Improved Beam Search for Hallucination Mitigation in Abstractive Summarization

Anonymous ACL submission

Abstract

Advancement in large pretrained language models has significantly improved their performance for conditional language generation tasks including summarization albeit with hallucinations. With the rise in the commercial use of text-generative applications, it has become necessary to have a component that ensures the factuality of the responses. To reduce hallucinations, conventional methods proposed improving beam search or using a fact checker as a postprocessing step. In this paper, we investigate using the Natural Language Inference (NLI) entailment metric to detect and prevent hallucinations in summary generation. We propose an inference time and easily generalizable NLI-assisted beam re-ranking mechanism by computing entailment probability scores between the input context and summarization model-generated beams during saliency-enhanced greedy decoding. We also investigate the limitations of existing academic factuality benchmarks and demonstrate that our proposed algorithm consistently outperforms the baselines in human evaluation on publicly available XSum and CNN/DM datasets.

1 Introduction

Pretrained sequence-to-sequence transformer-based models like BART (Lewis et al., 2019), Pegasus (Zhang et al., 2020) etc. have shown substantial improvements in the performance of NLP tasks like summarization, story generation, abstractive question answering, etc. Hallucination is a common issue observed during the generation process (Ji et al., 2022) as the pretraining is largely conducted on unlabeled data (Lewis et al., 2019). During this pretraining phase, the model learns the inaccuracies of training data along with its grammar and often generates words that are not pertinent to the given input during inference time.

Research has been conducted at curbing hallucination during the decoding phase. (King et al., 2022) proposed a modification to beam search by constraining the decoding step to focus on input-supported tokens. They hypothesize that the inaccuracies in gold summaries give rise to inconsistencies in the generated text. (Xu et al., 2020) and (van der Poel et al., 2022) investigated the relationship between hallucination and predictive uncertainty and proposed an extension to beam search by preferring low predictive uncertainty.

While there has been some success in constraining beam search using heuristics functions, they require manual inspection using intricate knowledge of the dataset, task and model to initialize their hyperparameters. PINOCCHIO (King et al., 2022) uses cosine distance to measure the consistency of generated word with context at each decoding step. As the dataset becomes more abstractive, relying solely on cosine distance and simple word-level heuristics is ineffective in steering the beam decoding factually. (Balachandran et al., 2022) proposed a Bart-based fact correction model that needs to be fine-tuned specifically for each dataset and adds to the overall model complexity. (Lango and Dusek, 2023) proposed a text classifier guided decoding to mitigate hallucinations in data-to-text application. Unlike our proposed decoding, their text classifier requires dataset specific supervised classifier data generation and training.

Our proposed approach overcomes the limitations of heuristics, extra parameter complexity without further training. To our knowledge, we are the first to introduce the semantically matching NLP task of Natural Language Inference (NLI) during decoding phase to re-rank the top few predictions of the model.

In this work, we compute NLI entailment scores at beam decoding step to provide the model an opportunity to change the beam track towards a less hallucinated region. Each intermediate beam is generated using greedy rollout decoding while attending to salient context parts. Then, the beams

are ranked using the NLI metric. We make the following contributions:
1) We show whether NLI can be detect hallucinations in abstractive summaries.
2) We develop a hallucination mitigation component for beam search that can modify the cumulative beam probability at the token level using the NLI metric. We also propose two versions of saliency mechanisms for effective encoding of context(i.e. article) at greedy rollout step.
3) We benchmark our performance against Pinocchio and Factedit on existing state-of-the-art factuality metrics including FactCC, SummaC and QGQA scores. We also provide demonstration examples in **??** to strengthen our case.
4) Using human evaluation, we demonstrate the effectiveness of our proposed approach against the baselines.

2 Related Work

2.1 Measuring and improving faithfulness

Faithfulness refers to how consistent the generated text is with respect to the input. Throughout this paper, we use the term factually inconsistent to be synonymous with hallucinated text. (Maynez et al., 2020) assessed the types of hallucinations produced by different abstractive summarizers (RNN-based Seq2Seq (See et al., 2017), GPT-tuned (Radford et al., 2019), BertS2S (Devlin et al., 2018)).

To measure factual inconsistency, (Kryscinski et al., 2020) trained FactCC – a Bart-Base model finetuned on synthetically hallucinated summaries using semantically variant and invariant transformations like Entity Swap, Sentence Negation, Paraphrasing and Noise Injection. But such a black box model lacks interpretation, has low generalizability to other datasets and is only adept at finding minor hallucinations like the transformations. (Durmus et al., 2020; Fabbri et al., 2022) have explored QA based metrics to measure factual inconsistencies by generating and comparing question-answer pairs across generated and ground truth summaries.

To generate faithful summaries, numerous architecture modifications have been proposed (Lyu et al., 2022; Zhang et al., 2022). Such modifications are neither generalizable nor feasible to be incorporated into industry text generation models with various constraints and train their existing summarizer models from scratch. Research has also been into improving the loss function component to improve overall factual accuracy. For example, (Kang and Hashimoto, 2020) demonstrated that the model shows increased factual accuracies by truncating the loss by adaptively removing high log loss examples.

