Hallucination Diversity-Aware Active Learning for Text Summarization

Yu Xia[1,2] **Xu Liu**[1] **Tong Yu**[3*] **Sungchul Kim**[3] **Ryan A. Rossi**[3]
Anup Rao[3] **Tung Mai**[3] **Shuai Li**[1]
[1]Shanghai Jiao Tong University [2]University of Michigan [3]Adobe Research
xiayuu@umich.edu {liu_skywalker, shuaili8}@sjtu.edu.cn
{tyu, sukim, ryrossi, anuprao, tumai}@adobe.com

Abstract

Large Language Models (LLMs) have shown propensity to generate hallucinated outputs, i.e., texts that are factually incorrect or unsupported. Existing methods for alleviating hallucinations typically require costly human annotations to identify and correct hallucinations in LLM outputs. Moreover, most of these methods focus on a specific type of hallucination, e.g., entity or token errors, which limits their effectiveness in addressing various types of hallucinations exhibited in LLM outputs. To our best knowledge, in this paper we propose the first active learning framework to alleviate LLM hallucinations, reducing costly human annotations of hallucination needed. By measuring fine-grained hallucinations from errors in semantic frame, discourse and content verifiability in text summarization, we propose **HA**llucination **D**iversity-**A**ware **S**ampling (HADAS) to select diverse hallucinations for annotations in active learning for LLM finetuning. Extensive experiments on three datasets and different backbone models demonstrate advantages of our method in effectively and efficiently mitigating LLM hallucinations.

1 Introduction

Despite the prominent capabilities of large language models (LLMs) in natural language generation (NLG) tasks (Brown et al., 2020; Chung et al., 2022; Touvron et al., 2023), a notable limitation of them lies in their propensity to hallucinate (Ji et al., 2023; Manakul et al., 2023; Zhao et al., 2023; Bang et al., 2023; Peng et al., 2023), where models generate seemingly plausible but ungrounded outputs that either contradict or cannot be verified by existing sources. The phenomenon of hallucination poses a crucial challenge to the real-world applications of LLMs, where the models' faithfulness and trustworthiness are emphasized (Yang et al., 2023; Zhao et al., 2023).

*Corresponding author

While many methods have been proposed recently to detect hallucinations in the outputs of LLMs (Manakul et al., 2023; Li et al., 2023; Mündler et al., 2024), how to efficiently and effectively alleviate hallucinations in LLMs remains a notable problem. Existing methods for hallucination mitigation often focus on finetuning LLMs with human feedback or human-annotated samples to align the models' outputs with human-plausible content (Ouyang et al., 2022; Sun et al., 2023; Wu et al., 2024). While these methods have proven effective, they often require large amounts of costly human annotations to identify and rectify hallucinations in LLM outputs (Zhao et al., 2023; Guerreiro et al., 2023; Perlitz et al., 2023; Xia et al., 2024). Moreover, most of them emphasize mitigating a specific type of hallucination, e.g., entity or token errors (Nan et al., 2021; Cao et al., 2022), which limits their applicability in addressing various types of hallucinations comprehensively.

Aiming to reduce the intensive amount of human annotations needed, in this paper we propose an active learning framework to finetune LLMs for hallucination mitigation. In this framework, we actively select samples that LLMs may hallucinate on for annotation and finetuning. As the text summarization task has gained wide attention in factuality evaluations, which measure whether the model's outputs are faithful to the source document (Maynez et al., 2020; Pagnoni et al., 2021; Ji et al., 2023; Li et al., 2023; Zhang et al., 2024), we instantiate our active learning framework to address LLM hallucinations in generated summaries. We revisit the different types of hallucinations in text summarization defined by Pagnoni et al. (2021). Then, we leverage corresponding detection models (Zhang et al., 2020; Zhong et al., 2022; Feng et al., 2023) to measure fine-grained hallucinations, including semantic frame errors, discourse errors, and content verifiability errors, for annotation sample selection.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 8662–8674

While measuring potential hallucinations of all three types, greedily choosing the samples most likely to exhibit hallucinations may result in an excessive focus on addressing a certain type of hallucination while overlooking others. For example, if the evaluation score for semantic frame hallucinations dominates among all three types, greedy selection would then lead to choosing samples that mostly result in semantic frame errors for human annotations. As a result, the finetuned LLMs may reduce semantic frame hallucinations effectively but still suffer from other types. To address this limitation and take into account the diversity of hallucination samples, we propose a sample selection strategy for our active learning framework, called **HA**llucination **D**iversity-**A**ware **S**ampling (HADAS). Extensive experiments demonstrate the advantage of our proposed method in alleviating hallucinations, while also limiting the amount of costly human annotations, compared with both the random sampling baseline and the existing sample selection approaches for text summarization.

In summary, we make the following contributions in this work: i) To our best knowledge, in this paper we propose the first active learning framework to alleviate LLM hallucinations, reducing the amount of human annotations needed; ii) We propose a sample selection strategy HADAS to select samples of diverse hallucination types; iii) We demonstrate with extensive experiments the effectiveness of our proposed active learning method in mitigating hallucinations in text summarization.

2 Related Work

2.1 Hallucination Mitigation in LLM

The hallucination problem has been a pressing topic in recent studies on LLMs, where models generate incorrect or non-existent information that either contradicts or is unsupported by existing sources (Ji et al., 2023; Yang et al., 2023). Although there is a growing number of studies on hallucination detection and evaluation in LLMs (Manakul et al., 2023; Bang et al., 2023; Guerreiro et al., 2023; Li et al., 2023; Mündler et al., 2024), how to effectively and efficiently mitigate hallucinations remains a notable challenge. A few recent works have explored addressing the hallucination problem during inference time via improved decoding strategies (Lee et al., 2022; Shi et al., 2023; Wan et al., 2023), retrieval augmentation (Shuster et al., 2021; Peng et al., 2023), and self-verification (Varshney et al., 2023; Dhuliawala et al., 2023). Another line of works focus on finetuning LLMs to hallucinate less with various learning paradigms. Wan and Bansal (2022) incorporate factual consistency as one of the training objectives during finetuning. Sun et al. (2023) use contrastive learning to reduce hallucination by comparing faithful samples with hallucinated samples. Roit et al. (2023) leverage reinforcement learning to align LLMs' outputs to be more factually consistent to the source document with a natural language inference model. While these methods have been proven effective, they typically require a large amount of costly human annotations. In comparison, our proposed active learning framework for LLM finetuning aims to mitigate hallucinations while minimizing the amount of human annotations needed.

