Enhancing Faithfulness in Abstractive Summarization via Span-Level Fine-Tuning

Sicong Huang **Qianqi Yan** **Shengze Wang** **Ian Lane**
University of California, Santa Cruz
{shuan213, qyan79, swang488, ialane}@ucsc.edu

Abstract

Abstractive summarization using large language models (LLMs) has become an essential tool for condensing information. However, despite their ability to generate fluent summaries, these models sometimes produce unfaithful summaries, introducing hallucinations at the word, phrase, or concept level. Existing mitigation strategies, such as post-processing corrections or contrastive learning with synthetically generated negative samples, fail to fully address the diverse errors that can occur in LLM-generated summaries. In this paper, we investigate fine-tuning strategies to reduce the occurrence of unfaithful spans in generated summaries. First, we automatically generate summaries for the set of source documents in the training set with a variety of LLMs and then use GPT-4o to annotate any hallucinations it detects at the span-level. Leveraging these annotations, we fine-tune LLMs with both hallucination-free summaries and annotated unfaithful spans to enhance model faithfulness. In this paper, we introduce a new dataset that contains both faithful and unfaithful summaries with span-level labels and we evaluate three techniques to fine-tuning a LLM to improve the faithfulness of the resulting summarization: *gradient ascent*, *unlikelihood training*, and *task vector negation*. Experimental results show that all three approaches successfully leverage span-level annotations to improve faithfulness, with *unlikelihood training* being the most effective.

1 Introduction

Abstractive summarization aims to condense a piece of text into a shorter version by distilling the key information from the source text and rewriting it in a concise manner. Recent advances in large language models (LLMs) such as GPT-4(o) (OpenAI et al., 2024b;a), Gemini (Team et al., 2024), Llama-3 (Grattafiori et al., 2024), and Qwen-2.5 (Qwen et al., 2025) have significantly enhanced the capabilities of summarization systems to produce fluent and coherent summaries. Additionally, the growing integration of retrieval-augmented generation (RAG) (Lewis et al., 2020; Siriwardhana et al., 2023; Zhang et al., 2024) has underscored the role of summarization as a critical component of modern interactive natural language systems.

However, despite the strong capabilities of LLMs, they still suffer from the problem of hallucination (Huang et al., 2023; Ji et al., 2023; Jiang et al., 2024), often referred to as unfaithfulness in summarization (Maynez et al., 2020; Goyal & Durrett, 2021; Kryscinski et al., 2020). This issue arises when the generated summary contains information that is neither grounded in nor aligned with the source document, limiting the practicality of deploying summarization systems in real applications. Figure 1 shows an example SAMSum (Gliwa et al., 2019) dialogue and its corresponding generated summaries that contains spans of hallucinated text.

A number of approaches have attempted to alleviate unfaithfulness with post-processing. For instance, Dong et al. (2020) and Cao et al. (2020) have suggested methods to edit and correct factual inaccuracies in summaries post-generation. Similarly, Madaan et al. (2023); Akyurek et al. (2023) employ the critique-and-refine process that generates critical feedback on the

Source Dialogue:

Pam:	Hey Robert, you said you could help with Tom's birthday?
Robert:	Sure, what do you need?
Pam:	I have to go shopping, cook, and clean, and I figured out I don't have time to pick up the balloons.
Robert:	From where?
Pam:	There's this store in the city centre that sells these awesome floating balloons.
Robert:	No problem, just text me the address.
Pam:	Bless you!
Robert:	;)

Baseline Summary:

Pam asked Robert for help with Tom's birthday celebration, as she needs to go shopping, cook, and clean, and doesn't have time to pick up floating balloons from a store in the city centre. Robert agreed to help by providing the address of the store .

Gradient Ascent Summary:

Pam asked Robert for help with Tom's birthday celebration, including picking up floating balloons from a store in the city centre. Robert agreed to help and requested the store's address.

Unlikelihood Summary:

Pam asked Robert if he could help with Tom's birthday celebration, specifically asking for his assistance in picking up floating balloons from a store in the city centre. Robert agreed to help and requested the store's address.

Task Vector Negation Summary:

Pam asked Robert for help with Tom's birthday celebration, including shopping, cooking, and cleaning. Robert agreed to help and Pam provided the address of a store in the city centre where she needed him to pick up floating balloons .

Figure 1: An example from the SAMSum dataset, showing a dialogue and corresponding summaries generated from models fine-tuned using four different approaches. Unfaithful spans are highlighted .

initial summary as a guide for the summarizer to refine the summary. Although effective, the extra post-processing steps required induce high latency and increase the computational demands during inference, restricting their applicable use cases. Another approach is to learn from negative samples. Cao & Wang (2021); Tang et al. (2022); Zhang et al. (2023); Qiu et al. (2024) synthetically create negative samples of unfaithful summaries. These samples are derived from reference summaries using strategies that mimic common error types. However, there are three problems with this approach:

1. Human reviewers generally prefer LLM summaries over standard reference summaries, even for large and well-established summarization datasets (Sottana et al., 2023; Goyal et al., 2023; Liu et al., 2024), rating them highly across all aspects of evaluation. This preference reveals the poor quality of reference summaries, raising concerns about the effectiveness of fine-tuning models on reference summaries and their perturbed version.
2. Synthetic negative samples generated from common approaches often fail to replicate the actual errors observed in model-generated summaries. Furthermore, the error distributions in generated summaries can vary significantly across different domains, rendering synthetic negative sample generation approaches insufficient to cover the wide variety of error types (Goyal & Durrett, 2021).
3. Contrastive learning approaches (Cao & Wang, 2021; Tang et al., 2022; Zhang et al., 2023) lack utilization of detailed, span-level information that could potentially improve summary faithfulness more effectively (Goyal & Durrett, 2021). In cases of unfaithfulness within LLM-generated summaries, typically only a few specific text spans are unfaithful. Only these problematic spans should be specifically targeted as negative examples during model training.