2.2 Hallucinations in abstractive summarization

(Ji et al., 2022) documented an extensive survey on hallucinations present in various NLP downstream tasks. They discuss the different hallucination mitigation methods with respect to these tasks and summarize the metrics to measure hallucination. An abstract summary is defined to be hallucinated if it has any spans of text not semantically supported by the input document. Hallucinations can be categorized into two major types – intrinsic and extrinsic. Intrinsic hallucinations refer to the contradictions in the abstract summary with respect to the input document: for example, using wrong pronouns, swapping names and verbs. Models like FactCC (Kryscinski et al., 2020) trained on minor text transformations can be used to detect such errors. Extrinsic hallucinations refer to the unsupported spans of text present in the generated summaries that cannot be verified only with the input document. It arises partly due to the extrinsic hallucinations present in human written summaries which the model overfits during training process. (Qiu et al., 2023) studied hallucination in multilingual setting and proposed a multilingual factuality metric by weighting English faithfulness metrics. Previous studies used NLI-trained models to rerank summaries like (Falke et al., 2019; Mitchell et al., 2022) and compared the complete summary candidates with context. They empirically show that NLI entailment probability alone isn't enough to differentiate the correct summary from incorrect ones. (Dziri et al., 2022) studied hallucinations in conversational models using their corresponding datasets and observed that Sequence-2-Sequence models like GPT-2 not only hallucinate but also amplify the percentage of hallucinations with respect to the training data. (Cao et al., 2022; Dong et al., 2022) inspect whether the hallucinations generated in summary align with world knowledge.

3 Using NLI to detect hallucinations

(MacCartney and Manning, 2008) defines Natural Language Inference as the task of determining whether a natural-language hypothesis can be inferred from a given premise. Given a premise and a

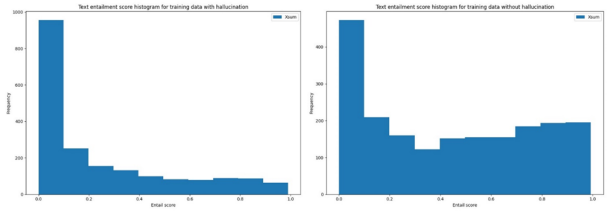


Figure 1: Histogram of entailment scores for the XSum training data (a) with and (b) without hallucinations

hypothesis, NLI computes the relationship between them in the form of three probabilities – entailment, contradiction and neutral. For Algorithm 1, we mainly consider the entailment score. In this section, we use the recognizing textual entailment task as defined in (MacCartney and Manning, 2008) to detect hallucinations in abstractive summarization task. Intrinsic hallucinations are harder to detect as they require more than lexical matching to deduce the relevance of a given word with context. To quantify the hallucinations, we consider only entity-based hallucinations for the purpose of analysis.

Dataset	Hallucinated	Not Hallucinated
XSum	0.243	**0.433**

Table 1: Average entailment scores of the XSum training data on 2000 samples

NLI models are fine-tuned to detect the relationship between single sentence premise and hypothesis. In order to verify whether their classification can extend to multisentence premise and single sentence hypothesis setting, we conducted an experiment to analyze the correlation between entailment scores and entity hallucinations on randomly selected 2000 training samples in the XSum dataset. It is crucial to note that Xsum is a single sentence summary dataset and . Hence, we effectively compare a single sentence against an array of sentences in the document. From Figure 1, it is evident that although there is a high frequency of low entailment scores for both data with and without hallucinations, the distinction between them becomes clearer at higher entailment scores. Indeed a higher entailment score correlates with low probability for entity hallucinations. This is also reflected in the average entailment scores as shown in Table 1. This analysis proves that using a multi-sentence premise, the NLI measure can detect entity-based hallucination in a single-sentence hypothesis. We use this argument and setting to introduce NLI during the beam decoding.

4 Methodology

In this section, we describe the components of our NLI-aided beam search re-ranker as shown in Figure 2. For all experiments, we use a BART-Base model (Lewis et al., 2019) finetuned with the given dataset. Architectures like BART (Lewis et al., 2019) have an autoregressive decoder that generates the output word by word conditioned on the input text and the words generated so far as shown in Equation 1. As mentioned in (Meister et al., 2020), Beam search performs a breadth-first search with limited branches with the beam size starting with the BOS token(Begin of sentence) and ending the search at EOS token(End of sentence). Each path from BOS to EOS is called a hypothesis.

We define intermediate beam or partial hypothesis as the sequence of sub paths of hypotheses starting at BOS and ending before EOS. The saliency enhanced greedy rollout component, explained in Section 4.1, attends over important parts of the context relevant to the intermediate beams and completes the beam till EOS.

The intermediate beams are sent to the saliency-enhanced greedy rollout to serve as a look-ahead mechanism to complete the beams. The completed candidate beams are scored using the entailment probabilities from the NLI model. Then the intermediate beams are re-ranked based on the weighted probabilities between entailment and the model probabilities. Detailed steps of our proposed algorithm are provided in Algorithm 1. We adopt required variable names from (King et al., 2022) for consistency.

$$P_\Theta(y|x) = \prod_{t=1}^{|y|} P_\Theta(y_t|x, y_{<t}), \qquad (1)$$

4.1 Saliency-enhanced Greedy rollout

Since it is difficult to perform the NLI task on partial hypotheses as the NLI models have been trained with complete sentences (MacCartney and Manning, 2008), we complete 2B intermediate beams as our first step where B is the beam size. Inspired by (Hargreaves et al., 2021), we use the greedy search on the intermediate beams to generate the remaining words and complete the partial hypotheses. In Algorithm 1, the saliency-enhanced greedy rollout (SGR) function takes the concatenated input of context, the intermediate beam and the next word sep-

