfCheckGPT with Claude-Sonnet-3.5, LLM-aaJ: LLM-as-a-Judge with Claude-Sonnet-3.5, FS: few-shot, PO: prompt-optimized, SFT: supervised fine-tuning.

Dataset	Model	F1	Precis.	Recall	Acc
HaluEval	SC-GPT (G)	0.870	0.891	0.848	0.872
	SC-GPT (C)	0.869	0.880	0.858	0.871
	LLM-aaJ	0.825	0.783	0.872	0.815
	FS	0.828	0.772	0.892	0.814
	PO	0.843	0.790	0.903	0.832
	SFT	0.996	0.999	0.993	0.996
RAGTruth	SC-GPT (G)	0.342	0.209	0.938	0.358
	SC-GPT (C)	0.346	0.213	0.919	0.383
	LLM-aaJ	0.617	0.482	0.856	0.806
	FS	0.628	0.490	0.875	0.811
	PO	0.538	0.373	0.963	0.697
	SFT	0.671	0.644	0.700	0.874
Dolly (NC)	SC-GPT (G)	0.664	0.547	0.845	0.619
	SC-GPT (C)	0.667	0.567	0.810	0.651
	LLM-aaJ	0.651	0.646	0.655	0.697
	FS	0.646	0.569	0.748	0.648
	PO	0.645	0.546	0.788	0.627
	SFT	-	-	-	-
TRIVIA+	SC-GPT (G)	0.614	0.486	0.835	0.636
	SC-GPT (C)	0.675	0.629	0.728	0.757
	LLM-aaJ	0.694	0.601	0.821	0.749
	FS	0.692	0.585	0.848	0.738
	PO	0.670	0.564	0.826	0.718
	SFT	0.663	0.581	0.772	0.727

Table 5 presents *measured* vs. *true* detection performances based on noisy vs. clean test labels, respectively. Here supervised models use clean training labels. We observe optimistically biased measured performances for LLM-based detectors when evaluated on LLM-based WS-labels. Results remain similar with DW and DL that also exhibit sample-dependent noise. We conjecture those involve hard samples near the decision boundary on which detectors make offsetting errors in both directions. In contrast, RF injects label noise into easy samples where detectors would otherwise perform accurately, leading to pessimistic under-reporting.

Table 6 presents results for supervised models

Table 5: **Noisy test labels lead to biased performance evaluation.** (top) Measured performances when detectors are evaluated on 4 different noisy test labels vs. (bottom) True performances on clean TRIVIA+ test labels. WS: Weak (LLM) Supervision, CM: Crowd Majority, DW: Dissenting Worker, DL: Dissenting Label, RF: Random Flip.

EVAL on:	**SC-GPT (C)**		**LLM-aaJ**		**FS**		**PO**		**SFT**	
	F1	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**
WS	0.763	0.747	n/a	n/a	0.929	0.930	0.912	0.913	0.708	0.685
DW	0.678	0.710	0.680	0.724	0.682	0.716	0.664	0.699	0.664	0.716
DL	0.644	0.692	0.651	0.707	0.636	0.684	0.622	0.670	0.612	0.665
RF	0.611	0.578	0.609	0.664	0.613	0.656	0.607	0.648	0.620	0.605
Clean	0.675	0.757	0.694	0.749	0.692	0.738	0.670	0.718	0.663	0.727

only when exposed to noisy vs. clean labels during training. Evaluation is on the *same* test data with *clean* labels. We find that FS and PO are inherently more robust to noise as they utilize only a few labeled examples *locally* as the in-context examples. In comparison, the *globally* fine-tuned SFT performance is hindered by label noise. We do not observe a significant difference between sample in/dependent noise schemes.

These results underscore needed research on RAG-based LLM hallucinations, where all models including supervised ones underperform and the latter can be further hindered by label noise.

Table 6: **Noisy train labels hinder supervised detectors;** especially those with global training (SFT) more so than local, in-context methods (FS, PO).

TRAIN on:	**FS**		**PO**		**SFT**	
	F1	**Acc**	**F1**	**Acc**	**F1**	**Acc**
WS	n/a	n/a	n/a	n/a	0.643	0.701
DW	0.692	0.744	0.675	0.729	0.638	0.698
DL	0.685	0.730	0.675	0.718	0.632	0.684
RF	0.687	0.733	0.656	0.698	0.631	0.732
Clean	0.692	0.738	0.670	0.718	0.663	0.727

6 Conclusion and Future Work

Hallucinations in LLM-generated content pose a pressing real-world problem that is likely to intensify as LLMs become more deeply integrated into everyday life. While RAG-based approaches have become increasingly prevalent for grounding LLMs in up-to-date domain-specific knowledge, LLM hallucinations still persist even under such grounding. Yet, the literature on hallucination detection benchmarks (HDB) with a focus on RAG-based tasks remain quite slim.