To address these challenges, we propose annotating span-level hallucinations in LLM-generated summaries and then leveraging this fine-grained information to update the model. Our contributions are threefold: (1) we construct a novel dataset containing LLM-generated

summaries, labeled at the span-level for faithfulness; (2) we evaluate three fine-tuning methods, namely gradient ascent, unlikelihood training and task vector negation to utilize the span-level unfaithful information to enhance LLM summarization faithfulness. We show that all three approaches improve faithfulness, with the latter two being more effective; and (3) we analyze how each method is affected by the weight given to negative samples (ϵ), revealing that unlikelihood training is the most stable of the three methods.

2 Related work

2.1 Improving summarization faithfulness

A number of prior approaches to improving faithfulness focus on post-processing. Dong et al. (2020) uses QA span fact correction models to revise entities in generated summaries to boost factual consistency. Similarly, Cao et al. (2020) corrects factual errors in generated summaries by training a corrector model on artificially created error data.

Other prior works leveraged synthetic negative sample summaries to improve faithfulness. Cao & Wang (2021) surveyed common errors that summarization models tend to make and designed strategies for constructing negative samples (e.g., entity swap, mask and regenerate) that corrupt the reference summaries. They then used contrastive learning to better discriminate between positive and negative examples, improving the representation and faithfulness during generation. Similarly, Tang et al. (2022) designed a linguistically informed taxonomy of factual errors for dialogue summaries and created synthetic negative samples based on the taxonomy before applying contrastive learning to improve faithfulness. Laban et al. (2023) proposed a cost effective protocol to create more natural sounding and manually verified synthetic negative samples, which can be used as training data for contrastive learning.

Chen et al. (2021) proposes to generate multiple contrastive candidate summaries featuring different entities from the source document, subsequently employing a discriminative model trained to differentiate between faithful summaries and synthetic negative ones, effectively ranking the candidates. Goyal & Durrett (2021) tried to tackle the problem from the perspective of training data. They demonstrated an approach to improve faithfulness by identifying unsupported facts in the training summaries and ignore the corresponding tokens during training. Goyal et al. (2022) took a closer look at the training dynamics and found that longer training on noisy datasets contributes to factual inconsistency. They show that using token sub-sampling to dynamically modify the loss computation during training, down weighting high-loss tokens, can substantially improve factual consistency.

Some more recent methods have focused on the critique-and-refine approach. Akyurek et al. (2023) implemented a reinforcement learning framework where a critic model provides feedback on generated summaries. The summarizer then refines its output based on this feedback, with the summarization task metric serving as a reward for the critic model. This process enables the critic to guide the summarizer towards improving specific performance metrics, fostering a cycle of continuous improvement.

2.2 Span-level hallucination labeling

There has been a large amount of work exploring evaluation metrics for summarization faithfulness (Zhang* et al., 2020; Yuan et al., 2021; Liu et al., 2023; Zha et al., 2023). However, the study of span-level hallucination labeling remains relatively underexplored. Zhou et al. (2021) proposed a task to predict token-level hallucinations in summarization and machine translation and introduced methods to create and fine-tune on synthetic data to solve the task. Goyal & Durrett (2021) presented an approach to label hallucinations in summaries at the dependency arc level, providing more fine-grained information. Though not specific to summarization, Liu et al. (2022) introduced a token-level reference-free hallucination detection benchmark and created multiple baselines. Niu et al. (2024) introduced RAGTruth, a large-scale hallucination corpus with span-level annotations designed for retrieval-augmented generation (RAG). While primarily focused on hallucination detection in RAG-based applications, the dataset includes extensive word-level annotations across various tasks, including summarization. Their experiments demonstrate that fine-tuning on high-quality annotated data can significantly improve hallucination detection over prompt-based and self-verification methods.

3 Methods

3.1 Problem formulation

As illustrated in Figure 3, training data comprises of both positive and negative samples of document-summary pairs. For each positive sample, the document D has a corresponding positive summary S_p where $S_p = (x_{p1}, x_{p2}, ..., x_{pT})$. Similarly, each negative sample includes a document D paired with a negative summary S_n where $S_n = (x_{n1}, x_{n2}, ..., x_{nT})$. Here, x denotes a token in the summary, and T denotes the number of tokens in the summary. Note that not all x_n tokens are considered hallucinated. Therefore, we use H_n to denote the set of hallucinated tokens in a negative sample summary S_n and $H_n \subseteq S_n$.

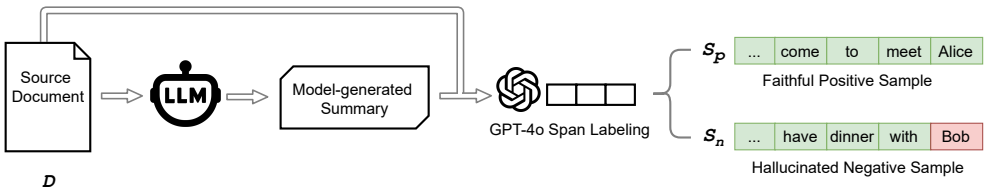

Figure 2: **Training data construction:** summaries of the source documents used for model training are generated using an LLM. Spans of text in the generated summaries that are unfaithful to the source document are automatically labeled using GPT-4o (using the prompt in Appendix A.3). Summaries that have no unfaithful spans labeled in their output are treated as positive training samples, and summaries that contain unfaithful spans are treated as negative training samples.

3.2 Fine-tuning methods to improve summarization faithfulness

We compare three methods for model fine-tuning that can take advantage of both positive faithful summaries and negative unfaithful summaries with span-level hallucination labels. To manage the influence of positive and negative samples, we introduce a hyperparameter $\epsilon \in [0, 1]$ to distribute the weights assigned to each type of summary. Specifically, ϵ specifies the weight of negative samples, and its complement $1 - \epsilon$ specifies the weight of positive samples.