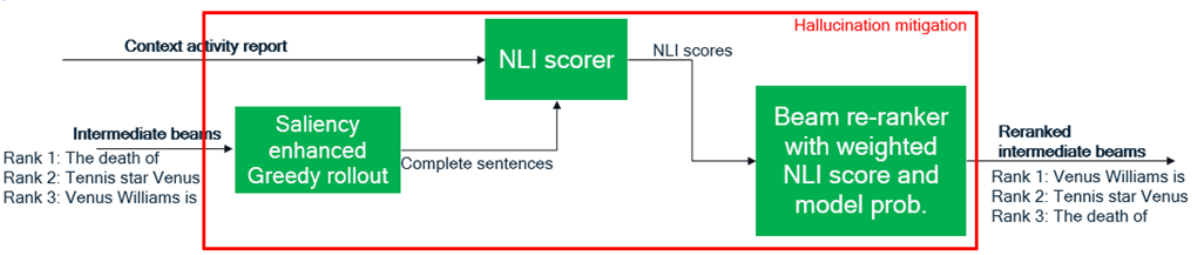


Figure 2: Proposed beam search re-ranker

arated by [SEP] token and generates the completed beams. During the greedy search, we empirically witnessed similar words being used to complete the beams regardless of the words in intermediate beams. This might be due to the long context and shorter attention span of pretrained transformers as mentioned in (Liu and Lapata, 2019). Thus the model might not effectively attend to the parts of context relevant to the words in the intermediate beam. To solve this problem, we take two steps.

First, inspired by (Cao and Wang, 2021), we enhance the effectiveness and diversity of greedy search, by introducing saliency on the context relative to the intermediate beam using attention head masking. We compute the saliency score for every word in context by averaging its cosine distance with each word in the intermediate beam. We propose two saliency versions v1 and v2 suitable for summaries with extractive and abstractive characteristics respectively. In Saliency v1, using a threshold as a hyperparameter, we compute mask matrix m (Equation 2) to selectively attend to words in the context relevant to the completion of the current intermediate beam. A hard masking is done so that it increases the probability of copying relevant words from the input. But in a highly abstractive summarization setting, attending over all words is crucial. Hence, we propose Saliency v2 that performs variable soft attention over the words in the input. In Saliency v2, inspired by GATE (Ahmad et al., 2021), we use the computed saliency scores to perform weighted attention (Equation 3) on the softmax output and normalize the scores(Equation 4).

Second, we perform the proposed re-ranking only if the hypothesis has a predefined minimum of words so that the beam doesn't converge to the same space during greedy search. This is because if the hypothesis has few words, the beam might not have the necessary entities suitable for measuring hallucination. In the future, we can identify the appropriate time steps suitable for re-ranking the hypothesis to avoid hallucinations.

$$\text{Attention(q,K,V)} = \text{softmax}(\frac{qK^T}{\sqrt{d_k}} + \text{m})\text{V} \quad (2)$$

Where q is query, K and V represent key and value matrices respectively, d_k is the scaling factor and m is the attention mask matrix.

$$\text{Attention(q,K,V)} = \text{F}(\text{softmax}(\frac{qK^T}{\sqrt{d_k}} + \text{m})\text{V}) \quad (3)$$

$$\text{F(P)}_{ij} = \frac{\text{P}_{ij}}{\text{Z}_i \text{D}_{ij}} \quad (4)$$

Where $Z_i = \sum_j \frac{P_{ij}}{D_{ij}}$ is the normalization factor and D_{ij} is the saliency score between ith token in intermediate beam and jth token in document.

4.2 Natural Language Inference(NLI) scorer

As a next step, we pass the greedy rollout beams to the NLI scorer. We obtain the entailment probability with the context as premise and the beam as hypothesis (MacCartney and Manning, 2008) as illustrated with Equation 4. The NLI function takes in Context C as the premise and rolled out beam R as the hypothesis and computes their relationship as entailment score. We hypothesize that the entailment probability is inversely proportional to the hallucination content of the beam.

4.3 Weighted Beam re-ranker

In order to incorporate the NLI score into the overall cumulative beam probability, we take a weighted average of entailment and model probabilities for each decoding step and add them to the cumulative beam probability. We then re-rank the beams based on the modified cumulative probability and select the top B candidates as re-ranked intermediate beams. The weights need to be normalized as we are adding two random variables. As mentioned in Equation 4, we consider weight(α)

Algorithm 1 NLI Assisted Beam Decoding

Input variables: Beam size B, Generative Model M, Vocab V, wait threshold δ
Input functions: Saliency enhanced greedy rollout function **SGR** returns completed beams,
Natural Language Inference function **NLI** returns entailment probability;
Initialize: I = {(Inter$_i$, Cumulative P$_i$) : Inter$_i$ $\in$ set of Intermediate beams} ;
Context C = $\{x_1, x_2, x_3, \ldots . x_n : x_i \in \text{V}\}$; Candidate Intermediate beams CI = {};
Current Completed beams CC = {}; Completed Generations CG = {};
Output: top-ranked elements of CG
while |CG| < B **do**
  for (Inter$_i$, Cumulative P$_i$) in I **do**
    T := { (t$_i$, P$_i$) : t$_i$ $\in$ Top 2B tokens of V predicted by Model M with P$_i$ probability}
    if |Inter$_i$| > δ **then**
      for (t$_i$, P$_i$) in T **do**
        R:= **SGR**(M, C, Inter$_i$, t$_i$)
        P$_{entail}$:= **NLI**(C, R)
        P$_{weighted}$:= (1-α)P$_i$ + α P$_{entail}$
        P$_i$:= P$_{weighted}$
      end for
    end if
    CI := {(Inter$_i$ + t$_i$, Cumulative P$_i$ + P$_i$): (t$_i$, P$_i$) $\in$ T}
    I := I U Top B beams from CI ranked by Cumulative P
    CC:= {Inter$_i$: for all beams Inter$_i$ $\in$ I ending with '<end>' token}
    I:= I - CC
    CG:= CG U CC
  end for
end while
return top ranked elements of CG

as a hyperparameter which can be increased up to 1.0 depending on the necessity of faithfulness in the generated text for a given task.