In this work, we first established a desiderata of desirable properties for HDBs to exhibit. While not necessarily exhaustive, the seven criteria in our desiderata offers a lens through which we revisited prominent existing HDBs, showing that none of them satisfies all the criteria. The largest gaps in the literature include: (1) RAG-based generations with long context, and (2) training labels with realistic noise. Notably, some HDBs even fail to meet two essential properties: organic hallucinations and human-verified evaluation labels.

In light of these findings, we designed and open-sourced a new RAG-based HDB called TRIVIA+, with long-context samples and four sets of noisy training labels, meeting all the criteria in our desiderata. Experiments revealed underwhelming performance by widely used un/supervised detectors on RAG-based HDBs, including TRIVIA+, leaving ample room for future work on detection.

Future work on HDBs could continue developing additional RAG-based HDBs with long context to foster research on mitigating hallucinations in widely-used RAG applications. It is equally important that future HDBs meet at least the criteria outlined in our desiderata. Among those, we highlight two that will likely benefit current and future LLM applications the most: First is HDBs that span multiple domains—fostering needed research on cross-domain generalization as LLMs are increasingly deployed in novel areas. Second is HDBs equipped with realistic label noise—promoting the application of robust learning under noisy labels (LNL) literature to this critical problem that often faces scarcity of reliably labeled data. Third, cross-dataset evaluation—training on one HDB (e.g. TRIVIA+) and evaluating on another (e.g. RAGTruth or Dolly)—is a promising direction for testing domain generalization of hallucination detectors.

7 Limitations

Our work and proposed hallucination detection benchmark (HDB) TRIVIA+ focused on RAG-based LLM generations. We identify three scope limitations. First, our definition is limited to faithfulness, i.e. consistency with retrieved reference context, rather than factuality. Assuming the reference context is accurate, faithfulness hallucinations subsume factual errors but also include those that contradict or have no basis in the context despite being factual. Second, we focused on knowledge-intensive question-answering (QA) tasks only. These typically necessitate retrieval of relevant, possibly lengthy context. Other reference based tasks such as summarization, translation, multi-turn dialog are not represented in TRIVIA+. Third, our work considers unimodal, text only generations. While there exist multi-modal benchmarks for image-text (Zhou et al., 2023; Chen et al., 2024b) and audio-visual input (Sun et al., 2024), for which faithfulness is an important notion, designing HDBs that combine multi-modal RAG with knowledge-intensive tasks, e.g. medical QA, would likely even better represent the future use of LLMs.

Regarding methodology, our desiderata (Section 2) are intended as a practical framework synthesized from observed literature gaps, not a formal or exhaustive specification. Each property is motivated by empirical evidence (D1–D2 by Tables 4 and 5; D3–D4 by Tables 4–6) or standard diversity considerations (D5–D7). Other properties, such as fine-grained labels or multi-task coverage, may also be desirable; we encourage future work to extend the desiderata as the field matures. Additionally, the ROUGE-based prefiltering strategy (ROUGE < 0.1) used to identify candidate hallucinated samples may bias the dataset toward low-overlap errors, potentially under-representing subtle hallucinations that maintain high lexical overlap with the ground-truth answer. However, stratified analysis shows that detector AUC-ROC differs minimally across ROUGE bins ($\Delta \leq 0.05$; see Appendix Table 12), suggesting the filtering does not bias toward easier-to-detect hallucinations.

Finally, while human-verified labels (D2) are the gold standard for evaluation, obtaining them is expensive and time-consuming, especially for long-context RAG tasks. LLM-as-a-judge labels offer a practical alternative but introduce sample-dependent noise (see Table 5). Dependence on proprietary LLMs for judge labels also limits the reproducibility of noise generation, though the resulting labels are released with the dataset. This cost–quality trade-off motivates both D2 (the need for human verification) and D4 (the need to study the effect of noisy labels).

8 Broader Impact & Ethical Considerations

Hallucinations generated by LLMs pose significant risks in sensitive domains such as healthcare, law, and education. Our work supports the development of more reliable and effective detection methods, potentially reducing harmful misinformation. All dataset contexts are publicly sourced and comply with their intended licenses. We acknowledge that LLM-generated content contains inaccuracies. To the best of our knowledge, there are no additional ethical concerns associated with this paper.