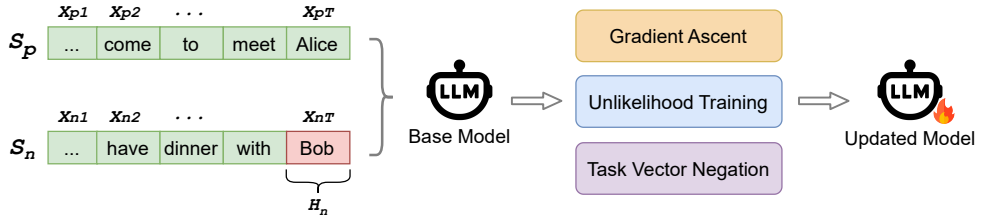

Figure 3: **Model update:** a base model is updated using both the faithful positive example summaries and the unfaithful negative example summaries with hallucination spans using one of three approaches we compare in this paper (1) Gradient Ascent, (2) Unlikelihood Training or (3) Task Vector Negation.

Gradient ascent Recent research by Yao et al. (2024) has demonstrated that unlearning undesirable behaviors in LLMs through gradient ascent is an effective alignment technique. This method has been used to steer LLM outputs away from harmful content, copyright infringements, and to reduce hallucinations. In our study, we apply gradient ascent to decrease hallucinated content in summarization by minimizing the probability of generating unfaithful tokens. This is achieved by reversing the sign of the cross-entropy loss. We define the gradient ascent loss as follows:

$$L_{ga} = \begin{cases} -(1-\epsilon) \sum\limits_{x_p \in S_p} \log p_\theta(x_p|.) & \text{if } S_p \\ \epsilon \sum\limits_{x_n \in H_n} \log p_\theta(x_n|.) & \text{if } S_n \end{cases}$$

Here, p_θ represents the model parameterized by θ. The loss computation varies depending on the sample type: for a positive sample S_p, the negative log-likelihood (NLL) loss is calculated with a weight of $(1-\epsilon)$; for a negative sample S_n, the loss is computed as the negative NLL, multiplied by the weight ϵ.

Unlikelihood training Unlikelihood training, initially introduced by Welleck et al. (2020), was designed to mitigate common issues such as dull and repetitive outputs from language models while maintaining perplexity. Subsequent studies (Li et al., 2020) have demonstrated its effectiveness in addressing problems like excessive copying from context, frequent word overuse, and logical inconsistencies in generated texts. The adaptability of this method allows for its application across various tasks, particularly in reducing undesirable outputs. In this method, the model is trained to assign lower probabilities to unwanted generations by maximizing the complement of the probability of generating such tokens. For our specific application in summarization, we focus on minimizing the generation of hallucinated tokens. We define the unlikelihood loss as follows:

$$L_{ul} = \begin{cases} -(1-\epsilon) \sum\limits_{x_p \in S_p} \log p_\theta(x_p|.) & \text{if } S_p \\ -\epsilon \sum\limits_{x_n \in H_n} \log(1 - p_\theta(x_n|.)) & \text{if } S_n \end{cases}$$

This configuration computes the negative log-likelihood (NLL) for positive samples, weighted by $(1-\epsilon)$, and for negative samples, it calculates the NLL of the complement probability $(1-p_\theta(x_n|.))$ weighted by ϵ. This approach ensures a strategic penalization of unfaithful tokens, thus aiming to enhance the overall faithfulness of the generated summaries.

Task vector negation Task vector arithmetic offers a straightforward method for modifying a pre-trained model to promote desired behaviors, as outlined by Ilharco et al. (2023). This technique involves calculating a task vector, τ_t, for a specific task t by subtracting the weights of the base pre-trained model (θ_{pre}) from the weights of a model that has been fine-tuned on that task (θ_{ft}). Formally, the task vector is defined as $\tau_t = \theta_{ft} - \theta_{pre}$. Building on this concept, Ilharco et al. (2023) demonstrated that negating a task vector fine-tuned on toxic language can effectively diminish the generation of toxic outputs. Inspired by this, we apply a similar approach to address the generation of unfaithful summaries. We define the resulting model's weights as:

$$\theta_{res} = \theta_{pre} + (1-\epsilon)\tau_{pos} - \epsilon\tau_{neg}$$

τ_{pos} represents the task vector derived from fine-tuning on positive faithful summaries, while τ_{neg} is obtained from fine-tuning on negative hallucinated summaries. This method allows us to manipulate the model composition to reduce the production of unfaithful summaries by strategically negating the influence of the undesired traits encoded in τ_{neg}.

4 Data

Recent studies (Sottana et al., 2023; Goyal et al., 2023) have show that even open-source LLMs outperform the reference summaries in well-established datasets, which are often rated poorly on *relevance*, *fluency*, *coherence*, and *consistency*. Additionally, prior methods that generate synthetic negative samples based on common error types (Cao & Wang, 2021; Tang et al., 2022; Zhang et al., 2023) fail to capture real hallucination patterns in LLM outputs (Goyal & Durrett, 2021). Error distributions also vary across domains, making synthetic samples insufficient. To address these issues, we construct a dataset of real LLM-generated summaries that are subsequently annotated for hallucinations at the span-level.

4.1 Dataset construction

The dataset construction process is illustrated in Figure 2. We construct the training and test set from CNNDM (Nallapati et al., 2016), SAMSum (Gliwa et al., 2019), and XSum (Narayan et al., 2018) covering both news and dialogue, two of the most studied domains of summarization research. Summaries of their source documents were generated using four LLMs, `Llama3.2-1b`, `SmolLM2-1.7b`, `OLMo2-7b`, and `Llama3.1-8b` (Grattafiori et al., 2024; Allal et al., 2025; OLMo et al., 2025), two models for each size class. We set top_p = 0.7

and temperature = 0.01 to minimize randomness. The prompts we used to generate these summaries are provided in Appendix A).

