ANAH: Analytical Annotation of Hallucinations in Large Language Models

Ziwei Ji[1,2*] Yuzhe Gu[1*] Wenwei Zhang[1†] Chengqi Lyu[1] Dahua Lin[1,3] Kai Chen[1†]
[1]Shanghai AI Laboratory [2]Hong Kong University of Science and Technology
[3]The Chinese University of Hong Kong
zjiad@connect.ust.hk {guyuzhe,zhangwenwei,lvchengqi,chenkai}@pjlab.org.cn

Abstract

Reducing the '*hallucination*' problem of Large Language Models (LLMs) is crucial for their wide applications. A comprehensive and fine-grained measurement of the hallucination is the first key step for the governance of this issue but is under-explored in the community. Thus, we present ANAH, a bilingual dataset that offers **AN**alytical **A**nnotation of **H**allucinations in LLMs within Generative Question Answering. Each answer sentence in our dataset undergoes rigorous annotation, involving the retrieval of a reference fragment, the judgment of the hallucination type, and the correction of hallucinated content. ANAH consists of ∼12k sentence-level annotations for ∼4.3k LLM responses covering over 700 topics, constructed by a human-in-the-loop pipeline. Thanks to the fine granularity of the hallucination annotations, we can quantitatively confirm that the hallucinations of LLMs progressively accumulate in the answer and use ANAH to train and evaluate hallucination annotators. We conduct extensive experiments on studying generative and discriminative annotators and show that, although current open-source LLMs have difficulties in fine-grained hallucination annotation, the generative annotator trained with ANAH can surpass all open-source LLMs and GPT-3.5, obtain performance competitive with GPT-4, and exhibits better generalization ability on unseen questions.[1]

1 Introduction

Large Language Models (LLMs) have achieved significant performance improvements across a diverse array of Natural Language Processing tasks (Petroni et al., 2021; Kamalloo et al., 2023; Sun et al., 2023; Chen et al., 2023, 2024). However, LLMs still face a worrisome problem that significantly hinders their real-world applications, *hallucination*, in which they produce plausible-sounding but unfaithful or nonsensical information (Ji et al., 2022; Bang et al., 2023) when answering the user questions, especially those require intensive knowledge. Given the fluency and convincing nature of the responses produced by LLMs, the detection of their hallucinations becomes increasingly difficult (Adlakha et al., 2023; Ren et al., 2023; Pezeshkpour, 2023). Such a challenge impedes the deep analysis and reduction of LLM hallucination and leads to extensive dissemination of misleading information as the user base widens and real-world applications proliferate (Mallen et al., 2023).

There have been extensive efforts on effectively

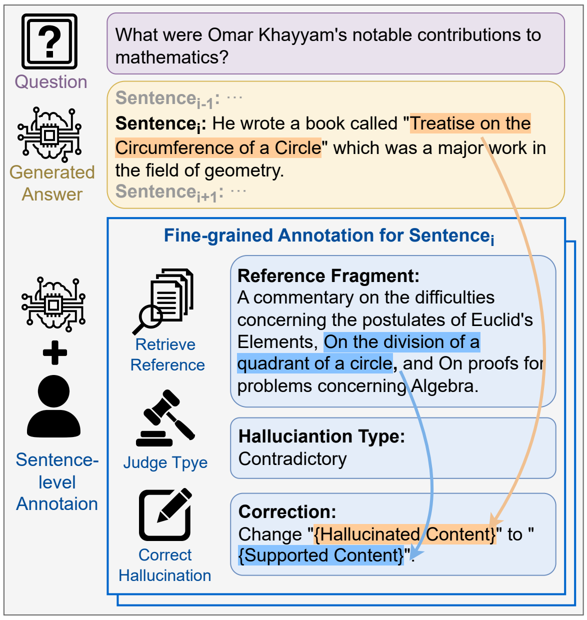

Figure 1: An example of ANAH for sentence-level hallucination annotation. Each sentence in a generated answer is annotated in fine-grained with Reference Fragment, Hallucination Type, and Correction. The hallucinated and supported content are highlighted in orange and blue, respectively.

*Equal contributions
†Corresponding author
[1]Please find the dataset, code, and model at https://github.com/open-compass/ANAH.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8135–8158
August 11-16, 2024

detecting and evaluating hallucination (Durmus et al., 2020; Mündler et al., 2023; Du et al., 2023a). However, most benchmarks were proposed before the advent of LLM and targeted specific English tasks (Dziri et al., 2021; Rohrbach et al., 2018), which are not challenging for current models. Recent benchmarks (Li et al., 2023a, 2024) for LLMs only categorize whether the entire response contains hallucinations without explanation and reference. This coarse-grained nature makes it difficult to trace the exact trigger of hallucinations and obstructs further mitigation of them.

Therefore, we establish a novel large-scale Chinese-English benchmark, named **ANAH**[2], that assesses the LLMs' ability to annotate the LLM hallucinations sentence-by-sentence, in the scenario of knowledge-based generative question answering. Rather than solely result-oriented, for each answer to a question, our approach prompts the model to annotate hallucination for **each sentence**, including retrieving **reference fragment** for the sentence, judging the **hallucination type** (No/Contradictory/Unverifiable Hallucinations, and No Fact), and **correcting** the sentence based on the reference fragment if hallucination exists (Fig. 1).

To facilitate the scale-up of datasets, we ensure the comprehensiveness and diversity of ANAH across various topics, questions, and answers. As shown in Fig. 2, first, we curate topics in both English and Chinese, encompassing a broad domain range including things, places, people, and historical events (Fig. 3). Second, we craft around three related questions for each topic to ensure originality and avoid contamination. Third, for each question, we construct a high-quality and a low-quality response with and without reference in generation, respectively, enabling a comparative analysis of hallucination distributions across different response scenarios. The final and pivotal stage is fine-grained hallucination annotation, as exemplified in Fig. 1. Eventually, we form ∼12k hallucination annotations of ∼4.3k answers to ∼2.2k questions spanning a broad domain range, which is challenging for hallucination detection.

Thanks to the completeness and fine-granularity of ANAH, the statistical results of the hallucination annotations quantitatively confirm that hallucinations progressively accumulate in the LLM responses. Furthermore, ANAH can be used to

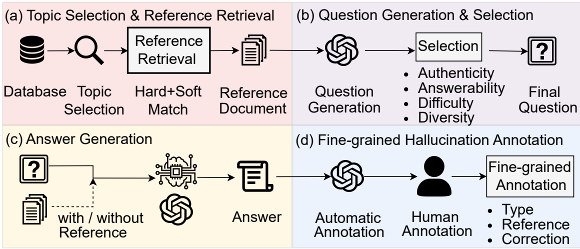

Figure 2: The overview of dataset establishment, comprising (a) Topic Selection and Reference Retrieval, (b) Question Generation and Selection, (c) Answer Generation, and (d) Fine-grained Hallucination Annotation.

train and evaluate hallucination annotators. We first discovered that only GPT-4 could do this task well. Thus, we further investigate training generative and discriminative hallucination annotators using ANAH and observe the advantages of generative annotators over discriminative annotators in handling the imbalance issue of hallucination types. Remarkably, our generative annotators achieve an accuracy of 81.01%, surpassing open-source models and rivaling GPT-4 (86.97%) in performance with a smaller size and lower source cost. We also observe that the hallucination annotators consistently exhibit better generalization regarding the number of questions than the breadth of topics, thereby guiding us toward prioritizing data scaling to cover a broader array of topics in future research.

2 Dataset Construction

ANAH's establishment contains four stages (Fig. 2): (1) selecting a broad range of topics to ensure comprehensiveness (§ 2.1), (2) constructing related questions whose responses can be fully supported by reference (§ 2.2), (3) generating answers from LLMs under different models and scenarios (§ 2.3), and (4) fine-grained hallucination annotation for further analysis and mitigation (§ 2.4).

2.1 Topic Selection and Reference Retrieval

The initial stage involves the selection of topics and corresponding references from knowledge-intensive datasets. To ensure diversified and wide-ranging information, our topic choices are categorized into celebrities, events, locations, and things. We also encompass various domains, including but not limited to Politics and Military, Art, Science and Technology, Religion, *etc.* (Fig. 3). Topics are meticulously chosen based on the frequency of their occurrence via Google Ngram Viewer [3]

[2] **ANAH** is short for **AN**alytical **A**nnotation of **H**allucinations.