References

Thomson reuters. introducing ai-assisted research: Legal research meets generative ai. https://legal.thomsonreuters.com/blog/legal-research-meets-generative-ai/. Date: 2023-11-15.

Maha Tufail Agro and Hanan Aldarmaki. 2023. Handling realistic label noise in bert text classification. In *ICNLSP*, pages 11–20. Association for Computational Linguistics.

Anthropic. 2025. Prompt improver. Accessed: 2025-05-08.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

Forrest Sheng Bao, Miaoran Li, Renyi Qu, Ge Luo, Erana Wan, Yujia Tang, Weisi Fan, Manveer Singh Tamber, Suleman Kazi, Vivek Sourabh, Mike Qi, Ruixuan Tu, Chenyu Xu, Matthew Gonzales, Ofer Mendelevitch, and Amin Ahmad. 2025. FaithBench: A diverse hallucination benchmark for summarization by Modern LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 448–461, Albuquerque, New Mexico. Association for Computational Linguistics.

Federico Borra, Claudio Savelli, Giacomo Rosso, Alkis Koudounas, and Flavio Giobergia. 2024. Malto at semeval-2024 task 6: Leveraging synthetic data

for llm hallucination detection. *arXiv preprint arXiv:2403.00964*.

Kedi Chen, Qin Chen, Jie Zhou, He Yishen, and Liang He. 2024a. DiaHalu: A dialogue-level hallucination evaluation benchmark for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9057–9079, Miami, Florida, USA. Association for Computational Linguistics.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*.

Ze Chen, Chengcheng Wei, Songtan Fang, Jiarong He, and Max Gao. 2024c. Opdai at semeval-2024 task 6: Small llms can accelerate hallucination detection with weakly supervised data. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 721–729.

Derek Chong, Jenny Hong, and Christopher D. Manning. 2022. Detecting label errors by using pre-trained language models. In *EMNLP*, pages 9074–9091. Association for Computational Linguistics.

A. P. Dawid and A. M. Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *ACL (Findings)*, pages 3563–3578. Association for Computational Linguistics.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proc. of NAACL*.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. In *EMNLP*, pages 2053–2070. Association for Computational Linguistics.

Dana Greenstein. LexisNexis launches second-generation legal AI assistant on Lexis+ AI. `https://tinyurl.com/28pzjm4b`. Date: 2024-04-23.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and 1 others. 2024. The hallucinations leaderboard–an open effort to measure hallucinations in large language models. *arXiv preprint arXiv:2404.05904*.

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

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975, Miami, Florida, USA. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8135–8158, Bangkok, Thailand. Association for Computational Linguistics.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating hallucination in large language models via self-reflection. *arXiv preprint arXiv:2310.06271*.

Ling Jiang, Keer Jiang, Xiaoyu Chu, Saaransh Gulati, and Pulkit Garg. 2024. Hallucination detection in llm-enriched product listings. In *Proceedings of the Seventh Workshop on e-Commerce and NLP@ LREC-COLING 2024*, pages 29–39.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, pages 1601–1611. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summa-

rization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less annotating, more classifying–addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli.

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

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024. Enhancing LLM factual accuracy with RAG to counter hallucinations: A case study on domain-specific queries in private knowledge-bases. *arXiv preprint arXiv:2403.10446*.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.

Junliang Luo, Tianyu Li, Di Wu, Michael Jenkin, Steve Liu, and Gregory Dudek. 2024. Hallucination detection and hallucination mitigation: An investigation. *arXiv preprint arXiv:2401.08358*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*, pages 9004–9017. Association for Computational Linguistics.

Andreas Marfurt and James Henderson. 2022. Unsupervised token-level hallucination detection from summary generation by-products. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 248–261, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Timothee Mickus, Elaine Zosa, Raúl Vázquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. Semeval-2024 task 6: Shroom, a shared-task on hallucinations and related observable overgeneration mistakes. In *SemEval@NAACL*, pages 1979–1993. Association for Computational Linguistics.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *First Conference on Language Modeling*.

Timo Möller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. Covid-qa: A question answering dataset for covid-19. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics.

Imran Rahman-Jones Natalie Sherman. Apple suspends error-strewn ai generated news alerts. `https://www.bbc.com/news/articles/cq5ggew08eyo`. Date: 2025-01-17.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *ACL*, pages 10862–10878. Association for Computational Linguistics.