$$\begin{aligned} \text{P}_{entail} &:= \text{NLI(C, R)} \\ \text{P}_{weighted} &:= (1-\alpha)\text{P}_i + \alpha \text{P}_{entail} \end{aligned} \tag{4}$$

Where NLI is the Natural Language Inference function taking Context C, and Intermediate beam Rollout R as inputs.

5 Experiments

5.1 Dataset

We use two datasets, namely, CNN/DM (Hermann et al., 2015) and XSum (Narayan et al., 2018), to evaluate our model performance. CNN/DM corpus is generated from human-written multi-line summaries for the CNN and Daily Mail news articles. It is under apache-2.0 license. It consists of over 285k training pairs, 13,368 validation pairs and 11,487 test pairs. The XSum dataset is made up of BBC articles and their one-line summaries. It is under MIT license. It comprises over 90k training samples and is more abstractive than CNN/DM as it contains 18.6% more novel unigrams. We aim to develop an approach that works consistently on both abstractive and extractive types of summaries.

5.2 Implementation Details

We used the PyTorch (Paszke et al., 2019) implementation of the BART base version (=140M parameters) from hugging face library (Wolf et al., 2020). For the decoding process, we use beam search with beam size 5 and a maximum length of 125 tokens after Byte Pair Encoding(BPE) tokenization. For NLI, we use BART-Large model finetuned on MNLI dataset (Williams et al., 2018). Unless explicitly specified, we use α = 0.8 for both the datasets. Due to resource constraints, we perform the evaluation on 820 randomly selected test samples from each of the datasets. For additional details, please refer to Appendix A.1.

5.3 Evaluation Metric

We use multiple factuality metrics including SummaC-Conv, SummaC-CZS (Laban et al., 2022), QGQA (Fischer et al., 2022) and FactCC (Kryscinski et al., 2020) to showcase the relative performance of baselines and our approach. FactCC is a binary classifier for factuality. Since we are comparing models against reference-less metrics, we consider the probability of not hallucinated as FactCC score. For QGQA, we use its F1 score. We penalize the cases when the algorithm doesn't generate a summary with corresponding low score. We provide a comprehensive human evaluation study consisting of 50 in-

stances randomly drawn from the XSum and CNNDM datasets equally. 4 human annotators in U.S. with more than 28 years of combined professional and academic English proficiency scored between 1(lowest) to 5(highest) to evaluate the faithfulness of summaries. For additional details, please refer to Appendix A.2.

6 Results

We demonstrate that our proposed beam search modification reduces hallucination during inference time by comparing it with multiple baselines including vanilla beam search, PINNOCHIO (King et al., 2022) and FactEdit (Balachandran et al., 2022). From Table 2, we can observe contrasting trends for both the datasets. For Xsum, our approach performs better than the baselines except for QGQA. While for CNN/DM, the metrics tend to favour FactEdit. We hypothesize that this behavior is due to vanilla beam search and FactEdit algorithms tending to generate slightly more extractive summaries in comparison to the other methods. Similar to the conclusion in a contemporary paper (Tang et al., 2023), we also observe that none of the factuality metrics are consistent across datasets for the Bart-based summarization model. As hypothesized in Section 4.1, with respective to contextualized reference-less factuality metrics, i.e, SummaC-Conv and SummaC-ZS scores, Saliency v1 version of our proposed decoding performs better for CNN/DM(extractive in nature) as it performs hard masking which increases chances of copying relevant words from input. While Saliency v2 versoin is preferred for Xsum(highly abstractive). In Table 3, We perform human evaluation to rate the summaries subjectively. The high faithfulness score of the proposed approach compared to other baselines bolsters the efficacy of our algorithm.

7 Analysis

In Table 7, we investigate the role of hyper parameter α in guiding the beams to factual generation by varying its values across the spectrum. For α = 0.0, all the generated beams are factually wrong. While for α = 0.2 and 0.8 the generated beams are factually consistent. The beams with non-zero α are inherently more diverse than the vanilla beam search generations.

Table 8 demonstrates how our proposed algorithm avoids an entity relation error made by vanilla beam search. In the article, "Theresa May" is the prime minister but vanilla beam search considers both as different entity. Baseline FactEdit fails to correct the hallucination while Pinnochio does not generate any summary. Our proposed algorithm guides the beam decoding into factually right area and correctly addresses "Theresa May" as Prime Minister. More examples are provided in the Appendix **??** section.

To quantify whether the beams were able to explore different regions of text space, we propose a diversity metric (Equation 5) to measure the average frequency of novel words across the beams. In future, we plan to evolve the set intersection operation to incorporate a semantic representation of words.

$$\text{Diversity} = \frac{\sum_{i=0}^{n-2} b_i \bigcup b_{i+1} - b_i \bigcap b_{i+1}}{{}^{n}_{2}\text{C}} \quad (5)$$

Where n is the beam size, b_i is the set of unique words in beam i.