[3] https://books.google.com/ngrams/

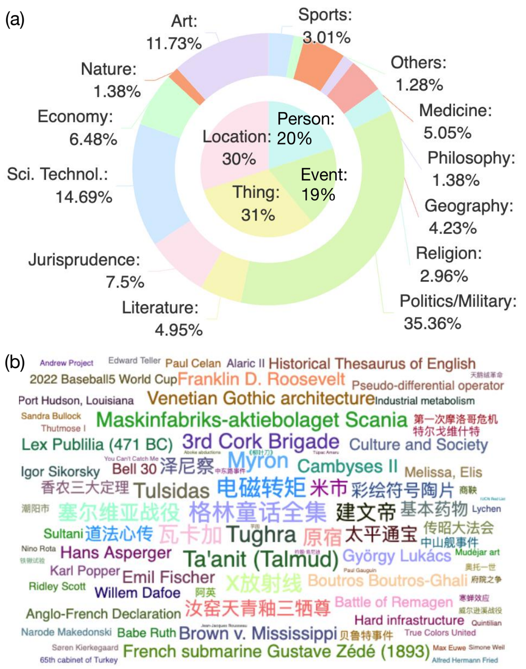

Figure 3: The topic distribution by chart of (a) categories (inner) and domains (outer), and (b) word cloud.

since topics that more frequently occur and are of public interest should be more important for real-world applications of LLMs. We also collect topics from publicly available summaries like historical timelines and the ranking of influential persons [4].

After selecting the topics, their corresponding reference documents are retrieved from pre-training databases (He et al., 2022), including Wikipedia[5], Baidu Baike[6], Encyclopedia Britannica[7]. We select the datasets that have been widely used in the pre-training stage of LLMs (Touvron et al., 2023) so that we can make sure that the model saw the truth, which is important for further analysis and mitigation of hallucinations.

During the reference retrieval process, the discrepancies in nomenclature across different sources and the potential of a single name having multiple meanings present challenges. To address these challenges, we adopt a strategy that progresses from hard to soft matching. First, we perform exact matching (i.e., hard matching) of the entries. Then, we sort the candidate entries according to the sentence semantic similarity and further judge them with InternLM (Team, 2023) to select the correct ones [8]. Finally, manual filtering is performed to iron out the problem of renaming. Overall, this phase establishes a robust foundation for the ensuing steps of benchmark construction.

2.2 Question Generation and Selection

The second stage involves the generation and selection of several questions based on the provided reference documents about a particular topic. To increase the possibility that the data is unseen and untainted, we create new questions rather than repurposing existing datasets. The questions are framed in a manner so that they can be fully answered exclusively grounded on the provided reference documents, avoiding being overly subjective or open-ended. To ensure diversity and comprehension across questions, they are designed to cover different types, such as 'what', 'when', 'where', 'why', etc, and perspectives such as descriptions, explanations, reasons, etc., encapsulating all facets of the information. The questions also traverse diverse levels of knowledge, ranging from basic, generic knowledge to more intricate, specialized knowledge or domain-specific expertise. The generation prompt is shown in Fig. A2.

To assure the uniqueness of each question and avoid duplication, we leverage CoSENT [9] for Chinese and MiniLM [10] for English, respectively, to calculate similarities among questions and sift out overly similar ones[11]. We then employ GPT-3.5 (OpenAI, 2023) to assess their answerability, *i.e.*, whether the given questions can be answered based solely on the provided reference documents. This ensures that the questions are fact-based, objective, and possess a definitive answer, thus increasing the reliability and consistency of the evaluation process. The prompt details are in Fig. A3.

Finally, we utilize GPT-4 (OpenAI, 2023) to select the top three questions from the above candidate questions, considering the following characteristics:

1. High authenticity: The questions should be free from any intentionally misleading, ambiguous, or false information.
2. High answerability: The questions exhibiting excessive subjectivity, controversy, or predic-

[4] For example, https://en.wikipedia.org/wiki/Timeline_of_Chinese_history and https://pantheon.world/explore/rankings.

[5] https://www.wikipedia.org/

[6] https://baike.baidu.com/

[7] https://www.britannica.com/

[8] Please refer to Fig. A1 for the prompt.

[9] https://huggingface.co/shibing624/text2vec-base-chinese

[10] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

[11] Pleaser refer to Appendix A for details.

Topic	Omar Khayyam
Reference	Ghiyāth al-Dīn Abū al-Fatḥ Umar ibn Ibrāhīm Nīsābūrī, commonly known as Omar Khayyam, was a polymath, known for his contributions to mathematics, astronomy, philosophy, and [0] Persian poetry... [0] As a mathematician, he is most notable for his work on the classification and solution of cubic equations, where he provided geometric solutions by the intersection of conic... ## Mathematics Khayyam was famous during his life as a mathematician. [1] His surviving mathematical works include: A commentary on the difficulties concerning the postulates of Euclid's Elements, On the division of a quadrant of a circle, and On proofs for problems concerning Algebra... ### Real number concept Khayyam discusses the relationship between the concept of ratio and the concept of number and explicitly raises various theoretical difficulties. [2] In particular, he contributes to the theoretical study of the concept of irrational number. ...
Question	What were Omar Khayyam's notable contributions to mathematics?
Answer	[0] Omar Khayyam was a mathematician and poet who is best known for his contributions to algebra and geometry. [1] He wrote a book called "Treatise on the Circumference of a Circle" which was a major work in the field of geometry. [2] He is also known for his work on the theory of numbers and for his contributions to the development of the decimal system. [3] I hope my reply is helpful.
Sent 0 Annotation	**<Reference>** Ghiyāth al-Dīn Abū al-Fatḥ Umar ibn Ibrāhīm Nīsābūrī, commonly known as Omar Khayyam, was a polymath, known for his contributions to mathematics, astronomy, philosophy, and Persian poetry.<SEP> As a mathematician, he is most notable for his work on the classification and solution of cubic equations, where he provided geometric solutions by the intersection of conic. **<Halluciantion>** None
Sent 1 Annotation	**<Reference>** A commentary on the difficulties concerning the postulates of Euclid's Elements, On the division of a quadrant of a circle, and On proofs for problems concerning Algebra. **<Halluciantion>** Contradictory **<Correction>** "Treatise on the Circumference of a Circle" to "On the division of a quadrant of a circle".
Sent 2 Annotation	**<Reference>** In particular, he contributes to the theoretical study of the concept of irrational number. **<Halluciantion>** Unverifiable **<Correction>** "and for his contributions to the development of the decimal system." to "".
Sent 3 Annotation	**<No Fact>**

Table 1: Examples of fine-grained hallucination annotation for each sentence in an answer. Related fragments for each sentence in reference are marked in the same colors (purple, blue, green, and grey for sentence 0, 1, 2, and 3, respectively).

tive nature should be excluded.

3. Difficulty: A certain level of difficulty should be guaranteed.
4. High diversity: Enhancement of overall diversity in terms of type, complexity, depth of knowledge, *etc*. Similar questions should be discarded.

The question selection prompt is in Fig. A4. This meticulous process of question generation and selection not only ensures the quality of the benchmark but also elevates its value in testing the model hallucinations.

2.3 Answer Generation

The third stage involves generating answers for each question with different LLMs. In this case, we use GPT-3.5 with a reference document to construct a high-quality answer and an early version of InternLM-7B without reference to generate a low-quality answer, respectively. Such a design allows to evaluation of the LLMs' hallucination annotation capability under different scenarios comprehensively. Please refer to Fig. A5 for details of answer generation with reference.

2.4 Fine-grained Hallucination Annotation

The final stage involves fine-grained hallucination annotation for the answers to each question generated in the previous stages. As shown in Tab. 1, we provide the annotators with documents on a specific topic and a related question. For each answer sentence, the complete annotation includes finding the exactly related reference fragments, assessing the hallucination type, and correcting the hallucinations accordingly. To reduce the extensive time and human labor[12] and keep accuracy, we adopt GPT-4 (OpenAI, 2023) for preliminary annotation, followed by the verification and refinement of human annotators.

Specifically, we first apply existing retrieval methods to determine a document window for

[12]typically 20 minutes per answer per annotator.

Language	# Topic	# Ans	# Sent	# Token (w/,w/o Ref)
English	476	2,626	6,606	4.1M / 642K
Chinese	324	1,772	5,582	2.8M / 683K

Table 2: Number of topics, annotated answers, annotated sentences, and tokens (with and without reference documents) for each language of ANAH.

each answer sentence that accurately encapsulates related information. We empirically choose BM25 (Robertson et al., 2009) for both language, and further apply two CoSENT mdoels[13] for Chinese, and MiniLM[14] for English, to rank reference fragments. The ensemble of multiple embedding models significantly improves retrieval accuracy, which serves as a foundation for accurate hallucination-type classification and hallucination correction and reduces the cost of human annotators to correct the reference fragment. Furthermore, to optimize resource utilization of GPT-4 without compromising the annotation accuracy, we empirically determine the context length of reference fragments to be 540 tokens for Chinese and 400 tokens for English. For the remaining unverifiable sentences due to the failure of retrieval, we extend the window length by sixfold for secondary annotation and finally fix the remaining cases after secondary annotation by human annotation.

Based on the document window for each answer sentence, GPT-4 is prompted to identify reference fragments and assess whether hallucinations exist. If the sentence contains factual information and aligns with the reference, its type is 'No Hallucination'. Annotators should also pinpoint the specific reference fragments from the original documents. If the sentence contradicts the reference, its type is 'Contradictory Hallucination'. The specific reference fragments and a suggestion on correcting the response are required. If the sentence lacks supporting evidence and cannot be verified, its type is 'Unverifiable Hallucination' and a revision suggestion is required. If the sentence does not contain any factual information for evaluation, it falls under the category of 'No Fact' without further annotation. See detailed GPT-4 prompts in Fig. A6. After preliminary annotation, human annotation is conducted following a similar workflow.