2.2 Active Learning in NLG

Active learning is a well-known technique employed in natural language processing to reduce annotation efforts by actively selecting informative samples (Zhang et al., 2022). In the context of language modeling, active learning is mainly used for text classification tasks (Ein-Dor et al., 2020; Margatina et al., 2021; Wu et al., 2022; Yu et al., 2022), such as named entity recognition (Shen et al., 2018; Radmard et al., 2021). A few recent works have explored active learning methods for NLG tasks. Tsvigun et al. (2022) propose the first effective diversity-based active learning query strategy for text summarization based on the embedding similarities between source documents. The authors report that the uncertainty-based strategy does not perform well and is outperformed by the random sampling baseline. Perlitz et al. (2023) evaluate the performance of existing active learning strategies across various NLG tasks such as paraphrase generation, summarization, and question generation. The authors suggest that compared to classification tasks, the lack of clearly defined ground-truth labels in NLG tasks poses difficulties in measuring uncertainty, which contributes to poor performance in uncertainty-based sample selection. As LLMs' hallucinations typically occur in NLG tasks, applying active learning for hallucination mitigation is an unexplored and non-trivial task. Our work proposes a diversity-based sampling strategy addressing LLMs' hallucinations in text summarization. Note that while Tsvigun et al. (2022) also proposes a diversity-based method for text summarization, it aims to select document samples that are

Source Document	
Heavy rains and flooding have forced hundreds of thousands of people from homes in southern Mexico's state of Tabasco over the past four days, with nearly as many trapped by the rising waters, state officials said Thursday. Officials say about 300,000 people are still trapped by the worst flooding in the region for 50 years ...	
Hallucination Type	**Example Summary**
Semantic Frame Error: The entity or predicate in the summary is not inconsistent with source document.	Recent heavy rains in northern Mexico have caused the worst flooding in 50 years.
Discourse Error: The statements or references in the summary are linked in an erroneous way.	Due to the worst flooding in 50 years in Tabasco, officials report that heavy rains began last Thursday.
Content Verifiability Error: The information in the summary is not present or verifiable in source document.	Due to heavy rains in southern Mexico, a state emergency was declared in Tabasco.

Table 1: Examples of three types of hallucinations in text summarization following the typology proposed by Pagnoni et al. (2021). The source document is from CNN-DailyMail (Hermann et al., 2015). We highlight the hallucinated content in red.

semantically diverse. In contrast, the diversity considered in our method focuses on various types of hallucinations in generated summaries. Thus, we make the first attempt towards an active learning paradigm for hallucination mitigation in NLG.

3 Hallucination Typology Revisit

Since LLMs may hallucinate in different forms (Ji et al., 2023), evaluations of hallucination have received increasing attention in recent studies. Particularly, a variety of factuality metrics have been developed (Maynez et al., 2020; Wang et al., 2020; Pagnoni et al., 2021; Zhong et al., 2022; Fabbri et al., 2022; Feng et al., 2023), including entity or token hallucination (Liu et al., 2022), sentence hallucination (Manakul et al., 2023), and relation hallucination (Zha et al., 2023).

Aiming to comprehensively mitigate hallucinations in text summarization, we follow the typology proposed in Pagnoni et al. (2021) and introduce three types of common hallucinations in text summarization: i) *Semantic Frame Error*, ii) *Discourse Error*, and iii) *Content Verifiability Error*. To provide more background information, we provide illustrative examples of these three types of hallucination in Table 1. Specifically, given a news article reporting the heavy rains and flooding in the southern Mexico area as the source document, a semantic frame error in the summary can be an entity or predicate incorrectly interpreted from the source, e.g. southern being misinterpreted as northern as illustrated in the first example summary. Additionally, a discourse error refers to the case when statements or claims in a sentence are linked together erroneously in terms of temporal ordering or causal links, e.g. the second example summary mistakenly states that the flooding was the cause of heavy rains. Moreover, a content verifiability error stands for the extrinsic information unverifiable from the source document, e.g., the declared state emergency in the third example summary is not mentioned in the source text. These examples demonstrate the varying forms of hallucinations that LLMs may generate in the summary.

Hence, by evaluating hallucinations in these various aspects, we can derive a more comprehensive understanding of hallucinations in LLM text summarization. This motivates us to capture and mitigate diverse types of hallucinations and enhance the faithfulness of LLMs in text summarization.

4 Methodology

We first present our proposed active learning framework for LLM finetuning in hallucination mitigation in Section 4.1. Section 4.2 details how we capture diverse hallucination types using off-the-shelf detection models. Then in Section 4.3, we describe in detail our hallucination diversity-aware sample selection strategy.

4.1 Active Learning for LLM Finetuning

We formulate our proposed active learning framework for LLM finetuning in text summarization with a feedback loop between the LLM and the annotator, as illustrated in Figure 1. We first introduce the necessary notations as follows.