Sottana et al. (2023) shows that top-tier LLMs correlate highly with human judgments when acting as a reviewer. Expanding upon this prior work we use GPT-4o (OpenAI et al., 2024a) to label LLM-generated summaries to identify "spans of text that are inconsistent with the source document." The prompt we use to label these spans is provided in A.4). When a summary has no labeled span in it's output, it is considered a positive example of a summary; if a span has been labeled as "inconsistent with the source document" the summary is treated as a negative example.

4.2 Dataset statistics

Our dataset consists of 111,897 training samples and 2,819 unlabeled test samples (1,000 CNNDM, 819 SAMSum and 1,000 XSum). Table 1 presents the training set's distribution of positive and negative samples across datasets and models. A clear trend emerges: smaller models are more prone to generating unfaithful summaries. Specifically, Llama3.2-1b produces unfaithful summaries in 64% of CNNDM, 83% of SAMSum, and 70% of XSum instances, whereas Llama3.1-8b exhibits significantly lower rate of unfaithfulness—17% for CNNDM, 30% for SAMSum, and 25% for XSum.

Additionally, Table 1 shows that smaller models not only generate more unfaithful summaries overall but also produce a higher proportion of hallucinated tokens, as measured by the Hallucinated Token Ratio (HTR). In SAMSum, the dataset where models tend to be the most unfaithful, Llama3.2-1b generates, on average, 42.6% hallucinated tokens in unfaithful summaries, compared to 22.9% for Llama3.1-8b. This pattern remains consistent across all three datasets, reinforcing the correlation between model size and summary faithfulness.

Model	**CNNDM**			**SAMSum**			**XSum**			**Total**
	Pos	**Neg**	**HTR**	**Pos**	**Neg**	**HTR**	**Pos**	**Neg**	**HTR**	
Llama3.2-1b	2751	4814	0.141	1276	6384	0.426	2470	5796	0.198	23491
SmolLM2-1.7b	5050	6067	0.106	5492	5981	0.300	2761	3576	0.155	28927
OLMo2-7b	11649	3328	0.098	2889	1264	0.255	5250	3372	0.142	27752
Llama3.1-8b	12287	2570	0.080	4707	2047	0.229	7637	2479	0.116	31727
Total	31737	16779		14364	15676		18118	15223		111897

Table 1: The training set's distribution of faithful (**pos**) and unfaithful (**neg**) summaries across datasets and models, along with Hallucinated Token Ratio (**HTR**), the proportion of hallucinated tokens within unfaithful summaries.

5 Experimental setup

5.1 Model and training

We perform experiments on all 4 models used for generating the training set (i.e. `Llama3.2-1b`, `SmolLM2-1.7b`, `OLMo2-7b`, and `Llama3.1-8b`). For efficient fine-tuning, we adopt Low-Rank Adaptation (LoRA) (Hu et al., 2022) and apply low-rank update matrices to all linear modules, with rank $r = 128$, $\alpha = 256$, and dropout probability of LoRA layer = 0.05. With these settings, trainable parameters ranges from 4% (Llama3.1-8b) to 7.8% (SmolLM2-1.7b). The training processes use an AdamW optimizer with learning rate = 5e-5 and linear learning rate scheduler with warm up ratio = 0.01. We set the batch size to 16 and train each model for one epoch on the training set for all experiments.

The baseline is supervised fine-tuning only on the positive portion of training data using regular cross-entropy loss, with no information from negative samples.

5.2 Evaluation

We evaluate summarization faithfulness using three reference-free automatic metrics: G-Eval, AlignScore, and BARTScore.

G-Eval (GE) (Liu et al., 2023) is a GPT-based evaluation metric that employs Chain-of-Thought (CoT) prompting (Wei et al., 2022) to generate intermediate reasoning steps before assigning a final score. We use GPT-4o as the underlying model and report its consistency score, which measures the factual alignment between the summary and the source document (Fabbri et al., 2021).

AlignScore (AS) (Zha et al., 2023) formulates faithfulness evaluation as a text-to-text information alignment problem. It estimates factual consistency by assessing how well the generated summary preserves key information from the source document. AlignScore is trained to generalize across diverse domains and has demonstrated strong performance in factuality alignment evaluation tasks.

BARTScore (BS) (Yuan et al., 2021) models evaluation as a sequence-to-sequence generation task. It measures faithfulness by computing the log probability of generating the source document from the summary using a BART model. A higher BARTScore indicates better faithfulness.

6 Results and discussion

The evaluation results for the supervised fine-tuning (SFT) baselines, in which the models have only been updated using faithful summaries (S_p) are shown in Table 2 and 3, and the results of the three model update methods explored in this paper are shown in Table 4 and 5. Additional results are provided in Appendix B.

Model	GE	AS	BS
Llama3.2-1b	4.035	0.868	-3.091
SmolLM2-1.7b	4.170	0.871	-3.150
OLMo2-7b	4.528	0.902	-3.200
Llama3.1-8b	4.581	0.900	-3.143

Table 2: SFT baseline results of four models on SAMSum.

Dataset	GE	AS	BS
CNNDM	4.733	0.900	-1.906
SAMSum	4.581	0.900	-3.143
XSum	4.585	0.906	-1.836

Table 3: SFT baseline results of Llama3.1-8b on three datasets.

6.1 Main results

Model	Gradient Ascent			Unlikelihood			Task Vector		
	GE	AS	BS	GE	AS	BS	GE	AS	BS
Llama3.2-1b	4.094↑	0.873↑	-3.099	**4.166**↑	**0.896**↑	-3.109	4.010	0.878↑	-3.160
SmolLM2-1.7b	4.176↑	0.876↑	-3.133↑	**4.342**↑	**0.896**↑	-3.139↑	4.231↑	0.890↑	-3.162
OLMo2-7b	4.546↑	0.906↑	-3.153↑	4.675↑	0.917↑	-3.164	**4.733**↑	**0.918**↑	-3.395
Llama3.1-8b	4.641↑	0.918↑	-3.158	4.749↑	0.918↑	-3.198	**4.760**↑	**0.928**↑	-3.213

Table 4: Results on the SAMSum dataset for the three studied methods on four models across G-Eval (GE), AlignScore (AS), and BARTScore (BS) evaluation metrics. Gradient ascent $\epsilon = 0.01$, unlikelihood $\epsilon = 0.5$, and task vector $\epsilon = 0.3$. The best result (in terms of G-Eval and AlignScore) for each model is **in-bold**. Scores higher than the baseline are marked with ↑.