[13] https://huggingface.co/shibing624/text2vec-base-chinese and https://huggingface.co/shibing624/text2vec-bge-base-chinese

[14] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Hallucination Type				Ref	Corr.
None	Cont.	Unver.	N.F.		
90.19	83.70	75.69	28.67	85.37	78.98

Table 3: Consistency between GPT-4 and human Annotations , where 'Cont.', 'Unver.', 'N.F.', 'Ref.', and 'Corr.' are abbreviations of Contradictory, Unverifiable, No Fact, Reference, and Correction, respectively.

2.5 Dataset Statistics

Eventually, our dataset covers both English and Chinese and comprises over 700 topics, $\sim$4.3k annotated answers, $\sim$12k annotated sentences, and $\sim$7M tokens with reference documents (Tab. 2). The topics also cover celebrities, events, locations, and things, from an array of domains, such as military/politics, health/medicine, and sports, as depicted in Fig. 3. The statistics underscore the comprehensiveness and extensive scale of our dataset.

We also verify the quality of GPT-4 generated annotations by analyzing their consistency with human annotations (the higher, the better). As shown in Tab. 3, the average consistency is 86.97% for hallucination type, 85.37% for reference, and 78.98% for correction. GPT-4 tends to erroneously annotate sentences as 'No Fact' when sentences contain referential ambiguity or summary discussion, while the type of 'No Fact' only accounts for $\sim$2% of annotated sentences. We provide inconsistent examples in §B.

Tab. 4 presents the proportions of hallucination type for answers generated by GPT-3.5 with reference and InternLM without reference. The hallucination proportions for answers generated with reference are much higher than those without. Such an observation which is consistent with recent research interests in retrieval augmented generation (RAG) (Lewis et al., 2020).

Accumulation Effect Thanks to the fine granularity of ANAH, we can quantitatively analyze the accumulation or snowball effect of hallucinations (Zhang et al., 2023). The probability of hallucinations occurring in the current sentence when the previous sentences contain hallucinations, $P(H_t|H_{[0:t-1]})$, is defined as

$$P(H_t|H_{[0:t-1]}) = \frac{P(H_t, H_{[0:t-1]})}{P(H_{[0:t-1]})}, \quad (1)$$

$$\text{where} \quad H_{[0:t-1]} = \exists t' \in [0:t-1] : H_{t'}.$$

H_t is a Boolean indicator that returns true if the current sentence is hallucinated. The hallucination probability is **58.51%** for English and **52.54%** for Chinese, while the hallucination probability when

Lang		None	Cont.	Unver.	N.F.
EN	w/ Ref	89.94	3.35	5.48	1.23
	w/o Ref	41.31	24.07	32.94	1.68
ZH	w/ Ref	74.86	8.04	16.05	1.05
	w/o Ref	31.82	28.07	35.86	4.25

Table 4: Proportion of each annotation type for answers generated with and without reference in English and Chinese.

the previous sentences don't contain, $P(H_t| \sim H_{[0:t-1]})$, is **14.61%** for English and **17.2%** for Chinese. $P(H_t|H_{[0:t-1]})$ is significantly higher than $P(H_t| \sim H_{[0:t-1]})$ indicates that the probability of hallucinations increases when the previous sentences contain hallucinations compared to when there are not, which quantitatively confirms the accumulation effect of hallucinations.

3 Hallucination Annotator

Taking advantage of the rich fine-grained annotations in ANAH, we explore training and evaluating both generative and discriminative annotators. The generative annotator generates textual annotations including reference fragments, hallucination type, and correction; while the discriminative annotator only focuses on discriminating hallucination type.

3.1 Generative Annotator

We adopt the same pipeline and prompts as the preliminary annotation of GPT-4 for the generative annotator. We first comprehensively analyze the current open-source and close-source LLMs' ability to generate fine-grained hallucination annotation using ANAH. Specifically, consistency with humans is assessed through the examination of an array of multilingual LLMs including Llama2 (Touvron et al., 2023), InternLM2, Qwen (Bai et al., 2023), Baichuan2 (Baichuan, 2023) in different sizes, GPT-3.5, and GPT-4.

In addition, we explore training hallucination annotators using InternLM on our dataset. The fine-grained annotation involves constructing multiple sentence annotations from each answer. When constructing the training data, each sentence from an answer forms a sample.

Data Augmentation We perform a multi-task setting where besides fine-grained hallucination annotation, we incorporate other tasks including question generation, question selection, answer generation from intermediate products of ANAH, and dialogue generation from ShareGPT (None, 2023) and Dolly (Conover et al., 2023). In addition, we apply prompt augmentation by the design of multiple prompts with varying instruction descriptions, relative locations of reference and question, etc. Please refer to § A.4 for details.

3.2 Discriminative Annotator

Recent works (Wu et al., 2023; Lightman et al., 2023; Uesato et al., 2022) explore process-supervised reword models to provide fine-grained signals in RLHF, which are also useful in hallucination mitigation process such as RLHF (Wu et al., 2023). Thus, we also explore training a sentence-level process-supervised discriminative annotator using InternLM, based on ANAH, which has the potential to be applied for fine-grained RLHF.

Following the sentence-level information including references and hallucination type of ANAH, the model is trained to categorize each sentence into one of four types: No/Contradictory/Unverifiable Hallucination, and No Fact. To enable process supervision and reuse the learned knowledge in LLMs, we replace the last layer of the pre-trained LLM with a four-category linear layer and load the remaining parameters of pre-trained LLMs for further training the annotators. This approach ensures that the scoring results are compatible with reward models in various aspects, including relevance and completeness (Wu et al., 2023). Additionally, the inference time of the discriminative annotator is significantly shorter than that of its generative counterparts.

4 Experiments

4.1 Implementation

Data Split ANAH is divided into training and testing sets. To investigate the direction of annotator generalization and dataset scaling, we further divide the testing set equally into unseen-topic and unseen-question groups. In the unseen-topic test set, the topics and corresponding references, questions, and answers remain unexposed during training. In the unseen-question test set, the topics have been exposed during training, while the questions remain unexposed.

Further details regarding the experimental implementation can be found in § C.1 for generative annotator and § C.2 for discriminative annotator.

4.2 Evaluation Protocols

For the hallucination type predicted by generative and discriminative annotators, we utilize **F1** and

Model	F1 ↑	ACC ↑	R ↑	BERT ↑	Pre4 ↑
GPT-3.5	48.01	47.94	29.4	78.78	64.25
GPT-4	87.11	86.97	86.32	96.21	86.44
Qwen-7B	8.46	4.67	24.28	77.28	44.89
Baichuan2-7B	9.63	5.50	4.21	10.65	39.82
LLama2-7B	13.76	8.31	4.37	19.93	8.26
InternLM2-7B	12.44	12.34	9.54	64.19	55.72
Qwen-14B	14.94	8.82	10.53	55.2	85.65
Baichuan2-13B	42.17	38.04	23.39	75.27	36.9
LLama2-13B	8.55	4.80	5.15	20.16	13.65
InternLM2-20B	61.49	63.17	46.36	84.68	94.93
Qwen-72B	58.27	55.69	35.96	79.21	77.19
Llama2-70B	18.42	12.53	7.13	20.95	43.31
ANAH-7B	78.69	79.92	58.51	87.27	94.90
ANAH-20B	80.49	81.01	58.82	88.44	94.86

Table 5: Automatic evaluation results for generative hallucination annotators based on different models, where 'R', 'BERT', and 'Pre4' refer to 'RougeL', 'BERTScore', and '4-gram Precision', respectively.

Accuracy to measure the quality of predicted categorization. As discriminative annotators can only classify hallucination types, we only evaluate reference fragments and corrections predicted by generative annotators and employ **RougeL** (Lin, 2004) and **BertScore** (Zhang* et al., 2020) to compare the generated text with gold-standard human reference in terms of gram, continuity, order and semantics. Since we aspire that the reference sentence predicted by generative annotators originate from the document, we also apply **n-gram Precision** to reflect fidelity to the source information.

4.3 Overall Results

Generative Annotator The results on the whole testing set in Tab. 5 show current open-source LLMs and GPT-3.5 struggle to follow the instructions to annotate hallucination in a fine-grained manner, while GPT-4 exhibits high consistency with humans. Consequently, we train our hallucination annotators utilizing the train split of ANAH. Remarkably, our ANAH-20B achieves an F1 of 80.49% and an accuracy of 81.01%, surpassing open-source models and rivaling GPT-4 in performance with a smaller size and lower source cost. We notice our model exhibits higher Precision but lower RougeL than GPT-4, indicating fidelity to the original documents but inaccurate identification of reference fragments and correction. Please refer to Tab. A4 for topic-specific analysis of ANAH-7B in Appendix D.