Given an LLM, we denote its weights as $\mathbf{W}$. We denote an input document as $\mathbf{x} = (x_1 \dots x_m)$ and the summary generated by the LLM as $\mathbf{y} = (y_1 \dots y_n)$, where m and n are the token lengths of the document and generated summary respectively. Suppose we have a total of N documents in the

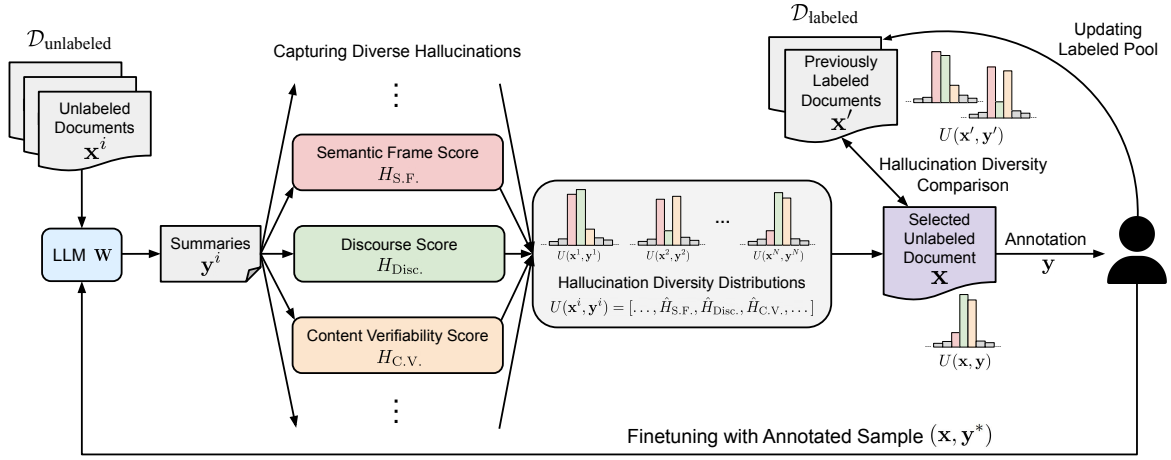


Figure 1: Overview of our hallucination diversity-aware active learning framework.

unlabeled pool, denoted as $\mathcal{D}_{\text{unlabeled}} = \{\mathbf{x}^i\}_{i=0}^N$. An unlabeled document means that no annotation is currently available to identify and correct the potential hallucination in LLM-generated summary $\mathbf{y}$ based on this document. We also keep track of a labeled pool $\mathcal{D}_{\text{labeled}} = \{(\mathbf{x}^j, \mathbf{y}^{*j})\}_{j=0}^M$, where $\mathbf{y}^*$ denotes the annotated summary and M is the size of labeled pool. We denote a sample selection strategy as $\mathcal{A}$ and the active learning loop consists of the following three main steps.

Sample Selection. To select a document $\mathbf{x}^i$ from the unlabeled pool $\mathcal{D}_{\text{unlabeled}}$, the LLM with weights $\mathbf{W}$ first generates a summary $\mathbf{y}^i$ for each document. Then, based on the selection strategy instantiated by the query function $\mathcal{A}$, we choose

$$(\mathbf{x}, \mathbf{y}) = \arg \max_{i \in 1,\ldots,N} \mathcal{A}((\mathbf{x}^i, \mathbf{y}^i) \mid \mathcal{D}_{\text{unlabeled}}, \mathbf{W}) \ , \tag{1}$$

which maximizes the designed criteria of $\mathcal{A}$ to choose the most informative samples for hallucination mitigation.

Human Annotation. The selected document-summary pair $(\mathbf{x}, \mathbf{y})$ is then annotated by examining and correcting $\mathbf{y}$ for hallucinated content based on the source document $\mathbf{x}$. The annotated summary denoted as $\mathbf{y}^*$ is collected. Subsequently, the document $\mathbf{x}$ is removed from the unlabeled pool and added to the labeled pool along with $\mathbf{y}^*$:

$$\mathcal{D}_{\text{unlabeled}} := \mathcal{D}_{\text{unlabeled}} \setminus \{\mathbf{x}\} \ ,$$

$$\mathcal{D}_{\text{labeled}} := \mathcal{D}_{\text{labeled}} \cup \{(\mathbf{x}, \mathbf{y}^*)\} \ .$$

Model Finetuning. After receiving the annotated document-summary pair $(\mathbf{x}, \mathbf{y}^*)$, the LLM is fine-tuned, and its weights are updated based on the document and the hallucination-annotated summary:

$$\hat{\mathbf{W}} = \arg \min_{\mathbf{W}} \mathcal{L}((\mathbf{x}, \mathbf{y}^*), \mathbf{W}) \ ,$$

where $\mathcal{L}$ is the loss function used for LLM finetuning, e.g., supervised finetuning objective. Then the updated LLM is evaluated on the validation or test set. Next, the LLM with the updated weights $\hat{\mathbf{W}}$ is used for the new round of sample selection with similar procedures as described previously.

Such iterative learning process is repeated until a stopping criterion is met, such as reaching a preset number of iterations or when the model's performance on the validation set no longer improves after a certain number of consecutive rounds.

4.2 Capturing Diverse Hallucination Types

As discussed in Section 3, hallucinations in LLMs can be of various types. To select samples that LLMs tend to hallucinate on, we aim to capture different types of hallucination in text summarization. Specifically, we adopt three hallucination detection methods measuring semantic frame errors, discourse errors, and content verifiability errors, respectively. The details are described as follows.

Note that we are fully aware that there are many emerging new types of hallucinations beyond the three types we considered here. As it is unrealistic to exhaustively take into account all hallucination evaluation methods, we follow a well-defined typology proposed by Pagnoni et al. (2021) to capture three common hallucinations in text summarization. Our contribution lies in developing a generic active learning framework for hallucination mitigation, which offers flexibility in that new measurements of hallucinations can be easily integrated.