First, in terms of G-Eval and AlignScore, regardless of the model, all studied methods show improvements from the baseline. Comparing Table 2 & 4, unlikelihood training yields greater improvements on smaller models (1b and 1.7b) and task vector negation show greater improvements on larger models (7b and 8b). The effectiveness of the studied methods is evident in Table 3 & 5 as well, observing improved G-Eval and AlignScore on all three datasets.

Second, in terms of G-Eval and AlignScore, unlikelihood training and task vector negation are more effective than gradient ascent. Gradient ascent, obtains higher scores than the

Dataset	Gradient Ascent			Unlikelihood			Task Vector		
	GE	AS	BS	GE	AS	BS	GE	AS	BS
CNNDM	4.737↑	0.901↑	-1.907	**4.796**↑	0.915↑	-1.875↑	4.785↑	**0.924**↑	-1.930
SAMSum	4.641↑	0.918↑	-3.158	4.749↑	0.918↑	-3.198	**4.760**↑	**0.928**↑	-3.213
XSum	4.632↑	0.909↑	-1.836	**4.702**↑	0.919↑	-1.808↑	4.686↑	**0.935**↑	-1.875

Table 5: Llama3.1-8b results of 3 studied methods on 3 datasets. Other settings are the same as Table 4. The best result (according to G-Eval and AlignScore) for each dataset is **in-bold**. Scores higher than the baseline are marked with ↑.

baseline across all models and datasets, but often only outperforms the baseline by a slight margin. This effect is further corroborated by Figure 4 that unlikelihood training and task vector negation show more consistent improvements than gradient ascent.

BARTScore, however, paints an inconsistent and often contradictory picture: the BARTScore of most results are lower than the baseline, while showing some improvements sporadically. Thus making it hard to draw conclusions from. We suspect the training objectives of the three methods we investigate render BARTScore unsuitable for evaluating faithfulness in this setting. We discuss this further in section 6.3.

6.2 Results of varying ϵ

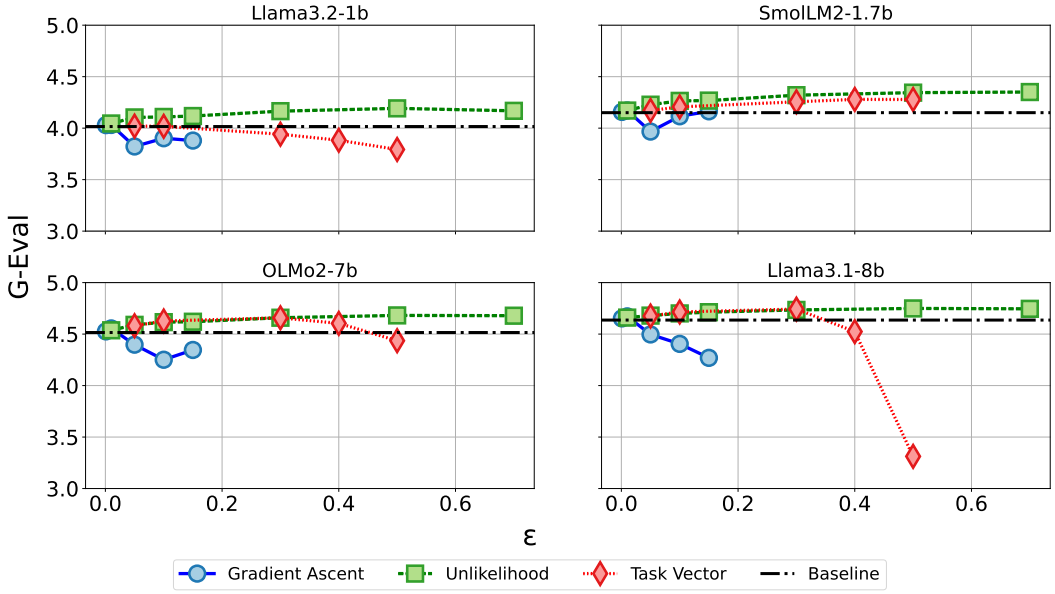

Figure 4: Average G-Eval on all three datasets with models trained using different ϵ.

In Section 3.2 we introduced a hyperparameter ϵ as the weight assigned to negative unfaithful samples during model fine-tuning. Figure 4 illustrates the effect of ϵ on the methods' performance. Gradient ascent slightly improves faithfulness when ϵ is small (0.01), but performance degrades as ϵ increases. Task vector negation can generally improve faithfulness, but performance degrades when ϵ becomes large (except for SmolLM2). Unlikelihood training is the most stable among the three methods, providing consistent improvements even with ϵ as large as 0.7. These results are consistent with the findings in Section 6.1 that all three methods can improve summarization faithfulness with unlikelihood training and task vector negation being more effective than gradient ascent.

6.3 Discussion

Effect of ϵ One prominent phenomenon we observe in Figure 4 and during training is the three effect of ϵ across the different methods. Gradient ascent only seems to perform well when ϵ is very small (i.e. 0.01). Increasing the weight of ϵ is detrimental to the resulting model's performance to the point of destroying the model's ability to generate coherent text. Task vector negation can tolerate moderate ϵ values and unlikelihood training is the most robust across different values of ϵ. We suspect the tolerance to ϵ is an important factor that determines a method's ability to leverage negative/unfaithful span-level information to improve faithfulness. This can help explain why gradient ascent only slightly improve faithfulness while the other two methods observe greater improvements.