Jio Oh, Soyeon Kim, Junseok Seo, Jindong Wang, Ruochen Xu, Xing Xie, and Steven Euijong Whang. 2024. ERBench: An entity-relationship based automatically verifiable hallucination benchmark for large language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhenting Qi, Xiaoyu Tan, Chao Qu, Yinghui Xu, and Yuan Qi. 2023. Safer: A robust and efficient framework for fine-tuning bert-based classifier with noisy labels. In *ACL (industry)*, pages 390–403. Association for Computational Linguistics.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Elisei Rykov, Yana Shishkina, Kseniia Petrushina, Kseniia Titova, Sergey Petrakov, and Alexander Panchenko. 2024. Smurfcat at semeval-2024 task 6: Leveraging synthetic data for hallucination detection. *arXiv preprint arXiv:2404.06137*.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Guangzhi Sun, Potsawee Manakul, Adian Liusie, Kunat Pipatanakul, Chao Zhang, Philip C. Woodland, and Mark J. F. Gales. 2024. Crosscheckgpt: Universal hallucination ranking for multimodal foundation models. *CoRR*, abs/2405.13684.

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

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Raúl Vázquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sánchez-Vega, Alessandro Raganato, Jindřich Libovickỳ, and 1 others. 2025. Semeval-2025 task 3: Mu-shroom, the multilingual shared task on hallucinations and related observable overgeneration mistakes. *arXiv preprint arXiv:2504.11975*.

Yixiao Xu, Peng Li, Jianghao Hao, and Hao Wang. 2025. LUMINA: Detecting hallucinations in RAG system with context–knowledge signals. *arXiv preprint arXiv:2509.21875*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. In *EMNLP (Findings)*, pages 3898–3908. Association for Computational Linguistics.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhongxiang Zhang, Jiazheng Liang, Junhao Ji, Fei Liu, Yuren Song, Qian Wu, and Xing Cheng. 2024. ReDeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. 2022. Is bert robust to label noise? a study on learning with noisy labels in text classification. In *Insights@ACL*, pages 62–67. Association for Computational Linguistics.

Zhiying Zhu, Yiming Yang, and Zhiqing Sun. 2025. Halueval-wild: Evaluating hallucinations of language models in the wild. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

A Appendix

A.1 LLM-as-a-Judge Prompt Template

```
You will be provided a article, a
question, and an answer. Your task is to
 determine which of the following is
true:
Using only the information in the
article, you are able to verify that the
 entire contents of the answer is indeed
 correct (every part of the answer is
supported by the contents of article).

CLASSIFICATION CRITERIA:
1. NOT HALLUCINATED:
   - Information explicitly stated in
   the source material
   - Information that can be directly
   inferred from the source material
   - Appropriate supplementary responses
    that do not make specific factual
   claims

2. HALLUCINATED:
   - Any non-supplementary claim not
   supported by the source material
   - Any claim that extends beyond
   reasonable inference from the context

Here is the article: <ARTICLE> {CONTEXT}
 </ARTICLE>
Here is the question: <QUESTION> {
QUESTION} </QUESTION>
Here is the answer: <ANSWER> {ANSWER} </
ANSWER>
Carefully analyze the article, question,
 and answer. Before answering the
question, please think about it step-by-
step within <THINKING></THINKING> tags.
Then provide your final response in the
format: <RESPONSE> $ANSWER </RESPONSE>,
where ANSWER must be either
HALLUCINATION, or NO HALLUCINATION.
```

A.2 Model Configuration Details

For SFT we do LoRA (r = 16, alpha = 32, on `q_proj` and `v_proj` modules, with a dropout of 0.05.) instruction fine-tuning on train split at response level using a "mistralai/Mistral-7B-Instruct-v0.2" model. Training used a batch size of 1, gradient accumulation of 4, `bf16` precision, and DeepSpeed optimization. The learning rate was linearly scheduled with 100 warm-up steps and 0.01 weight decay. The best model was selected based on evaluation AUC every 100 steps. For TRIVIA+ due to long context length we drop 16.13% longest samples from train split. The threshold is optimized to maximize F1 score on the test set. While we acknowledge that this may yield optimistic F1 values, the choice is motivated by differing target rates between the validation and test sets and does not detract from the takeaways. The same strategy is applied consistently across all methods to ensure a fair comparison.

A.3 Detection Results for Additional Metrics

Table 7 complements Table 4 with three additional metrics for detection performance, namely, ROC-AUC and PR-AUC (area under ROC and Precision-Recall curves) as well as Accuracy.