Next, we investigate whether our algorithm is able to guide the beam search towards faithful regions even if we increase the re-ranking interval i.e. performing the NLI-based re-ranking once for x number of tokens. We compare it against Pinochio due to both of them performing computations at decoding step. In Table 4, we can see that Pinnochio performs worse for Xsum while does well comparatively for CNN/DM compared to our method. Unlike our approach, Pinnochio took longer time for Xsum because it backtracks several times and eventually times out if it is unable to generate a faithful summary. We can see that the factuality metrics SummaC-Conv and SummaC-ZS pretty much stay consistent for reranking intervals from 8 to 64. It is clear that the beams do not fall off the right track even if we perform re-ranking at slightly longer intervals. This result is particularly important as we can tweak and increase re-ranking interval depending on the dataset characteristics to achieve lower inference latency. Rouge being a n-gram matching metric also stays similar for Xsum while produces lightly higher scores as we increase reranking interval for CNN/DM dataset. Due to CNN/DM being inclined towards extractive summarization, the language model might require lower guidance by our NLI assisted reranker to follow the correct summary path. In Table 5, we evaluate the diversity score for vanilla beam search and the proposed approach variations. Since, XSum has more chances of hallucination due to its abstractive nature, a higher diversity score would enable

Dataset	Decoding Algorithm	SummaC-Conv	SummaC-ZS	FactCC	QGQA
XSum	Vanilla beam search	0.244	-0.358	0.248	**0.843**
	FactEdit	0.245	-0.356	0.249	0.837
	PINNOCHIO	0.213	-0.363	0.205	0.786
	Ours (Saliency v1)	**0.248**	-0.289	**0.280**	0.841
	Ours (Saliency v2	0.248	**-0.279**	0.219	0.839
CNN/DM	Vanilla beam search	0.683	0.137	0.352	**0.948**
	FactEdit	0.536	**0.142**	**0.412**	0.945
	PINNOCHIO	**0.676**	0.141	0.344	0.942
	Ours (Saliency v1)	0.604	0.050	0.212	0.930
	Ours (Saliency v2	0.439	0.032	0.219	0.932

Table 2: Performance of baselines and our proposed approach on XSum and CNN/DM datasets.

Decoding Algorithm	Faithfulness
Vanilla beam search	3.37
Pinnochio	3.21
FactEdit	3.32
Ours (Saliency v2)	**3.48***

Table 3: Human Evaluation on faithfulness metric for baselines and proposed approaches. * denotes statistically significant over Pinnochio and FactEdit with 95% confidence interval on paired-t test.

Rerank Int.	S-Conv	S-ZS	R-1	R-2	R-L	Avg. Time
Pinocchio	0.213	-0.363	0.099	0.030	0.068	47.47
8	0.241	-0.354	0.109	0.036	0.075	8.404
16	0.240	-0.351	0.110	0.036	0.075	8.585
32	0.239	-0.349	0.109	0.035	0.073	8.584
64	0.240	-0.353	0.109	0.035	0.074	8.470

(a) Xsum

Rerank Int.	S-Conv	S-ZS	R-1	R-2	R-L	Avg. Time
Pinocchio	0.676	0.141	0.165	0.144	0.149	63.70
8	0.605	0.041	0.199	0.158	0.164	67.385
16	0.601	0.040	0.200	0.158	0.165	66.76
32	0.601	0.038	0.201	0.159	0.165	66.872
64	0.605	0.039	0.202	0.161	0.167	66.125

(b) CNN/DM

Table 4: Effect of re-ranking interval on overall performance for Xsum and CNN/DM datasets. S-Conv and S-ZS stands for SummaC-Conv and SummaC-ZS. R-1, R-2 and R-L stands for Rouge-1,2 and L. Avg. Time, in seconds, is the time taken on an average for one summary generation. Rerank Int. refers to Reranking Interval

the algorithm to explore more regions for reducing hallucination. Whereas, CNN/DM being highly extractive, doesn't require much re-ranking during the decoding and hence the low diversity.

Next, we inspect the role of the decoding algorithm during the rollout of intermediate beams on the overall performance of the algorithm. From Table 6, We can see that Top K and Greedy search

Decoding Algorithm	XSum	CNN/DM
Vanilla beam search	2.27	1.75
Ours (Saliency v1)	2.53	1.04
Ours (Saliency v2)	2.51	1.01

Table 5: Diversity scores of vanilla beam search and proposed approach.

Rollout decoding vs Dataset	XSum		CNN/DM	
	S-Conv	S-ZS	S-Conv	S-ZS
Random Sampling	0.247	-0.299	0.599	0.042
Top K	0.246	**-0.279**	**0.605**	**0.049**
Top P	0.247	-0.291	0.601	0.047
Greedy	**0.248**	-0.289	**0.605**	**0.049**

Table 6: Analysis of different decoding strategies for rollout component. S-Conv and S-ZS stands for SummaC-Conv and SummaC-ZS.

perform superior to their counterparts and can be concluded as better lookaheads for faithful decoding.