Discriminative Annotator Tab. 6 shows the F1 and the accuracy of the discriminative annotator is

Setting	F1↑		ACC↑		RougeL↑		Pre4↑	
	T	Q	T	Q	T	Q	T	Q
G-7B	75.93	77.24	77.89	78.12	58.02	57.76	95.62	95.17
G-20B	79.82	81.18	80.21	81.81	56.01	61.62	94.97	94.77
D-7B	66.20	68.53	69.15	70.86	-	-	-	-
D-20B	69.74	73.98	72.10	75.95	-	-	-	-

Table 6: Evaluation results for generative and discriminative annotators, noted by 'G' and 'D', respectively. 'T' represents the unseen-topic test set, while 'Q' represents the unseen-question test set. [15]

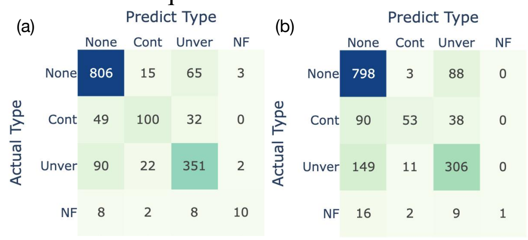

Figure 4: Hallucination Type Confusion Matrices for InternLM2-20B-based generative annotator (a) and discriminative annotator (b).

relatively lower than that of the generative annotator. Thus, we analyze the confusion matrices of hallucination type for both annotators. Fig. 4 shows the discriminative annotator is more prone to misjudge into the largest category (No Hallucination), with the 2nd to 4th row of the 1st column totaling 255, exceeding 147 for generative annotator, given the data imbalance issue depicted in Tab. 4. This suggests the current discriminative annotators are more affected by the imbalance issue of hallucination types and require further modification for improvements, which we leave for future research. Refer to § D for all confusion matrices.

Generalization Analysis Tab. 6 also indicates both generative and discriminative annotators perform better on the unseen-question test set than the unseen-topic test set in the hallucination-type classification task. This suggests leveraging prior knowledge learned from the same topic in training aids in handling exposed references in testing. This implies extending the breadth of topics has higher priority than extending questions of the same topic when scaling the data sizes of hallucination annotation in the future. In addition, we assess the generalization of ANAH annotator to other LLMs (e.g. Qwen-7B, Baichuan2-7B) in Appendix D.1.

4.4 Ablation Study

Data Augmentation As shown in the first two rows of Tab. 7, results are superior in the mix-task setting (introduced in § 3.1) compared to the single-task

[15]Due to the space limit, we put BERTScore in Tab A5.

[16]Due to the space limit, we put BERTScore in Tab A6.

Setting	F1↑		ACC↑		RougeL↑		Pre4↑	
	T	Q	T	Q	T	Q	T	Q
S.T.	75.93	77.24	77.89	78.12	58.02	57.76	95.62	95.17
M.T.	76.55	80.18	78.15	81.04	51.49	58.46	95.26	94.54
above + D.	74.62	78.33	69.97	76.48	52.18	56.78	95.06	95.33
M.T.+ P.A.	77.51	80.64	78.41	81.42	58.09	58.93	94.88	94.91
above + D.	76.8	80.44	77.76	81.30	57.98	58.99	94.72	94.93

Table 7: Ablation Study for Generative Annotator based on InternLM-7B in different settings. Here, 'S.T.' means single-task training, which only includes hallucination annotation task in training, while 'M.T.' adopts multi-task training, which further encompasses several generative tasks. "+ D" indicates that testing the annotations with prompt disturbance *i.e.*, the instructions used in testing are unseen in training. "P.A." indicates prompt augmentation is adopted in training. [16]

Model	F1 w/ Ref		ACC w/ Ref		F1 w/o Ref		ACC w/o Ref	
	T	Q	T	Q	T	Q	T	Q
G-7B	75.93	77.24	77.89	78.12	52.86	55.84	57.34	58.69
G-20B	79.82	81.18	80.21	81.81	58.06	59.95	59.51	61.2
D-7B	66.20	68.53	69.15	70.86	57.24	59.84	60.15	61.32
D-20B	69.74	73.98	72.10	75.95	60.26	61.85	63.75	64.37

Table 8: Evaluation results for generative and discriminative annotators. Here, "w/ Ref" means providing reference documents when annotating, while "w/o Ref" means without reference documents.

setting. This suggests that LLMs benefit from the multi-task shared representations and instruction-following ability.

In addition, to evaluate the robustness of generative annotators, we introduce disturbance by altering the test instruction descriptions, ensuring they differ from the training instructions. We compare the results obtained without and with prompt augmentation without and with disturbance in the last four rows of Tab. 7. The model trained with prompt argumentation declines due to perturbations, less than that with augmentation (0.39% vs. 6.37% in accuracy). It reveals models trained on diverse prompt formats increase robustness compared to their single prompt format-trained counterparts.

Reference We further examine the effectiveness of reference documents to the performance of the generative and discriminative annotators when judging the hallucination type. We test the annotators by compelling the model to rely solely on its parametric internal knowledge without any references. Tab. 8 reveals that only relying on its parametric knowledge decreases the prediction F1 and accuracy, indicating the importance of reference in annotating hallucinations.

5 Related Work

Hallucination Benchmarks can be broadly divided into two categories. One type of benchmark mainly constructs challenging queries in one/-multiple tasks and then evaluates the hallucination level in the responses (Lin et al., 2022; Dziri et al., 2022a,b, 2021; Rohrbach et al., 2018; Li et al., 2024). There are also domain-specific benchmarks curated recently, such as sports (Elaraby et al., 2023) and medical (Umapathi et al., 2023) domains. Besides these English benchmarks, a Chinese benchmark, HalluQA (Cheng et al., 2023), designs 450 adversarial questions spanning multiple domains. While these benchmarks lean toward arising hallucinations, ANAH aims to provide an analytical framework for hallucination annotation.

Another type of benchmarks can be used to train a hallucination detector/annotator and evaluate the hallucination level via the detector/annotator (Liu et al., 2021; Dziri et al., 2022a; Gupta et al., 2022; Laban et al., 2022; Durmus et al., 2020; Wang et al., 2020; Li et al., 2023a; Varshney et al., 2023; Yang et al., 2023; Muhlgay et al., 2023). All these works classify the whole response of LLMs as either hallucinatory or not. Such a coarse-grained nature makes it difficult to conduct more detailed statistical analysis. On the contrary, ANAH annotates hallucination for each sentence to different hallucination types with correction based on the retrieved reference documents. Furthermore, ANAH collects natural responses from LLMs instead of artificially guiding LLMs to produce hallucinatory responses (Li et al., 2023a; Muhlgay et al., 2023).

Hallucination Mitigation In the training stage, various techniques such as multi-task learning (Weng et al., 2020; Garg et al., 2019), model editing (Daheim et al., 2023; Ji et al., 2023a), and fine-grained RLHF (Wu et al., 2023) are proposed to mitigate hallucination. For inference time mitigation, different decoding strategies (Rebuffel et al., 2022; Chuang et al., 2023; Shi et al., 2023; Li et al., 2023b) are attempted. There are also multi-agent methods (Du et al., 2023b) and variants of the Chain-of-Thought approach involving verification or reflection (Dhuliawala et al., 2023; Lei et al., 2023; Ji et al., 2023b; Wang et al., 2023) proposed for LLMs. The hallucination annotators trained on ANAH have the potential to be integrated into the training and inference pipeline by offering fine-grained hallucination information for further mitigation.

6 Conclusion and Future Work

Hallucinations in generative tasks present substantial obstacles to the reliability and creditability of LLMs but lack a comprehensive and fine-grained detecting strategy. Thus, we present a bilingual dataset, ANAH for fine-grained hallucination annotation in GQA covering diverse topics, offering the opportunity to quantitatively analyze hallucination phenomena such as accumulation effect, and facilitating the development of state-of-the-art fine-grained hallucination annotators. Our generative hallucination annotators surpass all open-source LLMs and GPT-3.5 and obtain performance on par with GPT-4. Our generalization experiments indicate that improving the breadth of topics in the dataset is more important than extending questions under existing topics in the dataset.

This paper paves the way for further scaling up the dataset of ANAH to conduct a systematic evaluation and analysis of LLM hallucinations, with the trained hallucination annotators. The hallucination annotators also have the potential to be used in the hallucination mitigation pipeline in both the training and inference stages.

7 Limitations

This benchmark primarily incorporates the widely recognized and representative knowledge-intensive task, GQA. However, it does not encompass other tasks such as summarization and dialogue. During the dataset construction, we use GPT-3.5 with a reference document to construct a high-quality answer and an early version of InternLM-7B without reference to generate low-quality answers, respectively. Different models are used in that stage, we will further complete and analyze the other settings including GPT-3.5 without reference and InternLM-7B with reference.