Table 7: Detection performances on RAG-QA HDB datasets by SC-GPT (G): SelfCheckGPT with GPT-4-mini, SC-GPT (C): SelfCheckGPT with Claude-Sonnet-3.5, LLM-aaJ: LLM-as-a-Judge with Claude-Sonnet-3.5, FS: few-shot, PO: prompt-optimized, SFT: supervised fine-tuning.

Dataset	Model	ROC-AUC	PR-AUC	Acc.
HaluEval	SC-GPT (G)	0.908	0.924	0.872
	SC-GPT (C)	0.906	0.924	0.871
	LLM-aaJ	0.815	0.859	0.815
	FS	0.814	0.859	0.814
	PO	0.832	0.871	0.832
	SFT	0.999	0.999	0.996
RAGTruth	SC-GPT (G)	0.584	0.199	0.358
	SC-GPT (C)	0.586	0.199	0.383
	LLM-aaJ	0.825	0.682	0.806
	FS	0.836	0.694	0.811
	PO	0.800	0.671	0.697
	SFT	0.871	0.688	0.874
Dolly (NC)	SC-GPT (G)	0.708	0.673	0.619
	SC-GPT (C)	0.733	0.682	0.651
	LLM-aaJ	0.692	0.725	0.697
	FS	0.660	0.713	0.648
	PO	0.646	0.713	0.627
	SFT	-	-	-
TRIVIA+	SC-GPT (G)	0.729	0.547	0.636
	SC-GPT (C)	0.795	0.617	0.757
	LLM-aaJ	0.766	0.742	0.749
	FS	0.764	0.743	0.738
	PO	0.743	0.725	0.718
	SFT	0.777	0.632	0.727

A.4 More Details of TRIVIA+ Dataset Construction

To efficiently identify hallucinations given limited annotator resources, we selected candidate samples by comparing each LLM's answer to the human-written ground truth. We prioritized examples with the lowest ROUGE scores (i.e., least overlap), which increased the proportion of true hallucinations to approximately 35%. This filtering step identifies potentially hallucinatory examples without altering the LLM outputs, thereby preserving their "organic" nature.

We measured inter-annotator agreement using

Fleiss' Kappa on cases with three independent votes, obtaining a score of 0.46. This indicates fair-to-good agreement and reflects the challenging, ambiguous nature of the dataset. This difficulty motivates our focus on studying detection performance under noisy labeling conditions. The final dataset comprises 3,224 annotated instances, with 2,101 classified as supported and 1,123 as unfaithful.

A.5 Human Annotation Guidelines for TRIVIA+ Dataset

Annotation Guidelines

1. **Read the question, answer, and article.**
   - Understand the content of the question. Note that for the summarization use-case, you may not be provided with a question.
   - Read the provided answer carefully. Identify the key information, entities, and concepts that were addressed in the answer. Every detail in the answer is important for the assessment.
   - Read the article, focusing on sections relevant to the question and answer. If multiple references are provided, look for connections and relationships between the information provided by different reference materials that could help you to better assess the answer.
   - Resolve conflicting information as needed. If conflicting information is provided by different reference materials, consider factors such as the source of the information, the recency of the data, and the level of expertise or authority of the sources.
   - Pay close attention to the provided quotes and/or highlighted text from the article, but do not base your assessment solely on the quotes/highlights since they may miss important information in the full article.

2. **Highlight and tag phrases in the article.**
   - For each sentence in the provided answer, identify and highlight phrases that provide relevant details. Identify the relevant information that was used to answer the question. This could include facts, statistics, quotes, or other data points relevant to the answer.
   - If a phrase in the article supports multiple sentences, associate the phrase with the most relevant sentence, or choose the first relevant sentence if multiple sentences are equally relevant.
   - Keep the number of highlighted phrases to the minimum required to demonstrate supporting or contradicting facts.
   - If there is nothing to highlight, check the box "No entities to label".

3. **Choose a label for each sentence in the answer from the following:**

   Supported: If the information in the sentence is consistent with and supported by the information in the article.

   Contradicted: If the information in the sentence directly conflicts with information presented in the article.

   Not Mentioned: If the information in the sentence is neither confirmed nor refuted by the article.

   Supplementary: If the text provides supplementary information that does not pertain to the question, such as stating "The answer is based on my understanding of the article". Only select this label if the other labels do not apply.