Semantic Frame Score $H_{\text{S.F.}}$. As suggested and validated in Pagnoni et al. (2021), Ribeiro et al. (2022), and Feng et al. (2023), entailment-based models show clear advantages in detecting hallucinations on semantic frames, due to their fine-grained representation of facts, entities, and relations. Therefore, we adopt a recent entailment-based model FactKB (Feng et al., 2023) to evaluate semantic frame errors, which achieves state-of-the-art performances on factual consistency detection and high correlations with human judgments. The model takes the document-summary pair as input and outputs a probability of the summary being factually consistent with the document, which we denote as the semantic frame (S.F.) score $H_{\text{S.F.}}$.

Discourse Score $H_{\text{Disc.}}$. Different from semantic frames that focus on parts of a sentence like entities, detecting discourse errors such as erroneously connected claims requires a view of the entire sentence (Pagnoni et al., 2021). Therefore, sentence-level detection which is widely adopted in recent QA-based models (Wang et al., 2020; Fabbri et al., 2022; Zhong et al., 2022) comes in handy. The idea behind these methods is to compose each sentence of the models' outputs as a question and then ask a pre-trained QA model to answer whether this sentence is faithful to the source document. We adopt a recent QA-based method, UniEval (Zhong et al., 2022), which leverages a pretrained T5 model, further enhancing its natural language understanding ability at the sentence level. We denote the probability of the model answering "Yes" to the question as the discourse score $H_{\text{Disc.}}$.

Content Verifiability Score $H_{\text{C.V.}}$. For content verifiability, the main goal is to evaluate whether the information in the summary is present in the source document. Thus, as observed by Pagnoni et al. (2021), Ribeiro et al. (2022), and Feng et al. (2023), token-level similarity metrics such as BERTScore (Zhang et al., 2020) perform competitively well. Therefore, we choose BERTScore-Precision (BERT-P), which is more correlated with human judgments according to Feng et al. (2023), as our content verifiability score denoted as $H_{\text{C.V.}}$.

4.3 Hallucination Diversity-Aware Sampling

In this section, we describe in detail our proposed sample selection strategy that selects diverse hallucination samples for annotations.

Hallucination Score $H_{\text{Halu.}}$. With the above scores focusing on three different hallucination types, a natural idea for active learning sample selection strategy is to greedily select samples with the lowest total hallucination scores. Given a sample of document-summary pair $(\mathbf{x}, \mathbf{y})$, the hallucination score for this sample is calculated as

$$H_{\text{Halu.}}(\mathbf{x}, \mathbf{y}) = w_1 \hat{H}_{\text{S.F.}} + w_2 \hat{H}_{\text{Disc.}} + w_3 \hat{H}_{\text{C.V.}} \tag{2}$$

where $\hat{H}_{\text{S.F.}}$ is the min-max normalized value of $H_{\text{S.F.}}$ and similarly for $\hat{H}_{\text{Disc.}}$ and $\hat{H}_{\text{C.V.}}$, and w_1, w_2, and w_3 are the weights for each score. Note that for the three hallucination scores discussed in Section 4.2, the higher the score, the better. Thus, the lower the $H_{\text{Halu.}}$, the more hallucinations occur in the generated summary.

Such greedy exploitation of hallucination scores, however, might not lead to the most informative sample selections, as it might give excessive focus on a certain type of hallucination. For example, if semantic frame errors are more common and scores of $H_{\text{S.F.}}$ are consistently low, the hallucination score $H_{\text{Halu.}}$ would be predominantly influenced by $H_{\text{S.F.}}$. This could result in the selection of samples primarily exhibiting semantic frame errors, while neglecting other types of hallucinations.

Hallucination Diversity Score $H_{\text{Div.}}$. To further address the limitation of the greedy method, we propose a hallucination diversity-based sample selection strategy, **HA**llucination **D**iversity-**A**ware **S**ampling (HADAS). The main idea behind HADAS is that it would query the samples that have low hallucination scores while ensuring at the same time the hallucination types of selected samples as dissimilar (i.e., diverse) as possible.

To measure the similarity between hallucinations, we consider normalized scores of hallucination types as hallucination distribution U as

$$U(\mathbf{x}, \mathbf{y}) = [\hat{H}_{\text{S.F.}}, \hat{H}_{\text{Disc.}}, \hat{H}_{\text{C.V.}}] \,,$$

where additional hallucination metrics on other types can be easily included as illustrated in Figure 1. Then, we calculate the average Jensen-Shannon Divergence between the hallucination distribution of each unlabeled sample and all samples in the labeled pool as the diversity score $H_{\text{Div.}}$. Formally, given a unlabeled document and LLM-generated summary $(\mathbf{x}, \mathbf{y})$, its diversity score is calculated as

$$H_{\text{Div.}}(\mathbf{x}, \mathbf{y}) = \frac{\sum_{\mathcal{D}_{\text{labeled}}} \texttt{JSD}(U(\mathbf{x}, \mathbf{y}), U(\mathbf{x}', \mathbf{y}'))}{|\mathcal{D}_{\text{labeled}}|} \,, \tag{3}$$