8 Conclusion and Future Work

We propose a modification to the beam search decoding algorithm that guides beam generation to avoid falling into hallucination regions by re-ranking the beams based on NLI entailment scores computed on saliency-enhanced greedily rolled-out partial hypotheses. We present the issue of inconsistency of SOTA Summarization factuality metrics to motivate the development of a robust benchmark for detecting hallucinations. By human evaluation, we show that our NLI-based re-ranker consistently improves the faithfulness score. In the future, we intend to investigate NLI-based re-ranker's performance on other NLP downstream tasks such as story generation with prompt, question answering and query-focused summarization. Also, we plan to study more efficient methods to incorporate the NLI as a guidance mechanism for decoding algorithms. We also intend to study the efficacy of our

Gold Summary	US tennis star Venus Williams has been involved in a car accident that led to the death of a 78-year-old man.
α	**Generated Summary**
0.0	Tennis star Venus Williams has died in a car crash in Florida, police say. Tennis star Venus Williams has died in a car crash in Florida, US police say. Tennis star Venus Williams has died in a car crash in Florida, according to reports. Tennis star Venus Williams has died in a car crash in Florida, according to police. Tennis star Venus Williams has died in a car crash in Florida.
0.2	Tennis star Venus Williams was at fault for a car crash that killed a man in Florida, police say. Tennis star Venus Williams was at fault for a crash that killed a man in Florida, police say. Tennis star Venus Williams is being investigated over the death of a man in Florida, police say. Tennis star Venus Williams was at fault for a car crash that killed a man, police say. Tennis star Venus Williams is being investigated over the death of a man in a car crash in Florida.
0.8	Tennis star Venus Williams is being investigated over the death of a man in Florida, police say. Tennis star Venus Williams was at fault for a car crash that killed a man in Florida, according to reports. Tennis star Venus Williams was at fault for a car crash that killed a man in Florida, police say. Tennis star Venus Williams was at fault for a car crash that killed a man in Florida, according to police. Tennis star Venus Williams is being investigated for causing the death of a man by careless driving, police say.

Table 7: An example, from XSum dataset, illustrating the effect of hyperparameter α for beam size 5. Highlights in Red indicate factually inconsistent beams.

Article	Although there is some common ground between the two governments on, for example, the need for free trade within the single market, Carwyn Jones has complained that he didn't see the letter before it was published on Wednesday. (He has that in common with most of Mrs May's cabinet). The first minister told AMs: "I discussed the Article 50 letter in general terms with the prime minister when we met in Swansea last week. "I should be clear, though, that I didn't see the letter before today and we were not invited to contribute to its drafting. This is unacceptable and is the culmination of a deeply frustrating process in which the devolved administrations have persistently been treated with a lack of respect. "It is all the more regrettable given the UK government's stated aim was to develop a negotiating framework for the whole of the UK. "Mr Jones may have been playing to an audience, but Welsh Secretary Alun Cairns hit back: "I'm a bit disappointed in that. The prime minister has been in Wales three times in the last six weeks. "We've been talking about the contents of this letter for many months. "We've clearly all made our representations but, ultimately, the UK government needs to act in the interests of the whole of the UK and that's what we're doing, specifically with Wales being mentioned. " Mrs May did indeed mention Wales in the letter. She told Donald Tusk: "When it comes to the return of powers back to the United Kingdom, we will consult fully on which powers should reside in Westminster and which should be devolved to Scotland, Wales and Northern Ireland. "But it is the expectation of the government that the outcome of this process will be a significant increase in the decision-making power of each devolved administration. " That sentence may have been written more with Scotland in mind, but it does prompt the question: which powers? Farming? Economic aid? And will the money follow the powers? Alun Cairns wouldn't answer those questions, although Carwyn Jones has said he fears there won't be any money to accompany the powers after 2020. (Perhaps Mr Jones doesn't think Jeremy Corbyn will win power that year - Labour has pledged to maintain EU funding levels on regional aid beyond 2020). Some in Whitehall think the way EU money has been spent in Wales - check out Nick Clegg's film from Ebbw Vale - is an argument for transferring those powers to Westminster, but that looks politically less likely now. We may get some more details about the process in a white paper on the Great Repeal Bill on Thursday but, at the moment, the identity of the powers the UK government wants to see devolved is something of a mystery.
Gold Summary	Theresa May's letter triggering Article 50 may have attempted a more conciliatory tone but it does not seem to have worked with the Welsh Government.
Vanilla beam search	Theresa May's letter to the prime minister is "unacceptable" and "disappointing", according to the first minister.
Pinocchio	No summary
FactEdit	Theresa May's letter to the prime minister is "unacceptable" and "disappointing", according to the first minister.
Ours (Saliency V1)	Theresa May's letter to the Welsh secretary outlining her plans to leave the EU has been criticized by the Welsh Secretary.
Ours (Saliency V2)	Prime Minister Theresa May has been criticized by the first minister after the publication of a letter from the prime minister outlining her plans for devolution.

Table 8: An example, from XSum dataset, showing how the proposed method avoids hallucinations

approach on tasks with longer context like long document summarization using a retrieval augmented ranker.

Limitations

The re-ranking of intermediate beams at regular intervals might introduce some delay in the decoding process. We show in Table 4, that for smaller variations in re-ranking interval, the beams still follow the same faithfulness guided path. Thus, we believe that such latency could be offset by setting the appropriate interval size and parallel processing of the beams during inference time decoding.

Ethics Statement

We believe our work has no negative ethical impact on society and strongly hope our NLI-based reranker help in creating a positive impact by promoting faithfulness among text generative models conditioned on input text specifically commercial production running systems. Our work will aid in reducing false and biased information generated by LLMs. Our NLI and BART models are trained on open source datasets, CNNDM and XSum, which are said to contain hallucinations and biases from the news media. Whenever possible, it is recommended to retrain an unbiased NLI model.

References

Wasi Uddin Ahmad, Nanyun Peng, and Kai-Wei Chang. 2021. Gate: graph attention transformer encoder for cross-lingual relation and event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12462–12470.