4. **Select the best label even if multiple labels apply.**
   Take into account different interpretations and possible nuances and select the most relevant label. In particular, look out for these nuanced cases:

a. **Answer uses information from sections in the article that are either disjointed or truncated:** Carefully assess how each sentence in the article relates to each other, and how they should be interpreted (e.g., outdated information, tabular format) and combined. Once you have a good understanding of the relationships between the sentences/sections in the article, use the full set of information to select your label.
Example:
   - **Question:** Which team won the match between United States and Hungary in the Women's Water Polo Quarterfinal?
   - **Answer:** The US team made the most out of their extra player shots, scoring on 3 of their 5 attempts.
   - **Context:** [cite_1] United States beat Hungary 5-4 on Thursday to reach the women's water polo semifinal of the Paris Olympics.... [cite_2] The Hungarian squad managed to shoot the ball 31 times, while the Americans only got 22 shots up. The difference in the game offensively was that the US team made the most out of their extra player shots, scoring on 3
   - **Expected label:** Not mentioned.
   - **Reason:** It is not clear how many attempts were made by US during their extra player shots, we only know they got 5 out of the 22 attempted shots overall.

b. **Answer can be inferred but not explicitly mentioned in the article:** Consider if you can derive a number mathematically, paraphrase the reference information, or draw conclusions by reasoning with the knowledge provided in the article. If so, the answer can be marked as supported/contradicted depending on your derived/inferred findings, otherwise choose "not mentioned".
Example:
   - **Answer:** The company handles fulfillment logistics
   - **Context:** The company will pick, pack, ship, and provide customer service for those products....
   - **Expected label:** Supported.
   - **Reason:** The answer is a paraphrase of the context.

c. **Answer relies on some knowledge that is provided on the internet but not the article:** Check if the question is directly asking for this information, if so, you should not be using the internet for assessing the answer. If the knowledge is not directly asked for in the question, assess if the claim is about general terms and concepts, which you can safely read up to better assess the answer. If the claim is about specific events, dates, persons, do not use information from the internet to assess the answer, instead mark the case as "not mentioned".
Example:
   - **Question:** When Does 'Toy Story 5' Come Out? What We Know About New Movie?
   - **Answer:** Pete Docter, the Chief Creative Officer of Pixar, has stated that the sequel has the potential to take unexpected turns and surprise audiences.
   - **Context:** Docter defended the development of a fifth film, saying the sequel could head in unexpected directions and end up surprising audiences.
   - **Expected label:** Not men-

tioned.

- **Reason:** The name of the Chief Creative Officer of Pixar is not mentioned in the article.

d. **Answer uses specific dates but the article only mentions relative terms such as (today, tomorrow):** Check if a timestamp is provided for the reference in the article. If so, you can use the timestamp to assess if the answer is correct.

Example:

- **Answer:** The Bank of America Corp stock price today, August 08, 2024 is 39.53.
- **Context:** Written 11:29PM, Thursday, August 01 2024, PDT) [cite_3]: What Is the Bank of America Corp Stock Price Today? Monitor the latest movements within the Bank of America Corp real time stock price chart below. The Bank of America Corp stock price today is 39.53.
- **Expected label:** Contradicted.
- **Reason:** The timestamp date is August 01, 2024.

e. **Answer is clearly wrong based on some well known information, but supported by the article:** Do not penalize the answer for being faithful to the provided reference article. If the information can be supported by the article, select "Supported".

f. **Answer contains reference numbers in [], but do not match with the actual citation/url number.** Do not penalize the answer for incorrect reference numbers since these will be removed before displaying to the user. Instead, assess the answer for claims/information extracted from the references.

If multiple labels seem to be equally relevant, prioritize "Contradicted" over the rest, and "Not Mentioned" over "Supported".

If the facts are supported by the article, do not penalize the answer due to a lack of fluency or coherence or relevance when answering the question.

5. **Provide an overall assessment:**

Unanswerable question: If the question cannot be answered by the given article, check this box. This also includes cases where the article provides non-English text, or is full of special characters that make it difficult to read.

Answer not provided: Did the answer state that the question is not answerable, or indicates "I don't know"?

Note that you can select multiple checkboxes, and select relevant labels for each sentence. For example, an unanswerable question may have an answer containing information not mentioned in the article.

6. **Review, double-check, and assess the difficulty of the task:**

- Before finalizing your annotation, review the highlighted phrases and label to ensure accuracy of your annotations.
- Use similar criteria and judgement when evaluating different answers and articles.

Assess the difficulty of the task:

Not difficult: The text is easy to read and understand, and the label is obvious to me.