Model	Method	CNN-DailyMail				Multi-News				Gigaword			
		BERT-P	UniEval	FactKB	ROUGE-L	BERT-P	UniEval	FactKB	ROUGE-L	BERT-P	UniEval	FactKB	ROUGE-L
Flan-T5 Small	Random	73.30	60.12	69.83	13.76	67.84	46.68	62.60	9.63	56.71	36.72	7.50	23.06
	IDDS	74.92	63.96	76.58	14.63	66.96	50.06	66.00	9.40	57.42	39.22	9.00	23.67
	$\text{HADAS}_{\text{w/o Div.}}$	76.64	70.63	82.26	15.36	68.95	50.66	68.49	10.08	57.40	33.77	9.23	22.29
	HADAS	**78.63**	**75.75**	**87.46**	**16.55**	**70.26**	**56.40**	**74.22**	**11.04**	**61.06**	**40.53**	**10.85**	**23.89**
Flan-T5 Base	Random	69.26	58.65	69.25	15.12	65.51	47.71	52.69	7.45	56.33	42.00	7.33	27.29
	IDDS	70.64	63.95	74.22	15.42	62.22	40.17	41.74	6.68	56.77	43.97	6.49	27.09
	$\text{HADAS}_{\text{w/o Div.}}$	72.05	67.42	77.13	16.51	69.83	56.82	61.57	9.33	54.93	39.10	5.43	28.39
	HADAS	**73.74**	**70.31**	**80.73**	**17.19**	**70.82**	**61.12**	**66.39**	**9.87**	**59.36**	**47.98**	**9.18**	**29.25**
BART Base	Random	76.08	74.02	89.65	19.57	69.25	50.52	76.72	12.78	79.78	61.23	51.08	35.32
	IDDS	74.25	68.01	88.86	19.39	71.00	53.49	80.06	14.80	83.63	62.71	55.43	35.56
	$\text{HADAS}_{\text{w/o Div.}}$	77.56	75.42	92.82	19.95	68.68	50.78	75.81	13.28	85.59	56.60	69.44	35.11
	HADAS	**78.14**	**76.65**	**93.95**	**20.12**	**71.03**	**55.94**	**80.22**	**14.83**	**87.59**	**63.75**	**70.12**	**35.91**

Table 2: Main results of summarization factuality and quality metrics with 30% of hallucination annotations across models and datasets, where the best results are highlighted in bold, and the second-best are underscored.

where $(\mathbf{x}', \mathbf{y}')$ are samples from the labeled pool $\mathcal{D}_{\text{labeled}}$, and JSD represents the Jensen-Shannon Divergence measure. A higher $H_{\text{Div.}}$ value indicates higher diversity of the unlabeled sample $(\mathbf{x}, \mathbf{y})$ compared to previously labeled samples.

With the hallucination score defined in Equation 2 and the diversity score defined in Equation 3, we propose the following query function $\mathcal{A}$ that implements selection criteria in Equation 1:

$$\mathcal{A}(\mathbf{x}, \mathbf{y}) = \lambda H_{\text{Div.}}(\mathbf{x}, \mathbf{y}) - (1 - \lambda) H_{\text{Halu.}}(\mathbf{x}, \mathbf{y}), \quad (4)$$

where $\lambda \in [0, 1]$ is a hyperparameter. With the sample selection strategy implemented, we have completed the active learning framework for hallucination mitigation as formulated in Section 4.1 and illustrated in Figure 1.

5 Experiment Setup

5.1 Backbone Models

We conduct our experiments mainly on three backbone LLMs, Flan-T5 Small (Chung et al., 2022), Flan-T5 Base (Chung et al., 2022), and BART Base (Lewis et al., 2020). The models are selected following (Tsvigun et al., 2022) and based on their distinctive strengths in text summarization. For Flan-T5 Small and Flan-T5 Base, we directly prompt the models with the instruction "`Summarize:`" as they have been instruction-tuned to do summarization task. For BART Base, we follow Lewis et al. (2020) to use a BART Base model finetuned on XSum dataset to ensure the summarization quality.

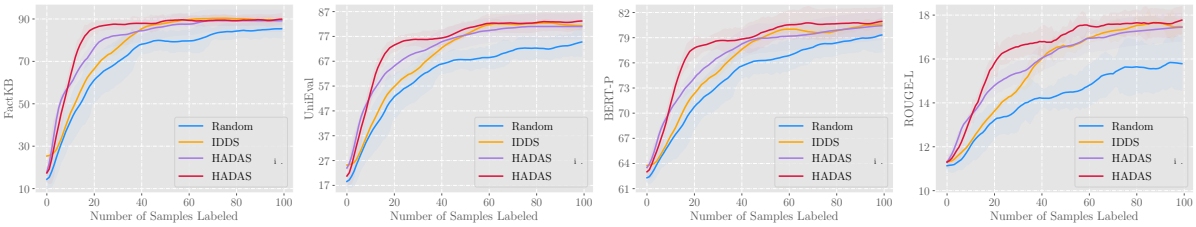
5.2 Datasets and Metrics

We choose three datasets, CNN-DailyMail (Hermann et al., 2015), Multi-News (Fabbri et al., 2019), and Gigaword (Rush et al., 2015). For computational efficiency, following Shen et al. (2018) and Tsvigun et al. (2022), we select a subset of samples from each dataset. Specifically, for CNN-DailyMail dataset, we randomly sample 5,000 samples from the training set, 500 from the test set, and 250 from the validation set. For both Multi-News and Gigaword datasets, we first randomly sample 2,000 samples from the training set. To better demonstrate improvements on these two datasets, we intentionally apply filtering to choose 200 samples from the test set and 100 samples from the validation set that are more prone to hallucinations by the models, as measured by the metrics introduced in Section 4.2, with $H_{\text{C.V.}}$ lower than 60 and both $H_{\text{Disc.}}$ and $H_{\text{S.F.}}$ lower than 40.

To evaluate the performance of our methods, we use the three hallucination detection metrics as introduced in Section 4.2: FactKB (Feng et al., 2023) for semantic frame, UniEval (Zhong et al., 2022) for discourse, and BERT-P (Zhang et al., 2020) for content verifiability. In addition, we also measure the ROUGE-L (Lin, 2004) score to assess the quality of generated summaries.