Vidhisha Balachandran, Hannaneh Hajishirzi, William Cohen, and Yulia Tsvetkov. 2022. Correcting diverse factual errors in abstractive summarization via post-editing and language model infilling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9818–9830, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Shuyang Cao and Lu Wang. 2021. Attention head masking for inference time content selection in abstractive summarization. In *NAACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yue Dong, John Wieting, and Pat Verga. 2022. Faithful to the document or to the world? mitigating hallucinations via entity-linked knowledge in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1067–1082, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

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

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Tim Fischer, Steffen Remus, and Chris Biemann. 2022. Measuring faithfulness of abstractive summaries. In *Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022)*, pages 63–73.

James Hargreaves, Andreas Vlachos, and Guy Emerson. 2021. Incremental beam manipulation for natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2563–2574, Online. Association for Computational Linguistics.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. 2022. Survey of Hallucination in Natural Language Generation. *arXiv e-prints*, page arXiv:2202.03629.

Daniel Kang and Tatsunori B. Hashimoto. 2020. Improved natural language generation via loss truncation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 718–731, Online. Association for Computational Linguistics.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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

Mateusz Lango and Ondrej Dusek. 2023. Critic-driven decoding for mitigating hallucinations in data-to-text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2853–2862, Singapore. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yuanjie Lyu, Chen Zhu, Tong Xu, Zikai Yin, and Enhong Chen. 2022. Faithful abstractive summarization via fact-aware consistency-constrained transformer. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1410–1419, New York, NY, USA. Association for Computing Machinery.

Bill MacCartney and Christopher D. Manning. 2008. Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528, Manchester, UK. Coling 2008 Organizing Committee.

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.

Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics.

Eric Mitchell, Joseph J. Noh, Siyan Li, William S. Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher D. Manning. 2022. Enhancing self-consistency and performance of pre-trained language models through natural language inference.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Yifu Qiu, Yftah Ziser, Anna Korhonen, Edoardo Ponti, and Shay Cohen. 2023. Detecting and mitigating hallucinations in multilingual summarisation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8914–8932, Singapore. Association for Computational Linguistics.

A. Radford, Jeffrey Wu, R. Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083. Association for Computational Linguistics.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers), pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281, Online. Association for Computational Linguistics.

Haopeng Zhang, Semih Yavuz, Wojciech Kryscinski, Kazuma Hashimoto, and Yingbo Zhou. 2022. Improving the faithfulness of abstractive summarization via entity coverage control. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 528–535, Seattle, United States. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

A Additional Implementation Details

For all experiments we used 1, 16 GB NVIDIA T4 GPU. Experiments were conducted in a private cloud infrastructure.

A.1 Evaluation Details

We use the bart base checkpoint[1] provided by VictorSanh in huggingface. Bart base has 140M parameters. For CNN/DM dataset related experiments, we use distilbart base checkpoint[2] provided by sshleifer in huggingface. Distilbart base contains 40% less parameters than bart-base. For both cases, max length used is 128 with beam size 5. Repetition penalty is 1.0 for XSUM and 3.0 for CNN/DM. During Greedy rollout, we use the setting maximum new tokens as 50 with early stopping true and beam size as 1. Whenever used, top k and top p values are set to 50 and 0.92 for uniformity across ablation studies. We use 0.15 as the threshold for Saliency v1 of greedy rollout. For the baselines, PINNOCHIO[3] and FactEdit[4] we get the model checkpoints from their official github repositories. For evaluation metrics, we use the official github implementations for SummaC-Conv[5], SummaC-ZS[6], FactCC[7] and QGQA[8].For compute ROUGE-1,2,3,L scores, we use the official pip package[9] with version 0.1.2.

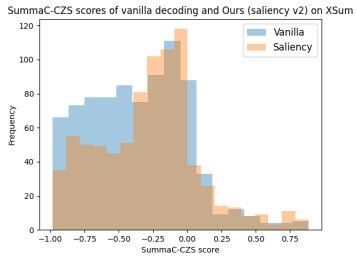
A.2 Additional Annotation Details

Each sample given to the consented annotators consists of an article and a randomly sampled summary from the summaries generated by proposed algorithms and baselines. They were informed to score 1 if the summary is highly unfaithful, i.e, critically changes the meaning of article and score 2-3, if noncritical new words appear like first names of entities like locations, people that do not present in the article but doesn't alter the meaning of the article. The annotators were recruited through advertisement inside the department and were compensated via credits. The tests were in the form of a survey and were done in a double blindfolded manner and cannot be traced back to the individuals. Therefore, not required by our IRB to be reviewed by them. The authors of this work are not lawyers. However, this opinion is based on United States federal regulation 45 CFR 46, under which this study qualifies for exemption via 46.104 Exempt research.

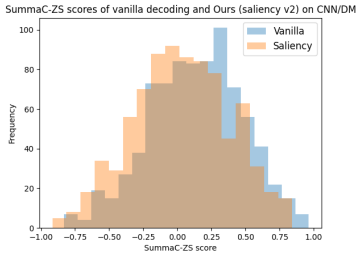
[1] https://huggingface.co/VictorSanh/bart-base-finetuned-xsum
[2] https://huggingface.co/sshleifer/distilbart-cnn-12-6
[3] https://github.com/allenai/pinocchio
[4] https://github.com/vidhishanair/FactEdit
[5] https://github.com/tingofurro/summac
[6] https://github.com/tingofurro/summac
[7] https://github.com/salesforce/factCC
[8] https://github.com/bigabig/faithfulness
[9] https://pypi.org/project/rouge-score/

A.3 Analysis

We perform a few additional ablation studies and analysis to understand our algorithm in depth. In Figure 3, we visualize the SummaC-ZS scores of vanilla beam search and Ours (saliency v2). In CNNDM dataset, the similar score distribution of vanilla beam search and our algorithm shows that a few outlier scores, especially in 0.25 to 0.50 bucket, have a high influence on the mean SummaC-CZS scores. We suspect this might be a possible explanation for the low performance of our algorithm on CNNDM dataset.