The inconsistency of BARTScore Although BARTScore is a widely used metric for evaluating summarization faithfulness, in our experiments, it performs inconsistently and does not correlate well with the other metrics used in this evaluation. For example, Table 2 & 4 shows that most finetuned models are lower in BARTScore than the baseline, and gradient ascent has higher BARTScores than unlikelihood training and task vector negation. These opposite conclusions from BARTScore make us suspect the negative token probability-minimizing objective of our training methods results in lower BARTScores, since BARTScore is essentially the average token probability. This finding challenges the premise of BARTScore: positive correlation between token probabilities and faithfulness. Thus we include BARTScore for reference but do not draw conclusions from it.

7 Conclusion

In this study, we propose leveraging span-level hallucination annotations to improve the faithfulness of LLM-generated summaries. To this end, we construct a dataset containing organically generated summaries labeled at the span level and evaluate three fine-tuning methods—*gradient ascent*, *unlikelihood training*, and *task vector negation*—that leverage unfaithful summary spans. Our results show that all three approaches improve summarization faithfulness, with unlikelihood training and task vector negation being more effective. Additionally, we analyze the impact of varying the weight of negative samples (ϵ) and find that unlikelihood training remains the most stable across a wide range of values. These findings highlight the effectiveness of span-level annotations in mitigating hallucinations and offer insights into robust fine-tuning strategies for improving LLM faithfulness.

8 Limitations

This work investigates the effectiveness of span-level faithfulness annotations in combination with fine-tuning methods that leverage the span-level information to reduce hallucinations in LLM-generated summaries. While our primary focus is on leveraging these annotations to improve faithfulness, the reliability of GPT-4o-based span annotation remains an open question. Although not a core contribution of this work, a more rigorous study of the annotation method is necessary to assess its accuracy and consistency. We acknowledge this limitation and intend to explore a more comprehensive validation in future work.

Moreover, Cao & Wang (2021) uses summary-level faithfulness information to improve summarization faithfulness, finding contrastive learning to be the most effective approach. Our work, being more fine-grained, focuses on leveraging span-level information. However, to perform contrastive learning on span (token)-level, requires a corresponding positive token for each negative token, which our current dataset does not provide. As a result, we are unable to include contrastive learning in this study.

Lastly, while our work explores fine-tuning techniques for improving faithfulness, we do not include comparisons with recent alignment methods such as Direct Preference Optimization (DPO Rafailov et al., 2023). We recognize the growing adoption of these techniques for aligning LLM behavior and plan to investigate their effectiveness in mitigating hallucinations in future work.

References

Afra Feyza Akyurek, Ekin Akyurek, Ashwin Kalyan, Peter Clark, Derry Tanti Wijaya, and Niket Tandon. RL4F: Generating natural language feedback with reinforcement learning for repairing model outputs. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7716–7733, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.427. URL https://aclanthology.org/2023.acl-long.427.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, Colin Raffel, Leandro von Werra, and Thomas Wolf. Smollm2: When smol goes big – data-centric training of a small language model, 2025. URL https://arxiv.org/abs/2502.02737.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. Factual error correction for abstractive summarization models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6251–6258, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.506. URL https://aclanthology.org/2020.emnlp-main.506.

Shuyang Cao and Lu Wang. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6633–6649, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.532. URL https://aclanthology.org/2021.emnlp-main.532.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5935–5941, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.475. URL https://aclanthology.org/2021.naacl-main.475.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. Multi-fact correction in abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9320–9331, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.749. URL https://aclanthology.org/2020.emnlp-main.749.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00373. URL https://doi.org/10.1162/tacl_a_00373.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-5409. URL https://aclanthology.org/D19-5409.

Tanya Goyal and Greg Durrett. Annotating and modeling fine-grained factuality in summarization. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur,

Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1449–1462, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.114. URL https://aclanthology.org/2021.naacl-main.114.

Tanya Goyal, Jiacheng Xu, Junyi Jessy Li, and Greg Durrett. Training dynamics for text summarization models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2061–2073, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.163. URL https://aclanthology.org/2022.findings-acl.163.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3, 2023.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, ..., and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6t0Kwf8-jrj.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. A survey on large language model hallucination via a creativity perspective, 2024.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.750. URL https://aclanthology.org/2020.emnlp-main.750.

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9662–9676, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.600. URL https://aclanthology.org/2023.emnlp-main.600.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In

H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4715–4728, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.428. URL https://aclanthology.org/2020.acl-main.428.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6723–6737, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.464. URL https://aclanthology.org/2022.acl-long.464.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153.

Yixin Liu, Kejian Shi, Katherine He, Longtian Ye, Alexander Fabbri, Pengfei Liu, Dragomir Radev, and Arman Cohan. On learning to summarize with large language models as references. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8647–8664, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.478. URL https://aclanthology.org/2024.naacl-long.478/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/K16-1028. URL https://aclanthology.org/K16-1028.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1797–1807, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1206. URL https://aclanthology.org/D18-1206/.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2 olmo 2 furious, 2025. URL https://arxiv.org/abs/2501.00656.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, ..., and Yury Malkov. Gpt-4o system card, 2024a. URL https://arxiv.org/abs/2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, ..., and Barret Zoph. Gpt-4 technical report, 2024b.

Haoyi Qiu, Kung-Hsiang Huang, Jingnong Qu, and Nanyun Peng. AMRFact: Enhancing summarization factuality evaluation with AMR-driven negative samples generation. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 594–608, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.33. URL https://aclanthology.org/2024.naacl-long.33/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics*, 11:1–17, 2023. doi: 10.1162/tacl_a_00530. URL https://aclanthology.org/2023.tacl-1.1.