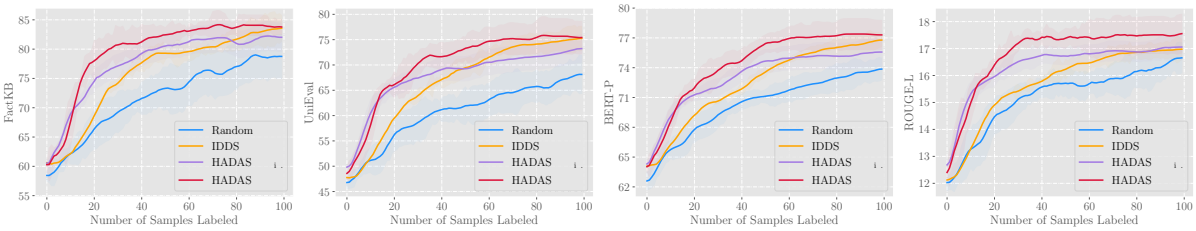
5.3 Baselines and Variants

We compare our proposed HADAS method with the following baselines and variants. **Random**: A canonical active learning baseline that randomly selects from samples without requiring any additional information. **IDDS** (Tsvigun et al., 2022): A recent diversity-based sampling strategy con-



(a) FactKB (b) UniEval (c) BERT-P (d) ROUGE-L

Figure 2: Factuality and quality curves over full hallucination annotations of Flan-T5 Small on CNN-DailyMail.



(a) FactKB (b) UniEval (c) BERT-P (d) ROUGE-L

Figure 3: Factuality and quality curves over full hallucination annotations of Flan-T5 Base on CNN-DailyMail.

sidering semantic similarities between documents for text summarization. **HADAS$_{\text{w/o Div.}}$**: A variant of our proposed method that do not consider hallucination diversity. **HADAS$_{\text{w/ S.F.}}$**: A variant of our proposed method based solely on semantic frame scores. **HADAS$_{\text{w/ Disc.}}$**: A variant of our proposed method based solely on discourse scores. **HADAS$_{\text{w/ C.V.}}$**: A variant of our proposed method based solely on content verifiability scores.

For the hyperparameters of HADAS and HADAS$_{\text{w/o Div.}}$, we set $w_1 = w_2 = w_3 = 0.33$ assuming their losses contribute equally to the hallucination generation (Groenendijk et al., 2021). For HADAS specifically, we did a grid search on the value of λ across $[0.25, 0.33, 0.5, 0.67, 0.75]$ and found $\lambda = 0.33$ would yield good performance across most models and datasets.

5.4 Active Learning Setting

As mentioned in Section 1, a notable difference between traditional NLP tasks such as NER and the hallucination mitigation we are considering is the difficulty of annotation. Annotating for hallucination is far more challenging than annotating for NER or other classification tasks. In hallucination mitigation, there is no clear standard of what is correct or incorrect, making hallucination annotation a highly demanding task for annotators (Zhao et al., 2023; Guerreiro et al., 2023). With this consideration in mind, we design a low-resource active learning setting similar to Tsvigun et al. (2022) and Jukić and Snajder (2023) that models the difficulty of obtaining human annotations for hallucination, thereby approximating a practical scenario.

Specifically, in each active learning iteration, only 1 sample from the unlabeled pool will be selected and annotated, which is approximately 0.05% of the total data samples. Following the convention of previous active learning works on annotation emulation (Shen et al., 2018; Ein-Dor et al., 2020; Radmard et al., 2021; Shelmanov et al., 2021; Tsvigun et al., 2022), we use the ground truth, i.e., gold summaries in text summarization, to emulate the human-annotated samples. After annotation, the model is finetuned with the annotated sample. Note that we use standard supervised finetuning here with selected samples in each step. Following Radmard et al. (2021), we then evaluate the model on a validation set and load the previous optimal model weights if the performance decreases after finetuning. The model is then evaluated on the test set and the performance is recorded. Following Tsvigun et al. (2022), we run the active learning loop for 100 iterations with 100 annotations in total for each experiment. We use an AdamW optimizer with a learning rate of 5e-5. All experiments run on 8×RTX2080Ti GPUs and are repeated 5 times.

6 Results

6.1 Performance Comparison

We present the main evaluation results in Table 2, using 30% of the annotation budget to assess methods in a low-resource setting, accounting for

Model	Method	CNN-DailyMail				Multi-News				Gigaword			
		BERT-P	UniEval	FactKB	ROUGE-L	BERT-P	UniEval	FactKB	ROUGE-L	BERT-P	UniEval	FactKB	ROUGE-L
Flan-T5 Small	$\text{HADAS}_{\text{w/ S.F.}}$	76.96	72.34	84.23	15.81	69.82	53.98	71.37	10.72	55.57	36.04	9.03	21.92
	$\text{HADAS}_{\text{w/ Disc.}}$	76.62	73.91	83.14	16.45	67.80	50.77	67.38	9.84	59.49	**41.95**	8.90	23.69
	$\text{HADAS}_{\text{w/ C.V.}}$	77.84	73.82	84.22	16.55	69.84	49.10	66.66	9.89	60.94	31.88	8.31	21.40
	HADAS	**78.63**	**75.75**	**87.46**	**16.55**	**70.26**	**56.40**	**74.22**	**11.04**	**61.06**	40.53	**10.85**	**23.89**
Flan-T5 Base	$\text{HADAS}_{\text{w/ S.F.}}$	72.54	67.08	**80.93**	16.64	68.65	58.75	62.76	9.35	55.75	43.37	6.19	27.32
	$\text{HADAS}_{\text{w/ Disc.}}$	72.44	67.64	75.19	16.92	69.28	60.59	65.14	9.18	58.16	41.70	8.88	29.14
	$\text{HADAS}_{\text{w/ C.V.}}$	72.56	66.68	79.34	15.97	69.26	53.53	60.43	8.97	58.27	44.74	7.66	27.09
	HADAS	**73.74**	**70.31**	80.73	**17.19**	**70.82**	**61.12**	**66.39**	**9.87**	**59.36**	**47.98**	**9.18**	**29.25**
BART Base	$\text{HADAS}_{\text{w/ S.F.}}$	76.85	74.02	93.00	19.41	67.35	53.15	**81.47**	13.10	85.39	55.37	61.95	34.43
	$\text{HADAS}_{\text{w/ Disc.}}$	76.68	76.42	92.39	17.47	68.41	53.63	70.55	11.33	82.98	60.27	53.69	35.65
	$\text{HADAS}_{\text{w/ C.V.}}$	76.86	71.61	86.59	19.38	69.58	52.69	72.28	12.37	78.47	61.60	45.19	34.35
	HADAS	**78.14**	**76.65**	**93.95**	**20.12**	**71.03**	**55.94**	80.22	**14.83**	**87.59**	**63.75**	**70.12**	**35.91**

Table 3: Ablation results of summarization factuality and quality metrics with 30% of hallucination annotations across models and datasets, where the best results are highlighted in bold, and the second-best are underscored.

the challenge of annotating hallucinations. Results utilizing the full annotation budget are presented in Figures 2 and 3 and discussed in 6.2.