Quite difficult: The text is quite hard to read and I had to read carefully and/or look up the unknown terms, get familiar with the topic, and checked the article very carefully. But once I understood the text, the label is quite obvious to me.

> **Very difficult:** The text is hard to understand, and even after looking up terms, there are still multiple words in the article/question/answer that do not make sense to me. After carefully inspecting the text, it seems like multiple labels can apply.

A.6 Annotation Example and Label Aggregation

We illustrate the sentence-level annotation and response-level aggregation process with a concrete example from TRIVIA+.

Context (excerpt): *"The Battle of Gettysburg was fought July 1–3, 1863, in and around the town of Gettysburg, Pennsylvania. The battle involved the largest number of casualties of the entire war and is often described as the war's turning point."*

Question: *When was the Battle of Gettysburg fought?*

Answer (3 sentences):

1. *"The Battle of Gettysburg was fought from July 1 to July 3, 1863."* → **Supported**
2. *"It took place near Gettysburg, Pennsylvania, and resulted in over 50,000 casualties."* → **Not Mentioned** (the specific number is not stated in the context)
3. *"It is widely regarded as the turning point of the Civil War."* → **Supported**

Response-level aggregation: We apply the strictest-label rule: Contradicted > Not Mentioned > Supported. Since sentence 2 is labeled "Not Mentioned" (unfaithful), the entire response is aggregated as **unfaithful**, following the standard practice used by FACTS Grounding (Jacovi et al., 2025) and RAGTruth (Niu et al., 2024).

A.7 Stratified Analysis of TRIVIA+

We provide stratified analyses of TRIVIA+ across multiple dimensions. Tables 8–10 report hallucination rates by domain, context length, and generating LLM, respectively. Table 11 breaks down detection rate by hallucination type. Table 12 verifies that ROUGE-based filtering does not bias detection difficulty.

A.8 More Details of ReDEEP

Finally, Table 13 reports results for the unsupervised ReDeEP detector.

Table 8: Hallucination rate by domain in the TRIVIA+ test split.

Domain	N	Hall. Rate	95% CI
COVID	139	14.4%	[9.5%, 21.2%]
MS-MARCO	763	18.1%	[15.5%, 21.0%]
NQ	674	29.4%	[26.1%, 32.9%]
TriviaQA	309	31.7%	[26.8%, 37.1%]
DROP	1,339	50.0%	[47.3%, 52.6%]

Table 9: Hallucination rate by context length in the TRIVIA+ test split.

Context Length	N	Hall. Rate	95% CI
Short (<1K)	886	40.0%	[36.8%, 43.2%]
Medium (1K–5K)	1,355	36.6%	[34.1%, 39.2%]
Long (>5K)	983	27.8%	[25.1%, 30.7%]

Table 10: Hallucination rate by generating LLM in the TRIVIA+ test split.

LLM	N	Hall. Rate	95% CI
SOTA	1,006	18.4%	[16.1%, 20.9%]
Gemma	532	41.2%	[37.1%, 45.4%]
Mixtral-8x7B	1,686	42.6%	[40.3%, 45.0%]

Table 11: Detection rate (%) by hallucination type on TRIVIA+. Most detectors show no significant difference (Fisher's exact test, $p > 0.05$) between Contradicted and Not Mentioned types, except SC-GPT (G).

Method	Contradicted	Not Mentioned	Fisher p
SFT	78.2	70.4	0.46
SC-GPT (C)	74.6	59.3	0.11
SC-GPT (G)	86.8	59.3	0.001
LLM-aaJ	83.2	74.1	0.28
FS	85.3	81.5	0.57
PO	82.7	81.5	0.79

Table 12: AUC-ROC stratified by ROUGE score bin on the TRIVIA+ test split. Detector performance shows minimal difference across bins ($\Delta \leq 0.05$), suggesting the ROUGE-based filtering does not bias toward easier-to-detect hallucinations.