In addition, our focus predominantly lies on the answer generation stage, without considering other stages such as the model's ability to recognize adversarial questions (Kumar et al., 2023; Zhu et al., 2023), red teaming (Ganguli et al., 2022), acknowledge unknown knowledge (Yin et al., 2023; Rajpurkar et al., 2018; Amayuelas et al., 2023), and retrieve accurate external knowledge once they realize their parametrical knowledge is not enough.

8 Ethical Considerations

We used publicly available reference documents for our benchmarks, effectively circumventing any possible harm toward individuals or groups. The generated data by LLMs were carefully selected and processed by humans to secure privacy and confidentiality. No personal identification information was involved, and all data were made anonymous before any analysis was conducted.

References

Vaibhav Adlakha, Parishad BehnamGhader, Xing Han Lu, Nicholas Meade, and Siva Reddy. 2023. Evaluating correctness and faithfulness of instruction-following models for question answering. *CoRR*, abs/2307.16877.

Alfonso Amayuelas, Liangming Pan, Wenhu Chen, and William Yang Wang. 2023. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. *CoRR*, abs/2305.13712.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Zehui Chen, Weihua Du, Wenwei Zhang, Kuikun Liu, Jiangning Liu, Miao Zheng, Jingming Zhuo, Songyang Zhang, Dahua Lin, Kai Chen, et al. 2023. T-eval: Evaluating the tool utilization capability step by step. *arXiv preprint arXiv:2312.14033*.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Wenwei Zhang, Jiangning Liu, Dahua Lin, Kai Chen, and Feng Zhao. 2024. Agent-flan: Designing data and methods of effective agent tuning for large language models. *arXiv preprint arXiv:2403.12881*.

Qinyuan Cheng, Tianxiang Sun, Wenwei Zhang, Siyin Wang, Xiangyang Liu, Mozhi Zhang, Junliang He, Mianqiu Huang, Zhangyue Yin, Kai Chen, et al. 2023. Evaluating hallucinations in chinese large language models. *arXiv preprint arXiv:2310.03368*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Nico Daheim, Nouha Dziri, Mrinmaya Sachan, Iryna Gurevych, and Edoardo M Ponti. 2023. Elastic weight removal for faithful and abstractive dialogue generation. *arXiv preprint arXiv:2303.17574*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Li Du, Yequan Wang, Xingrun Xing, Yiqun Ya, Xiang Li, Xin Jiang, and Xuezhi Fang. 2023a. Quantifying and attributing the hallucination of large language models via association analysis. *CoRR*, abs/2309.05217.

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

Esin Durmus, He He, and Mona Diab. 2020. Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070.

Nouha Dziri, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy. 2022a. Faithdial: A faithful benchmark for information-seeking dialogue. *Transactions of the Association for Computational Linguistics*, 10:1473–1490.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R Zaiane, and Siva Reddy. 2022b. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021. Evaluating groundedness in dialogue systems: The begin benchmark. *Findings of ACL*.

Mohamed Elaraby, Mengyin Lu, Jacob Dunn, Xueying Zhang, Yu Wang, and Shizhu Liu. 2023. Halo: Estimation and reduction of hallucinations in open-source weak large language models. *arXiv preprint arXiv:2308.11764*.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *CoRR*, abs/2209.07858.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Dialfact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801.

Conghui He, Wei Li, Zhenjiang Jin, Bin Wang, Chao Xu, and Lin. 2022. Opendatalab: Empowering general artificialintelligence with open datasets. `https://opendatalab.com`.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *ACM Computing Surveys*.

Ziwei Ji, Zihan Liu, Nayeon Lee, Tiezheng Yu, Bryan Wilie, Min Zeng, and Pascale Fung. 2023a. RHO: Reducing hallucination in open-domain dialogues with knowledge grounding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4504–4522, Toronto, Canada. Association for Computational Linguistics.

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

Ehsan Kamalloo, Nouha Dziri, Charles L. A. Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5591–5606. Association for Computational Linguistics.

Aounon Kumar, Chirag Agarwal, Suraj Srinivas, Soheil Feizi, and Hima Lakkaraju. 2023. Certifying

LLM safety against adversarial prompting. *CoRR*, abs/2309.02705.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Deren Lei, Yaxi Li, Mingyu Wang, Vincent Yun, Emily Ching, Eslam Kamal, et al. 2023. Chain of natural language inference for reducing large language model ungrounded hallucinations. *arXiv preprint arXiv:2310.03951*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. *arXiv preprint arXiv:2401.03205*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. 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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023b. Inference-time intervention: Eliciting truthful answers from a language model. *arXiv preprint arXiv:2306.03341*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2021. A token-level reference-free hallucination detection benchmark for free-form text generation. *arXiv preprint arXiv:2104.08704*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9802–9822. Association for Computational Linguistics.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2023. Generating benchmarks for factuality evaluation of language models. *arXiv preprint arXiv:2307.06908*.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin T. Vechev. 2023. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *CoRR*, abs/2305.15852.

None. 2023. Sharegpt.

OpenAI. 2023. Chatgpt: Optimizing language models for dialogue.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544.

Pouya Pezeshkpour. 2023. Measuring and modifying factual knowledge in large language models. *CoRR*, abs/2306.06264.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics.

Clément Rebuffel, Marco Roberti, Laure Soulier, Geoffrey Scoutheeten, Rossella Cancelliere, and Patrick Gallinari. 2022. Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 36(1):318–354.

Ruiyang Ren, Yuhao Wang, Yingqi Qu, Wayne Xin Zhao, Jing Liu, Hao Tian, Hua Wu, Ji-Rong Wen, and Haifeng Wang. 2023. Investigating the factual knowledge boundary of large language models with retrieval augmentation. *CoRR*, abs/2307.11019.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. Trusting your evidence: Hallucinate less with context-aware decoding. *arXiv preprint arXiv:2305.14739*.

Kai Sun, Yifan Ethan Xu, Hanwen Zha, Yue Liu, and Xin Luna Dong. 2023. Head-to-tail: How knowledgeable are large language models (llm)? aka will llms replace knowledge graphs? *arXiv preprint arXiv:2308.10168*.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities. https://github.com/InternLM/InternLM.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *CoRR*, abs/2307.15343.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2023. Unleashing cognitive synergy in large language models: A task-solving agent through multi-persona selfcollaboration. *arXiv preprint arXiv:2307.05300*, 1(2):3.

Rongxiang Weng, Heng Yu, Xiangpeng Wei, and Weihua Luo. 2020. Towards enhancing faithfulness for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2675–2684.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *arXiv preprint arXiv:2306.01693*.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. *arXiv preprint arXiv:2310.06498*.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

A Dataset Construction

A.1 Topic Selection and Reference Retrieval

We use InternLM to assess whether a query and its candidate entries are synonymous via the prompt in Figure A1.

If the sentence similarity between two questions exceeds the threshold, we consider them overly similar. The threshold is 300 for Chinese (via CoSENT) and 0.9 for English (via MiniLM), which are selected by case study.

A.2 Question Generation and Selection

First, we generate multiple questions based on the reference documents via prompts in Figure A2.

We use GPT-3.5 to filter the open-ended subjective questions and make sure of their answerability via the prompts in Figure A3.

We use GPT-4 to select the final questions based on authenticity, answerability, difficulty, and variety via prompts in Figure A4.

A.3 Answering under Different Models and Scenarios

We generate answers with the document via prompts in Figure A5.

A.4 Fine-grained Hallucination Annotation

We utilize GPT-4 to generate fine-grained hallucination annotation via prompts in Figure A6 to A10.

English Prompt:
I will provide two entries along with introductions. Please determine if the two entries are synonymous, i.e., if the two entries refer to the same event, object, person, or location, etc.
Entry 1: {name1}
Introduction 1: {doc1}
Entry 2: {name2}
Introduction 2: {doc2}
Are the two entries synonymous?

Figure A1: Prompts for Reference Retrieval.

B Case Study

Table A1, A2, and A3 show the examples where the GPT-4 generated annotation is inconsistent with human annotation.

C Implementation Details

C.1 Generative Annotator

The maximum sequence length is set to 16k. This setting is also held constant in baselines. We load the pre-trained InternLM2-7B model and train it with the following settings and hyper-parameters: the epoch is 1, the batch size is 2, the learning rate is 4e-5, and the AdamW optimizer is with a linear scheduler. We generate responses using sampling implemented via the LMDeploy library[17]. Our model is trained on 8 NVIDIA A800 GPUs. It takes approximately 1 hour to train.

C.2 Discriminative Annotator

We use InternLM2-7B and 20B as the base model for training. We train the discriminative annotator on our benchmark with the following settings and hyper-parameters: the epoch is 2, the batch size is 8, the learning rate is 1e-5, the AdamW optimizer is with a linear scheduler, and the maximum sequence length is 16k. Our model is trained on 8 NVIDIA A800 GPUs.

D Results and Analysis

Topic-specific automatic evaluation results for generative hallucination annotators are shown in Tab. A4. The trained ANAH-7B performs best on location topics while struggling with event topics.