As shown in Table 2, HADAS consistently achieves the best results in hallucination evaluation metrics, spanning three different types, across all metrics and datasets. It also maintains high summarization qualities as measured by ROUGE-L. This demonstrates the effectiveness of our hallucination diversity-aware sample selection strategy. We also observe that, while IDDS shows a consistent advantage over the random baseline, its improvements are modest compared to those of HADAS. Moreover, the variant of our proposed method, $\text{HADAS}_{\text{w/o Div.}}$, shows clear improvements on CNN-DailyMail. However, it does not consistently outperform IDDS on the other two datasets, and it even performs worse than the random baseline in Multi-News with the BART model. We attribute the unsatisfying performance to the greedy strategy of selecting samples that do not adequately cover the different hallucination types. We attribute this unsatisfying performance to the greedy strategy of selecting samples that do not adequately cover different hallucination types. As a result, LLMs may not comprehensively encounter cases prone to hallucination during the finetuning process. This underscores the importance of considering hallucination diversity in HADAS.

6.2 Efficiency Comparison

In Figure 2 and 3, we present the performance curves based on full hallucination annotations. Due to limited space, we only show representative curves for Flan-T5 Small and Flan-T5 Base on the CNN-DailyMail dataset and similar trends are observed in other experiments.

From Figure 2 and 3, we observe that HADAS's performance increases rapidly in the early stages, indicating that it selects more informative hallucination samples. Although most methods converge to comparable performance levels with more annotations, the swift improvement of HADAS underscores the efficiency of our method. This efficiency is particularly valuable in practical applications, given the high costs and challenges associated with hallucination annotations. Additionally, we note that while $\text{HADAS}_{\text{w/o Div.}}$ also shows quick initial growth, its pace slows down, and it eventually gets outperformed by IDDS with more annotations. This phenomenon suggests that while a greedy selection may be beneficial in the short term, it might not lead to better outcomes in the long run, emphasizing the importance of considering the diversity of hallucination samples.

6.3 Ablation Studies

To further demonstrate the effectiveness of considering hallucination diversity, we conducted ablation experiments evaluating HADAS's performance when measuring only a single type of hallucination. The results, presented in Table 3, clearly show that focusing on a single hallucination type does contribute to reducing that specific type of hallucination. For instance, $\text{HADAS}_{\text{w/ S.F.}}$ mostly achieves the best or second-best performances on FactKB, as it specifically targets semantic frame errors. Similar patterns are observed for $\text{HADAS}_{\text{w/ Disc.}}$ and $\text{HADAS}_{\text{w/ C.V.}}$. However, these singular measurements alone are not sufficient for comprehensive hallucination mitigation, as some performances are even worse than the random baseline in addressing

different types of hallucinations. These ablation results further highlight HADAS's advantage in considering hallucination diversity during sample selection, consistently achieving most of the best performances across all metrics.

7 Conclusion

In this work, we propose the first active learning framework to mitigate hallucinations in LLMs, reducing the need for intensive human annotation efforts. By measuring various types of hallucinations in text summarization and developing a novel hallucination diversity-aware sample selection method, we effectively and efficiently mitigate LLM hallucinations in summarizations in a comprehensive manner. Extensive experiments on several datasets and backbone models demonstrate the advantages of our method across various factuality metrics while maintaining high summarization quality.

Limitations

Despite the promising results, our proposed method depends on existing hallucination detection methods to identify diverse hallucinations. The selection of appropriate hallucination detection metrics requires extra attention to ensure they can effectively capture various types of hallucinations. As we discussed in Section 4.2, we have selected three types of hallucination detection models based on empirical results from previous works. However, these models may not be perfectly suited for our purposes in detecting specific hallucination types.

Our primary contribution is the development of a generic active learning framework for hallucination mitigation, which offers the flexibility to easily integrate additional hallucination detection methods. We plan to conduct more comprehensive experiments using more fine-grained and interpretable hallucination detection methods in future work.

Additionally, in our experiments, we followed the practices of prior active learning studies by using ground-truth data to emulate human annotations, specifically gold summaries in our context. However, recent works suggest that these gold summaries might also contain hallucinated content. Although we have intentionally chosen datasets with more reliable gold summaries, conducting experiments with actual human annotations would be highly beneficial to further evaluate the effectiveness of our active learning framework.

Ethical Consideration

Active learning inherently involves biased sampling, which can potentially result in datasets with biased annotations. Consequently, this approach can be intentionally employed to amplify existing biases within datasets. Our research enhances the effectiveness of hallucination mitigation, thereby also increasing its capacity to introduce hallucinations more efficiently. Therefore, extra cautions are needed for any practical application of our method.

References

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. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, Nusa Dua, Bali. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Meng Cao, Yue Dong, and Jackie Cheung. 2022. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3354, Dublin, Ireland. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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*.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky,

Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir R. Radev. 2019. Multi-news: a large-scale multi-document summarization dataset and abstractive hierarchical model.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. 2023. FactKB: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, Singapore. Association for Computational Linguistics.