Andrea Sottana, Bin Liang, Kai Zou, and Zheng Yuan. Evaluation metrics in the era of GPT-4: Reliably evaluating large language models on sequence to sequence tasks. In

Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8776–8788, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.543. URL https://aclanthology.org/2023.emnlp-main.543.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5657–5668, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.415. URL https://aclanthology.org/2022.naacl-main.415.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap ..., and Oriol Vinyals. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning, 2024.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. BARTScore: Evaluating generated text as text generation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=5Ya8PbvpZ9.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. AlignScore: Evaluating factual consistency with a unified alignment function. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11328–11348, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.634. URL https://aclanthology.org/2023.acl-long.634.

Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. FaMeSumm: Investigating and improving faithfulness of medical summarization. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10915–10931, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.673. URL https://aclanthology.org/2023.emnlp-main.673.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. Raft: Adapting language model to domain specific rag, 2024.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.),

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1393–1404, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.120. URL https://aclanthology.org/2021.findings-acl.120.

A LLM prompts used in experimentation

A.1 Prompt for LLM to generate summary - CNNDM

```
System prompt: You are an accurate summarizer that always writes concise
summaries of text that is as consistent with the source text as possible.
You will be given a piece of news article in the user prompt, and you need
to reply with a concise summary of the article as a response.

User prompt: {source_doc}
```

A.2 Prompt for LLM to generate summary - SAMSum

```
System prompt: You are an accurate summarizer that always writes concise
summaries of text that is as consistent with the source text as possible.
You will be given a dialogue conversation in the user prompt, and you need
to reply with a concise summary of the conversation in 2 sentences.

User prompt: {source_doc}
```

A.3 Prompt for LLM to generate summary - XSum

```
System prompt: You are an accurate summarizer that always writes concise
summaries of text that is as consistent with the source text as possible.
You will be given a piece of news article in the user prompt, and you need
to reply with a concise summary of the article as a response.

User prompt: {source_doc}
```

A.4 Prompt for GPT-4o to label spans

```
System prompt:

Your input fields are:
1. `source` (str): The source document.
2. `summary` (str): The summary of the source document.

Your output fields are:
1. `reasoning` (str)
2. `hallucinated_spans` (list[str]): Spans of the summary text that are
hallucinated i.e. unfaithful to the source (each span must be a substring of the
 summary text).

All interactions will be structured in the following way, with the appropriate
values filled in.
```

```
[[ ## source ## ]]
{source}

[[ ## summary ## ]]
{summary}

[[ ## reasoning ## ]]
{reasoning}

[[ ## hallucinated_spans ## ]]
{hallucinated_spans}        # note: the value you produce must be parseable
according to the following JSON schema: {"type": "array", "items": {"type":
"string"}}

[[ ## completed ## ]]

In adhering to this structure, your objective is:
    Analyze the provided source text and its summary to identify any portions of
 the summary that do not accurately reflect the source content. Provide a
detailed reasoning for your analysis and list any hallucinated spans-these are
sections of the summary text that introduce information not present in the
source or misrepresent the source details. Ensure that your reasoning aligns
with the source content to justify the identification of hallucinated spans, if
any.
```

Demo user prompt 1:

```
This is an example of the task.

[[ ## source ## ]]
Lenny: Babe, can you help me with something?
Bob: Sure, what's up?
Lenny: Which one should I pick?
Bob: Send me photos
Lenny: <picture>
Lenny: <picture>
Lenny: <picture>
Bob: I like the first ones best
Lenny: But I already have purple trousers. Does it make sense to have two pairs?
Bob: I have four black pairs :D :D
Lenny: yeah, but shouldn't I pick a different color?
Bob: what matters is what you'll give you the most outfit options
Lenny: So I guess I'll buy the first or the third pair then
Bob: Pick the best quality then
Lenny: ur right, thx
Bob: no prob :)

[[ ## summary ## ]]
Lenny asked Bob for help picking between three pairs of trousers, and Bob
recommended choosing the best quality option. Lenny was unsure about picking two
 pairs of the same color and Bob agreed, suggesting that having more options is
important.

Respond with the corresponding output fields, starting with the field `[[ ##
reasoning ## ]]`, then `[[ ## hallucinated_spans ## ]]` (must be formatted as a
valid Python list[str], and the list of spans must follow original order that
they appear in the summary), and then ending with the marker for `[[ ##
completed ## ]]`.
```

Demo model response 1:

[[## reasoning ##]]
Lenny was unsure about picking two pairs of the same color, but Bob did not agree with Lenny's reasoning. Instead, Bob suggested that having more options is important.

[[## hallucinated_spans ##]]
["Bob agreed"]

[[## completed ##]]

Demo user prompt 2:

This is an example of the task.

[[## source ##]]
ARLINGTON, Va. (AP) - Salvage crews have recovered an engine, large sections of fuselage, and are working to retrieve a wing from the wreckage of a commercial airliner involved in last week's midair collision near Washington's Reagan National Airport, officials said Monday.

More human remains have been recovered from the Potomac River, bringing the number of identified victims to 55 of the 67 aboard the two aircraft. The collision between an American Airlines jet and an Army Black Hawk helicopter was the deadliest U.S. air disaster since 2001.

Authorities say the operation to remove the plane will take several days, after which the helicopter wreckage will be retrieved. More than 300 responders are participating in the recovery, with two Navy barges assisting in lifting heavy debris. Divers and salvage workers are adhering to strict protocols, pausing recovery efforts when human remains are found.

The American Airlines jet, carrying 64 people, was en route from Wichita, Kansas, preparing to land when it collided with the Army helicopter, which had three personnel aboard on a training mission. There were no survivors. Passengers included figure skaters returning from the 2025 U.S. Figure Skating Championships and a group of hunters.

Family members of victims were escorted by police to the Potomac River on Sunday to pay their respects. Officials hope to recover the jet's cockpit on Tuesday, with parts of both aircraft transported to a hangar for investigation.