Figure A11 shows the confusion matrices of hallucination type for annotators in different sizes. Figure A12 and A13 show the confusion matrices for discriminative annotators under different scenarios in different sizes.

D.1 Generalization on other LLMs

To assess generalizability, we sample 100 sentence-level annotations for answers generated by other models (Qwen-7B, Baichuan2-7B). We manually check the quality of ANAH-7B annotator as in Table A7. The accuracy for other models is similar to that of GPT3.5 and InternLM. It proves that our annotator is still relatively stable on other models.

We find that for the same query, generated answers from different models are around the topic and they are not far apart. Thus, in the context of factual QA, the divergence is not substantial and the answers are relatively in domain. Please find some examples in Tab. A8.

E Human Annotation

The annotation platform is developed internally by the laboratory. Human annotators, comprising well-educated undergraduates. Their salary is 300 yuan per day which is adequate given the participants' demographic. An ethics review board approved the data collection protocol.

Human annotation involves two stages: (1) screening topics and references; and (2) fine-grained hallucination annotation. We provide comprehensive instructions for each task, including task descriptions, precautions, estimated time, three examples, and three negative cases, to facilitate understanding.

We also employ a double annotation process during human annotation: (1) Annotators fix the GPT4 pre-annotations. (2) Experienced annotators (selected by platform) review the annotations and give feedback. Multiple rounds of (1) and (2) are performed until the platform deems the annotation

[17] https://github.com/InternLM/lmdeploy

English Prompt:
I would like you to act as a question generator. I will provide references and you will generate 10 questions about "{topic}" based on the reference. The specific requirements are as follows:
1. the questions can be fully answered based only on the reference document, i.e. the answers to the questions are fully contained in the reference document. The questions should be objective and not too subjective or open-ended.
2. the 10 questions should be of as many different types as possible, e.g. what, when, where, why. Questions can be asked from different perspectives, e.g. descriptions, explanations, reasons, etc. Ensure that the questions are of different types and cover all aspects of the information.
3. 10 questions can cover different levels of knowledge, from general, basic knowledge to more specialized, complex subject knowledge or domain knowledge.
4. have only one question per item.
Reference: {reference document}
Please list the 10 questions directly based on the above reference without any explanation:

Chinese Prompt:
我希望你充当一个问题生成器。我将提供参考资料，你将根据资料生成关于“{topic}”的10个问题。具体要求如下：
1. 只根据参考资料，完全可以回答问题，即问题的答案完全包含在参考资料中。问题要客观，不要太过主观和开放。
2. 10个问题尽量是不同类型的，比如：什么、何时、何地、为什么。问题可以从不同的角度出发，例如描述、解释、原因等。确保问题类型多样，覆盖资料的各个方面。
3. 10个问题可以涉及不同层次的知识，从常识性、基本性的知识，到更专业化、复杂化的学科知识或领域知识。
4. 每条只有一个问题。
参考资料：{reference document}
请根据以上参考资料，不做说明直接列出10个问题：

Figure A2: Prompts for Question Generation.

acceptable. (3) NLP experts check the annotation quality to finally decide whether to accept it. The pass rate is 85% and unqualified samples are redone until accepted.

English Prompt:
I would like you to act as a question judge. Given several questions, determine if each question meets all of the following conditions: objective, about facts, has a definitive answer, and not open-ended.
{questions}
Please answer "yes" or "no" in label order, separated by line breaks and without any explanation.
Chinese Prompt:
我希望你充当一个问题判断器。分别判断下列问题是否满足以下所有条件：客观的、关于事实的、有确切答案的、非开放的。
{questions}
请按标号顺序回答“是”或“否”，用换行符隔开，不加任何解释说明。

English Prompt:
I would like you to act as a question answerability judge. I will provide a question and reference document, and you will judge whether the question is fully answerable based only on the reference document, i.e., whether the answer is included in the reference.
Reference document: {reference document}
Question: {question}
Is it possible to answer the question at all, based only on the reference document? Please answer "yes" or "no" directly without any explanation.
Chinese Prompt:
我希望你充当一个问题可回答性判断器。我将提供问题和参考资料，你将判断只根据参考文档，是否完全可以回答问题，即答案是否包含在参考资料中。
参考文档：{reference document}
问题：{question}
只根据参考文档，是否完全可以回答问题？请直接回答“是”或“否”，不加任何解释说明。

Figure A3: Prompts for Question Answerability Judge.

English Prompt:
Good questions have the following characteristics: 1. high degree of truthfulness: the question contains no intentionally misleading, ambiguous or false information. 2. high answerability: remove questions that are too subjective, controversial, or predictive. 3. have a certain level of difficulty for the model. 4. increase the overall diversity (in terms of type, complexity, depth of knowledge, etc.), and remove questions that are similar to other questions. Combine the above evaluation metrics and select the 3 best problems among these. Please respond directly to the question numbers, separated by commas, without any explanation.

Chinese Prompt:
好的问题具有以下特征： 1. 真实度高：问题中有没有故意误导、含糊不清或者虚假的信息。 2. 可回答性高：去掉过于主观、有争议、预测类的问题。 3. 对于模型有一定的难度。 4. 增加整体的多样性（类型、复杂度、知识深度等方面）,去除和其他问题相似的问题。 综合以上评价指标，在这些问题中选择3个最好的问题。请直接回复问题编号，用逗号隔开，不加任何解释说明。

Figure A4: Prompts for Question Selection.

English Prompt:
Reference document: {reference document}
Please answer the question based on the above reference: {question}

Chinese Prompt:
参考资料：{reference document}
请根据以上参考资料，回答问题：{question}

Figure A5: Prompts for Answering.

English Prompt:
I would like you to act as a hallucination annotator in an answer. I will provide a reference document and a question about "{name}" and you will judge whether the answer point contains hallucinations. The specific requirements are as follows:
1. If the point is supported by and consistent with the reference document, please write <Hallucination> None. And write the specific reference segment: <Reference> XXX. If there are multiple reference segments, please use "<SEP>" to separate them. Reference segments should be copied directly from the original text without modification.
2. If the point contradicts the reference document, please write: <Hallucination> Contradictory. And write the specific reference segment: <Reference> XXX. Also, write how to modify the answer: <Correction> "XXX" to "YYYY". If you need to delete XXX, write: <Correction> "XXX" to "".
3. If the point cannot be verified and there is no evidence in reference to support it, please write: <Hallucination> Unverifiable. And write the specific reference segment: <Reference> XXX. Also, write how to modify the answer: <Correction> "XXX" to "YYYY". If you need to delete XXX, write: <Correction> "XXX" to "".
4. If the point does not contain any factual information to be judged, please write: <No Fact>.
Question: {question}
Reference: {reference document}
Point: {answer sentence}
Please annotate:

Chinese Prompt:
我希望你充当一个回答中的幻觉标注器。我将提供关于“{name}”的参考资料和问题，你将判断回答的要点是否含有幻觉。具体要求如下：
1. 如果要点与参考文档一致，请写：<幻觉>无。并注明参考片段：<参考>XXX。如果有多个参考片段，请用“<SEP>”分隔。参考片段应直接从原文复制，不需修改。
2. 如果要点与参考文档矛盾，请写：<幻觉>矛盾。并注明参考片段：<参考>XXX。同时说明如何修改回答：<改正>“XXX”改为“YYY”。如需删除内容XXX，请写：<改正>将“XXX”改为“”。
3. 如果要点无中生有，找不到证据支撑，无法验证，请写：<幻觉>无法验证。并注明参考片段：<参考>XXX。同时说明如何修改回答：<改正>“XXX”改为“YYY”。如需删除内容XXX，请写：<改正>将“XXX”改为“”。
4. 如果要点不包含待判断的事实信息，请写：<无事实>。
问题：{question}
参考文档：{reference document}
回答要点：{answer sentence}
请标注:

Figure A6: Prompts for Fine-grained Hallucination Annotation.