Rick Groenendijk, Sezer Karaoglu, Theo Gevers, and Thomas Mensink. 2021. Multi-loss weighting with coefficient of variations. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1468–1477.

Nuno M. Guerreiro, Duarte M. Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André F. T. Martins. 2023. Hallucinations in Large Multilingual Translation Models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Josip Jukić and Jan Snajder. 2023. Parameter-efficient language model tuning with active learning in low-resource settings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5061–5074, Singapore. Association for Computational Linguistics.

Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale N Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. *Advances in Neural Information Processing Systems*, 35:34586–34599.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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.

Yixiao Li, Yifan Yu, Chen Liang, Nikos Karampatziakis, Pengcheng He, Weizhu Chen, and Tuo Zhao. 2024. Loftq: LoRA-fine-tuning-aware quantization for large language models. In *The Twelfth International Conference on Learning Representations*.

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

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.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *The Twelfth International Conference on Learning Representations*.

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 2021. Entity-level factual consistency of abstractive text summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2727–2733, Online. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023. Active learning for natural language generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9862–9877, Singapore. Association for Computational Linguistics.

Puria Radmard, Yassir Fathullah, and Aldo Lipani. 2021. Subsequence based deep active learning for named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4310–4321, Online. Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Paul Roit, Johan Ferret, Lior Shani, Roee Aharoni, Geoffrey Cideron, Robert Dadashi, Matthieu Geist, Sertan Girgin, Leonard Hussenot, Orgad Keller, Nikola Momchev, Sabela Ramos Garea, Piotr Stanczyk, Nino Vieillard, Olivier Bachem, Gal Elidan, Avinatan Hassidim, Olivier Pietquin, and Idan Szpektor. 2023. Factually consistent summarization via reinforcement learning with textual entailment feedback. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6252–6272, Toronto, Canada. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V. Dylov, and Alexander Panchenko. 2021. Active learning for sequence tagging with deep pre-trained models and Bayesian uncertainty estimates. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1698–1712, Online. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

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*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):13618–13626.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro,

Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Akim Tsvigun, Ivan Lysenko, Danila Sedashov, Ivan Lazichny, Eldar Damirov, Vladimir Karlov, Artemy Belousov, Leonid Sanochkin, Maxim Panov, Alexander Panchenko, Mikhail Burtsev, and Artem Shelmanov. 2022. Active learning for abstractive text summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5128–5152, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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*.

David Wan and Mohit Bansal. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Junda Wu, Rui Wang, Tong Yu, Ruiyi Zhang, Handong Zhao, Shuai Li, Ricardo Henao, and Ani Nenkova. 2022. Context-aware information-theoretic causal de-biasing for interactive sequence labeling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3436–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36.

Yu Xia, Fang Kong, Tong Yu, Liya Guo, Ryan A Rossi, Sungchul Kim, and Shuai Li. 2024. Which llm to play? convergence-aware online model selection with time-increasing bandits. *arXiv preprint arXiv:2403.07213*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2023. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*.

Yue Yu, Lingkai Kong, Jieyu Zhang, Rongzhi Zhang, and Chao Zhang. 2022. AcTune: Uncertainty-based active self-training for active fine-tuning of pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1422–1436, Seattle, United States. Association for Computational Linguistics.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

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*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57.

Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A survey of active learning for natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6166–6190, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv:2303.18223*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A LLaMa-2 Experiments

We also conduct some experiments on the LLaMa-2 7B to show the effectiveness of our method on larger models in mitigating hallucinations. Similarly in Section 5.1, we first finetune the model on

Model	Method	CNN-DailyMail				Multi-News				Gigaword			
		BERT-P	UniEval	FactKB	ROUGE-L	BERT-P	UniEval	FactKB	ROUGE-L	BERT-P	UniEval	FactKB	ROUGE-L
LLaMa-2 7B	Random	57.72	52.42	75.44	14.28	55.69	62.02	60.74	10.33	60.78	58.28	21.38	21.13
	IDDS	58.61	58.45	79.80	16.15	53.73	61.04	61.08	**11.35**	63.20	60.61	21.12	23.77
	$\text{HADAS}_{\text{w/o Div.}}$	58.19	58.90	78.84	15.58	58.67	63.82	61.73	10.57	63.56	60.12	24.02	23.49
	HADAS	**60.34**	**60.61**	**83.52**	**16.75**	**58.79**	**64.28**	**64.89**	10.93	**64.77**	**62.35**	**24.26**	**23.81**

Table 4: Preliminary results of summarization factuality and quality metrics with 30% of hallucination annotations of LLaMa-2 7B on all datasets, where the best results are highlighted in bold, and the second-best are underscored.

the XSum dataset to ensure summarization quality. Specifically, we use LoRA (Hu et al., 2022) finetuning with a randomly selected subset of 5000 samples with the prepended instruction prompt `"Summarize the following article:"`. The LoRA rank is set to 8 and alpha is set to 16. An AdamW optimizer is used with a learning rate of 5e-4. We then use the LLaMa-2 7B model finetuned on the XSum dataset as the initial point for active learning. The same instruction prompt is used for experiments on all three datasets. The rest of the active learning settings, dataset construction, and hyperparameter selection in the experiments remain the same as detailed in Section 5, except that we also apply LoRA finetuning with the above configuration during the active learning process. The experiments run on 2×A40 GPUs and are repeated 3 times. The results are shown in Table 4.

From Table 4, we observe that HADAS consistently outperforms baselines on most evaluation metrics. The results again validate advantages of our method as similarly observed in Section 6. Note that in these experiments we do not optimize the training configurations and parameters, which may influence the LoRA finetuning performances of LLaMa-2 models as suggested by Li et al. (2024) and lead to sub-optimal results. We leave more comprehensive evaluations of our hallucination mitigation method on LLMs of larger parameter sizes as future work.