ROUGE bin	Halluc%	SFT	SC-GPT(C)	SC-GPT(G)
$[0, 0.01)$	37%	0.775	0.791	0.722
$[0.01, 0.1)$	20%	0.723	0.772	0.722

ReDeEP (Zhang et al., 2024) is an unsupervised hallucination detector that combines two signals extracted from a single forward pass of a LLaMA backbone: an attention-based external context score (ECS) and a parametric knowledge score (PKS) derived from divergence between in-

termediate and final layer logits. We focus on the chunk version as it shows superior performances in (Zhang et al., 2024). RAGTruth contains responses from six LLMs, including LLaMA-2-7B and 13B. This enables two evaluation settings: *self-detect*, where each backbone processes only the 450 test responses generated by that same model (the default in the official implementation[8]), and *all responses*, where the backbone processes all 2,700 test responses regardless of the generating model. Since TRIVIA+ contains no LLaMA-generated responses, only the latter setting applies. On RAGTruth, the AUC drop from self-detect to all-response is modest for 7B (0.747 vs. 0.737) but larger for 13B (0.798 vs. 0.727), suggesting the 13B parametric knowledge signal is less informative for text generated by other models. ReDeEP requires storing full attention matrices (`output_attentions=True`), which scales quadratically with sequence length. On TRIVIA+, 14% (7B) to 22% (13B) of test responses exceed the memory of a single NVIDIA RTX PRO 6000 GPU (95 GB) and cannot be processed. These responses receive an uninformative default score of 0.5 for evaluation; all 645 test responses are included in the reported metrics. This structural limitation further highlights the challenge that long-context benchmarks like TRIVIA+ pose for existing detection methods.

Table 13: ReDeEP detection performance. *Self-detect*: backbone processes only responses from the same model (450 per model). *All*: backbone processes all test responses.

Backbone	RAGTruth Self-detect (450 resp.)	RAGTruth All (2,700 resp.)	TRIVIA+ All (645 resp.)
AUC-ROC			
LLaMA-2-7B	0.747	0.737	0.593
LLaMA-2-13B	0.798	0.727	0.600
F1 / Precision / Recall			
LLaMA-2-7B	0.717 / 0.660 / 0.783	0.630 / 0.597 / 0.667	0.535 / 0.405 / 0.787
LLaMA-2-13B	0.731 / 0.686 / 0.783	0.612 / 0.535 / 0.715	0.531 / 0.368 / 0.957

A.9 Annotation UI

We illustrate our annotation UI in Figure 3.

[8] https://github.com/Jeryi-Sun/ReDEeP-ICLR, accessed April 2026.

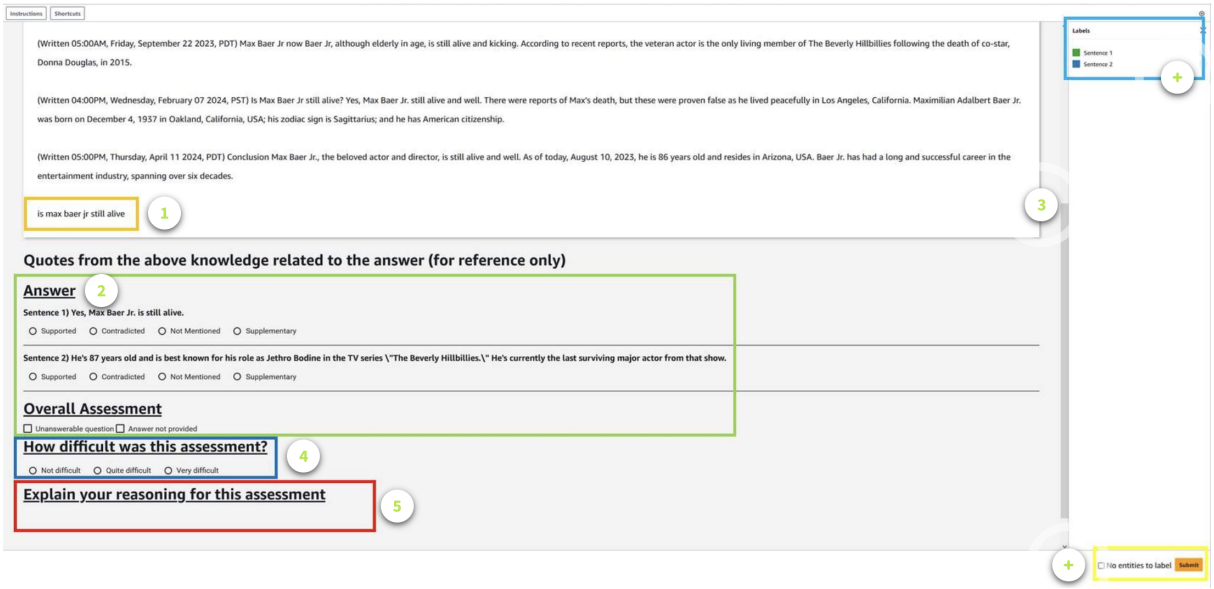

Figure 3: UI for human annotators.