English Prompt:
I would like you to act as a hallucination annotator in an answer. I will provide a reference document and a question about "{name}" and you will judge whether the answer point contains hallucinations. The specific requirements are as follows:
1. If the point is supported by and consistent with the reference document, please write <Hallucination> None. And write the specific reference segment: <Reference> XXX. If there are multiple reference segments, please use "<SEP>" to separate them. Reference segments should be copied directly from the original text without modification.
2. If the point contradicts the reference document, please write: <Hallucination> Contradictory. And write the specific reference segment: <Reference> XXX. Also, write how to modify the answer: <Correction> "XXX" to "YYYY". If you need to delete XXX, write: <Correction> "XXX" to "".
3. If the point cannot be verified and there is no evidence in reference to support it, please write: <Hallucination> Unverifiable. And write the specific reference segment: <Reference> XXX. Also, write how to modify the answer: <Correction> "XXX" to "YYYY". If you need to delete XXX, write: <Correction> "XXX" to "".
4. If the point does not contain any factual information to be judged, please write: <No Fact>.
Reference: {reference document}
Question: {question}
Answer: {answer sentence}
Please annotate:

Chinese Prompt:
我希望你充当一个回答中的幻觉标注器。我将提供关于“{name}”的参考资料和问题，你将判断回答的要点是否含有幻觉。具体要求如下：
1. 如果要点与参考文档一致，请写：<幻觉>无。并注明参考片段：<参考>XXX。如果有多个参考片段，请用“<SEP>”分隔。参考片段应直接从原文复制，不需修改。
2. 如果要点与参考文档矛盾，请写：<幻觉>矛盾。并注明参考片段：<参考>XXX。同时说明如何修改回答：<改正>“XXX”改为“YYY”。如需删除内容XXX，请写：<改正>将“XXX”改为“”。
3. 如果要点无中生有，找不到证据支撑，无法验证，请写：<幻觉>无法验证。并注明参考片段：<参考>XXX。同时说明如何修改回答：<改正>“XXX”改为“YYY”。如需删除内容XXX，请写：<改正>将“XXX”改为“”。
4. 如果要点不包含待判断的事实信息，请写：<无事实>。
参考文档：{reference document}
问题：{question}
回答要点：{answer sentence}
请标注:

Figure A7: Prompts for Fine-grained Hallucination Annotation.

English Prompt:
I would like you to act as a hallucination annotator in an answer. I will provide a reference document and a question about "name" and you will judge whether each point of the answer contains hallucinations. The specific requirements are as follows:
1. If the point does not contain any factual information to be judged, please write: <No Fact>. And end the annotation.
2. If the point contains factual information, please find the specific reference segment and write: <Reference> XXX. If there are multiple reference segments, please use "<SEP>" to separate them. Reference segments should be copied directly from the original text without modification.
3. If the point is supported by and consistent with the reference document, please write: <Hallucination> None.
4. If the point contradicts the reference document, please write: <Hallucination> Contradictory. Also, write how to modify the answer: <Correction> "XXX" to "YYYY". If you need to delete XXX, write: <Correction> "XXX" to "".
5. If the point cannot be verified and there is no evidence in reference to support it, please write: <Hallucination> Unverifiable. Also, write how to modify the answer: <Correction> "XXX" to "YYYY". If you need to delete XXX, write: <Correction> "XXX" to "".
Question: {question}
Reference: {reference document}
Please annotate the answer:{answer sentence}

Chinese Prompt:
我希望你充当一个回答中的幻觉标注器。我将提供关于“name”的参考资料和问题，你将判断回答的每个要点是否含有幻觉。具体要求如下：
1. 如果要点不包含待判断的事实信息，请写：<无事实>，并结束标注。
2. 如果要点包含事实信息，请找相关的参考片段，请写：<参考>XXX。如果有多个参考片段，请用“<SEP>”分隔。参考片段应直接从原文复制，不需修改。
3. 如果要点与参考文档一致，请写：<幻觉>无。
4. 如果要点与参考文档矛盾，请写：<幻觉>矛盾。同时说明如何修改回答：<改正>“XXX”改为“YYY”。如需删除内容XXX，请写：<改正>将“XXX”改为“”。
5. 如果要点无中生有，找不到证据支撑，无法验证，请写：<幻觉>无法验证。同时说明如何修改回答：<改正>“XXX”改为“YYY”。如需删除内容XXX，请写：<改正>将“XXX”改为“”。
问题：{question}
参考文档：{reference document}
请标注要点：{answer sentence}

Figure A8: Prompts for Fine-grained Hallucination Annotation.

English Prompt:
Imagine you are a detective who specializes in identifying hallucinations. I will provide you with reference documents and questions about "name" and you will need to evaluate each point of information in the responses for the presence of hallucinations. Please follow the steps below:
- If the information point does not contain a fact that can be judged, mark: <No Fact> and end the annotation.
- If the information point contains a fact, list the corresponding reference: <Reference> XXX. If there is more than one, separate them with "<SEP>". Please ensure that the reference information is copied directly from the original text and does not need to be altered.
- If the information point is consistent with the reference, please mark: <Hallucination> None.
- If the information point contradicts the reference, please mark it as <Hallucination> Contradictory and include a correction: <Correction> "XXX" to "YYYY". When something needs to be eliminated, write: <Correction> "XXX" to "".
- If the information point cannot find relevant evidence, or cannot be verified, please mark: <Hallucination> Unverifiable, and include a correction: <Correction>"XXX" to "YYYY". When you need to eliminate something, please write: <Correction> "XXX" to "".

Question: {question}
Reference: {reference document}
Please annotate the information point: {answer sentence}

Chinese Prompt:
想象你是一个专门鉴别幻觉的侦查员。我将向你提供关于“name”的参考文档和问题，你需要评估回答中的每个信息点是否存在幻觉。请按以下步骤进行：
- 如信息点不包含可判断的事实，请标明：<无事实>，并结束评估。
- 如信息点包含事实，请列出相应的参考信息点：<参考>XXX。若有多个，请以“<SEP>”分隔。请确保参考信息直接复制自原文，无需更改。
- 如信息点与参考内容一致，请标注：<幻觉>无。
- 如信息点与参考内容相矛盾，请标注：<幻觉>矛盾，并附上改正方法：<改正>“XXX”改为“YYY”。需要剔除某内容时，请写：<改正>将“XXX”改为“”。
- 如信息点无法找到相关证据，或无法验证，请标注：<幻觉>无法验证，并附上改正方法：<改正>“XXX”改为“YYY”。需要剔除某内容时，请写：<改正>将“XXX”改为“”。

问题：{question}
参考文档：{reference document}
请标注信息点：{answer sentence}

Figure A9: Prompts for Fine-grained Hallucination Annotation.

English Prompt:
You are now a hallucination detection system. I will provide you with a reference document and a question on the topic "name". Your task is to analyze the responses to the question and determine whether or not there is a hallucination for each point. The steps of the assessment are as follows:
- If it does not contain factual information that needs to be judged, write: <No Fact> and stop the assessment.
- If facts are included, identify the relevant reference clip. Write: <Reference> XXX. Separate multiple references with "<SEP>". Please copy the reference fragment directly from the original without modification.
- If the points are identical to the reference, write: <Hallucination> None.
- If the main points are contradictory to the reference document, write: <Hallucination> Contradictory. Include a suggestion for revision: <Correction> "XXX" to "YYY". If a section needs to be deleted, write: <Correction> "XXX" to "".
- If no evidence can be found to support a point, or if it cannot be verified, write: <Hallucination> Unverifiable, with a suggested change: <Correction> "XXX" to "YYYY". If a section needs to be deleted, write: <Correction> "XXX" to "".
Question: {question}
Reference: {reference document}
Please analyze the point:{answer sentence}

Chinese Prompt:
你现在是一个幻觉检测系统。我会为你提供关于主题“name”的一篇参考文档和一个问题。你的任务是分析问题的回答，判断每个要点是否存在幻觉。评估步骤如下：
- 如果没有包含需要判断的事实信息，请写：<无事实>，并停止评估。
- 如果包含事实，找出相关参考片段。请写：<参考>XXX。多个参考片段请用"<SEP>"分隔。参考片段请直接从原文复制，不要修改。
- 如果要点与参考完全一致，请写：<幻觉>无。
- 如果要点与参考文档存在矛盾，写：<幻觉>矛盾。并附上修改建议：<改正>“XXX”改为“YYY”。如果需要删除某部分，写：<改正>将“XXX”改为“”。
- 如果无法找到证据支持要点，或无法验证，写：<幻觉>无法验证，并附上修改建议：<改正>“XXX”改为“YYY”。如果需要删除某部分，写：<改正>将“XXX”改为“”。
问题：{question}
参考文档：{reference document}
请分析要点：{answer sentence}

Figure A10: Prompts for Fine-grained Hallucination Annotation.

Topic	Fear of crime
Reference	**Reference 0:** The fear of crime refers to the fear of being a victim of crime as opposed to the actual probability of being a victim of crime. The fear of crime, along with fear of the streets and the fear of youth, is said to have been in Western culture for "time immemorial". While fear of crime can be differentiated into public feelings, thoughts and behaviors about the personal risk of criminal victimization, distinctions can also be made between the tendency to see situations as fearful, the actual experience while in those situations, and broader expressions about the cultural and social significance of crime and symbols of crime in people's neighborhoods and in their daily, symbolic lives... # Interpersonal communication and the mass media Hearing about events and knowing others who have been victimised are thought to raise perceptions of the risk of victimisation. This has been described as a 'crime multiplier', or processes operating in the residential environment that would 'spread' the impacts of criminal events... A subject's criminal risk perception is exaggerated by peer-communication on crime and only moderated by the own experience. Public perceptions of the risk of crime are no doubt also strongly shaped by mass media coverage.
Question	What are the three different components of risk perception that may interact in relation to the likelihood of falling victim to crime?
Selected Ans Sent	The three different components of risk perception that may interact in relation to the likelihood of falling victim to crime are: 1. Perceived Risk: This refers to an individual's subjective assessment of the likelihood of being a victim of crime. It is influenced by factors such as personal experience, media coverage, and social norms. 2. Actual Risk: This refers to the objective probability of being a victim of crime, based on statistical data and crime patterns. 3. Perceived Benefits and Costs: This refers to an individual's assessment of the benefits and costs of taking precautions to reduce their risk of being a victim of crime. It is influenced by factors such as the perceived effectiveness of different precautions, the perceived costs of taking precautions, and the perceived benefits of not taking precautions.
GPT-4 Annotation	**<No Fact>**
Human Annotation	**<Reference>** Reference 0: A subject's criminal risk perception is exaggerated by peer-communication on crime and only moderated by the own experience. Public perceptions of the risk of crime are no doubt also strongly shaped by mass media coverage. **<Halluciantion>** None

Table A1: An example for inconsistent GPT-4 and human annotation.