Federal investigators are analyzing conflicting altitude data from the collision. The jet's flight recorder showed an altitude of 325 feet (99 meters), while control tower data placed the Black Hawk at 200 feet (61 meters), the maximum permitted altitude for helicopters in the area. About a second before impact, the jet's pitch changed, but investigators have not determined whether it was an evasive maneuver.

Authorities warn against premature speculation about the cause of the crash or why the helicopter may have been above its altitude limit. "There are all kinds of reasons for altitude deviations-something as simple as a flock of birds or an obstacle," said Col. Mark Ott, deputy director of Army aviation.

The National Transportation Safety Board (NTSB) is working to reconcile the conflicting altitude readings using data from the helicopter's black box and refining the tower's data. All five air traffic controllers on duty at Reagan Airport during the collision have been interviewed. A preliminary report is expected within 30 days, though a full investigation may take over a year.

```
[[ ## summary ## ]]
A midair collision near Washington's Reagan National Airport between an American
 Airlines jet and an Army Black Hawk helicopter resulted in 76 fatalities,
making it the deadliest U.S. air disaster since 2001. Salvage crews have
recovered major wreckage, and the entire wreckage will be recovered within a
day. More human remains have been found, with 55 victims identified.
Investigators are analyzing conflicting altitude data, as the jet's flight
recorder showed 325 feet while control tower data placed the helicopter at 200
feet. The flight data recorder also indicated the jet made an evasive maneuver
right before the crash. Authorities caution against speculation as the National
Transportation Safety Board (NTSB) investigates, with a preliminary report
expected in 30 days.

Respond with the corresponding output fields, starting with the field `[[ ##
reasoning ## ]]`, then `[[ ## hallucinated_spans ## ]]` (must be formatted as a
valid Python list[str], and the list of spans must follow original order that
they appear in the summary), and then ending with the marker for `[[ ##
completed ## ]]`.
```

Demo model response 2:

```
[[ ## reasoning ## ]]
1. There were 67 people in total on board the two aircrafts, resulting in 67
fatalities. Not 76.
2. Though the engine and fuselage were recovered, and a wing is being retrieved,
 the claim that the entire wreckage will be recovered within a day is incorrect.
3. The flight data recorder indicated the jet's pitch changed a second before
impact, but the investigators have not determined whether it was an evasive
maneuver.

[[ ## hallucinated_spans ## ]]
["76", "the entire wreckage will be recovered within a day", "the jet made an
evasive maneuver"]

[[ ## completed ## ]]
```

User prompt:

```
[[ ## source ## ]]
{source}

[[ ## summary ## ]]
{summary}

Respond with the corresponding output fields, starting with the field `[[ ##
reasoning ## ]]`, then `[[ ## hallucinated_spans ## ]]` (must be formatted as a
valid Python list[str], and the list of spans must follow original order that
they appear in the summary), and then ending with the marker for `[[ ##
completed ## ]]`.
```

B More results

Here are more results on CNNDM and XSum datasets all showing similar trends.

Model	GE	AS	BS
Llama3.2-1b	4.139	0.851	-1.857
SmolLM2-1.7b	4.292	0.866	-1.878
OLMo2-7b	4.641	0.892	-1.917
Llama3.1-8b	4.733	0.900	-1.906

Table 6: SFT baseline results of four models on CNNDM.

Model	GE	AS	BS
Llama3.2-1b	3.875	0.855	-1.772
SmolLM2-1.7b	3.993	0.864	-1.778
OLMo2-7b	4.379	0.893	-1.849
Llama3.1-8b	4.585	0.906	-1.836

Table 7: SFT baseline results of four models on XSum.

Model	Gradient Ascent			Unlikelihood			Task Vector		
	GE	AS	BS	GE	AS	BS	GE	AS	BS
Llama3.2-1b	4.171↑	0.860↑	-1.869	**4.306**↑	**0.885**↑	-1.817↑	4.084	0.847	-1.923
SmolLM2-1.7b	4.309↑	0.869↑	-1.872	**4.455**↑	**0.894**↑	-1.874↑	4.443↑	0.880↑	-1.934
OLMo2-7b	4.693↑	0.898↑	-1.906↑	**4.767**↑	0.918↑	-1.852↑	4.709↑	**0.927**↑	-1.971
Llama3.1-8b	4.737↑	0.901↑	-1.907	**4.796**↑	0.915↑	-1.875↑	4.785↑	**0.924**↑	-1.930

Table 8: Results on CNNDM for three studied methods on four models according to G-Eval (GE), AlignScore (AS), and BARTScore (BS). Gradient ascent $\epsilon = 0.01$, unlikelihood $\epsilon = 0.5$, and task vector $\epsilon = 0.3$. The best result (according to G-Eval and AlignScore) for each model is **in-bold**. Scores higher than the baseline are marked with ↑.

Model	Gradient Ascent			Unlikelihood			Task Vector		
	GE	AS	BS	GE	AS	BS	GE	AS	BS
Llama3.2-1b	3.838	0.858↑	-1.767↑	**4.100**↑	**0.895**↑	-1.761↑	3.741	0.841	-1.824↑
SmolLM2-1.7b	4.031↑	0.868↑	-1.780	**4.237**↑	**0.898**↑	-1.803	4.089↑	0.885↑	-1.798
OLMo2-7b	4.424↑	0.906↑	-1.817↑	**4.601**↑	**0.921**↑	-1.794↑	4.545↑	0.909↑	-1.950
Llama3.1-8b	4.632↑	0.909↑	-1.836	**4.702**↑	0.919↑	-1.808↑	4.686↑	**0.935**↑	-1.875

Table 9: Results on XSum for three studied methods on four models according to G-Eval (GE), AlignScore (AS), and BARTScore (BS). Gradient ascent $\epsilon = 0.01$, unlikelihood $\epsilon = 0.5$, and task vector $\epsilon = 0.3$. The best result (according to G-Eval and AlignScore) for each model is **in-bold**. Scores higher than the baseline are marked with ↑.