Topic	Neo-Freudianism
Reference	**Reference 0:** Neo-Freudianism is a psychoanalytic approach derived from the influence of Sigmund Freud but extending his theories towards typically social or cultural aspects of psychoanalysis over the biological...Karen Horney theorized that to deal with basic anxiety, the individual has three options: Moving toward others: Accept the situation and become dependent on others. This strategy may entail an exaggerated desire for approval or affection. Moving against others: Resist the situation and become aggressive. This strategy may involve an exaggerated need for power, exploitation of others, recognition, or achievement. Moving away from others: Withdraw from others and become isolated. This strategy may involve an exaggerated need for self-sufficiency, privacy, or independence...
Question	What are the three options that Karen Horney theorized individuals have to deal with basic anxiety?
Selected Ans Sent	Karen Horney theorized that individuals have three options to deal with basic anxiety: 1. Moving towards people: This option involves seeking out social connections and relationships in order to feel safe and secure. Individuals who choose this option may become overly dependent on others and struggle with feelings of inadequacy. 2. Moving against people: This option involves competing with others and striving for power and control in order to feel safe and secure. Individuals who choose this option may become overly aggressive and struggle with feelings of insecurity. 3. Moving away from people: This option involves withdrawing from social connections and relationships in order to feel safe and secure. Individuals who choose this option may become overly detached and struggle with feelings of loneliness and isolation.
GPT-4 Annotation	**<Halluciantion>** Unverifiable **<Correction>** "Individuals who choose this option may become overly aggressive and struggle with feelings of insecurity." to "Karen Horney theorized that to deal with basic anxiety, the individual has three options: Moving toward others (becoming dependent on others), moving against others (becoming aggressive), or moving away from others (becoming isolated)."
Human Annotation	**<Reference>** Reference 0: Moving against others: Resist the situation and become aggressive. This strategy may involve an exaggerated need for power, exploitation of others, recognition, or achievement. **<Halluciantion>** None

Table A2: An example for inconsistent GPT-4 and human annotation.

Topic	Psychosis
Reference	**Reference 0:** Psychosis, any of several major mental illnesses that can cause delusions, hallucinations, serious defects in judgment and other cognitive processes, and the inability to evaluate reality objectively... Other medical conditions associated with psychoses include thyroid disease, vitamin deficiencies, liver disease, epilepsy, Parkinson disease, and encephalitis. Disorders of sleep or prolonged sensory deprivation can also cause hallucinations and delusions. [1,6]Many drugs can alter brain function and cause symptoms of psychoses.
Question	How are hallucinations related to abnormal brain activity?
Selected Ans Sent	Hallucinations are a symptom of abnormal brain activity. Hallucinations can be caused by a variety of factors, including mental illness, substance abuse, and neurological disorders. When the brain is not functioning properly, it can produce hallucinations as a result of abnormal activity in the visual, auditory, or other sensory systems. For example, in schizophrenia, the brain's dopamine system is overactive, leading to hallucinations and delusions. Similarly, in Parkinson's disease, abnormal activity in the basal ganglia can cause visual hallucinations. Other conditions, such as epilepsy and migraines, can also cause hallucinations as a result of abnormal brain activity. Overall, hallucinations are a symptom of abnormal brain activity and can be caused by a variety of underlying conditions.
GPT-4 Annotation	**<Reference>** Reference 0: Other medical conditions associated with psychoses include thyroid disease, vitamin deficiencies, liver disease, epilepsy, Parkinson disease, and encephalitis. Disorders of sleep or prolonged sensory deprivation can also cause hallucinations and delusions. **<Halluciantion>** None
Human Annotation	**<Reference>** Reference 0: Other medical conditions associated with psychoses include thyroid disease, vitamin deficiencies, liver disease, epilepsy, Parkinson disease, and encephalitis. **<Halluciantion>** Unverifiable **<Correction>** "and migraines" to "".

Table A3: An example for inconsistent GPT-4 and human annotation.

Topic	F1 ↑	ACC ↑	R ↑	BERT ↑	Pre4 ↑
Person	75.8	76.58	53.69	87.69	52.79
Event	70.48	73.33	52.26	80.70	42.94
Location	83.34	83.81	75.39	92.77	67.86
Thing	79.36	80.23	58.19	87.34	29.53

Table A4: Topic-specific automatic evaluation results for generative hallucination annotators ANAH-7B, where 'R', 'BERT', and 'Pre4' refer to 'RougeL', 'BERTScore', and '4-gram Precision', respectively.

Setting	BERT↑	
	T	Q
G-7B	87.29	87.27
G-20B	87.96	88.93

Table A5: Evaluation results for generative annotators, noted by 'G'. 'T' represents the unseen-topic test set, while 'Q' represents the unseen-question test set.

Setting	BERT↑	
	T	Q
S.T.	87.29	87.27
M.T.	85.94	87.55
above + D.	86.05	86.71
M.T.+ P.A.	86.95	87.6
above + D.	86.89	87.58

Table A6: Ablation Study for Generative Annotator based on InternLM-7B in different settings. Here, 'S.T.' means single-task training, which only includes hallucination annotation task in training, while 'M.T.' adopts multi-task training, which further encompasses several generative tasks. "+ D" indicates that testing the annotations with prompt disturbance *i.e.*, the instructions used in testing are unseen in training. "P.A." indicates prompt augmentation is adopted in training.

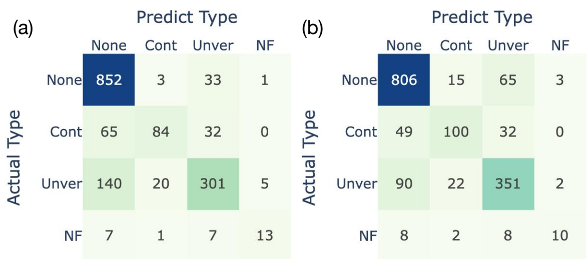

Figure A11: Hallucination Type Confusion Matrices for Generative Annotators. (a) InternLM2-7B-based annotator (b) InternLM2-20B-based annotator

Model	F1↑	ACC↑
Qwen-7B	73.80	76.76
Baichuan2-7B	75.97	77.55

Table A7: Evaluation results for generative annotators.

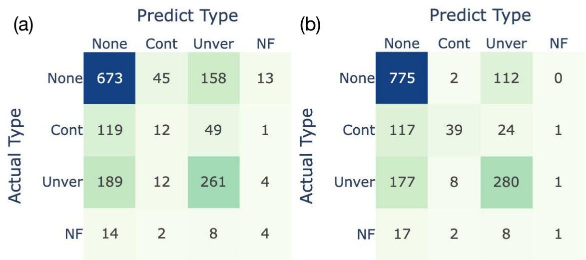

Figure A12: Hallucination Type Confusion Matrices for Discriminative Annotators based on InternLM2-7B. (a) without reference (b) with reference

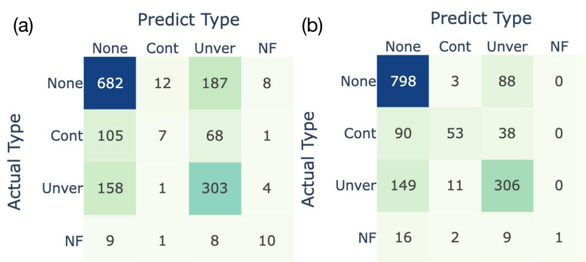

Figure A13: Hallucination Type Confusion Matrices for discriminative annotators based on InternLM2-20B. (a) without reference (b) with reference

Question	How did Victor Hugo gain wider fame in 1831?
InternLM2-7B	Victor Hugo gained wider fame in 1831 with his historical novel Notre-Dame de Paris (The Hunchback of Notre-Dame), an evocation of life in medieval Paris during the reign of Louis XI...
Qwen-7B	Victor Hugo gained wider fame in 1831 with his historical novel Notre-Dame de Paris, which is an evocation of life in medieval Paris during the reign of Louis XI...
Baichuan2-7B	Victor Hugo gained wider fame in 1831 when he published his novel Les Misérables. This novel was published in 1831 and was well received by critics and the public...

Table A8: Generated answers from different models for the same query.