(a) XSum



(b) CNNDM

Figure 3: SummaC-ZS distribution of vanilla beam search and Ours (Saliency v2)

Next, we investigate whether our algorithm is able to guide the beam search towards faithful regions even if we increase the re-ranking interval i.e. performing the NLI-based re-ranking once for x number of tokens. From Table 4a and 4b, it is evident that the beams do not fall off the right track even if we perform re-ranking at slightly longer intervals. This result is particularly important as we can tweak and increase re-ranking interval to achieve lower inference latency.

Finally, we study if utilizing both entailment and contradiction probabilities aids in re-ranking the beams. $\mathrm{P}_{weighted}$ in Algorithm 1 will be modified using Equation 5. For this experiment, we assigned 0.6 and 0.2 to α_1 and α_2 respectively. From Table 9, we see that combining the contradiction probability of NLI doesn't yield consistently better results than the proposed approach across the datasets.

$$\begin{aligned} \mathrm{P}_{prob} &:= \alpha_1 \mathrm{P}_{entail} + (1-\alpha_1)\mathrm{P}_{contradiction} \\ \mathrm{P}_{weighted} &:= \alpha_2 \mathrm{P}_i + (1-\alpha_2)\mathrm{P}_{prob} \end{aligned} \tag{5}$$

Where P_{entail} and $\mathrm{P}_{contradiction}$ denote the entailment and contradiction probabilities from NLI and P_i refers to the model probability.

A.4 Demonstration Examples

Decoding Algorithm vs Dataset	XSum				CNN/DM			
	Diversity	**S-Conv**	**S-ZS**	**QGQA**	**Diversity**	**S-Conv**	**S-ZS**	**QGQA**
Vanilla beam search	2.27	0.244	-0.358	**0.843**	1.75	**0.683**	**0.137**	**0.948**
Ours : E	2.54	**0.248**	-0.289	0.841	1.04	0.605	0.049	0.930
Ours : E & C	2.54	0.247	**-0.270**	0.841	0.99	0.604	0.053	0.930

Table 9: Effect of **E**ntailment and **C**ontradiction NLI probabilities on overall performance

Article	Diego Simeone's side surrendered a 15-game unbeaten run as the Brazilian fired home an 88th-minute volley. A draw would have seen Atletico climb above Barcelona to the summit going into the 10-day winter break. But they had to hold on after losing skipper Gabi Fernandez to two yellows cards in the space of five minutes early in the second half. In a sometimes bad-tempered encounter, the visitors were up against it from the moment Fernandez was cautioned for a foul, then ordered in the 56th minute after committing a handball near the halfway line. Charles saw two first-half chances saved by Atletico keeper Jan Oblak in a game short on clear openings. But the Brazilian had the final say, finding the net with the aid of a deflection off defender Diego Godin for his sixth goal of the season. Barcelona have a game in hand after being without a domestic fixture due to their involvement in the Club World Cup final against River Plate.
Gold Summary	Atletico Madrid missed a chance to go top of La Liga after falling to a late winner from Malaga striker Charles.
Vanilla beam search	Cristiano Ronaldo scored the only goal of the game as Atletico Madrid came from behind to draw 1-1 with 10-man Charles.
Pinocchio	No summary
FactEdit	Cristiano Ronaldo scored the only goal of the game as Atletico Madrid came from behind to draw 1-1 with Charles in La Liga.
Ours (Saliency V1)	Atletico Madrid's La Liga play-off hopes suffered a setback as a late goal by Neymar saw them draw with Charles.
Ours (Saliency V2)	Atletico Madrid's La Liga title hopes suffered a blow as they lost 1-0 to Charles at the Nou Camp.

Article	It comes after Camden residents tried to save the St James' Gardens site, close to London Euston, which was a burial ground from 1790 until 1853.Local church warden Dorothea Hackman said it was "quite outrageous" they were going to "dig up our dead".HS2 Ltd said the work would be done with "dignity, respect and care". Notable people buried in the gardens include Capt Matthew Flinders, the first person to circumnavigate and name Australia, and Bill Richmond, one of the first black boxers."[These people] shouldn't be disturbed by spurious activities like this," said Ms Hackman, who helped organise the service, which was expected to be attended by 40 people. "And just think of the detrimental effect removing the benefit of the trees and green space will have on the area in terms of air quality. "There has not been destruction on this scale since the Sixties. Government has run roughshod over democracy. "Resident Marian Kamlish, 92, said that "in times of austerity", such a "vanity project" was an insult to those who work for the likes of the NHS, fire and police forces.HS2 Ltd will excavate sections of the burial ground to enable it to plan the removal of the remains prior to their subsequent re-interment elsewhere. A spokesman stressed the grounds had not been in use for more than a century. "We will ensure that we treat the site with dignity, respect and care," he said. "As such, we will continue to work closely with the local community, the Archbishops' Council, the local parish, Historic England and other organisations as we proceed with the next phase of the project. "In February, Parliament granted powers to build Phase 1 of the line - between London and Birmingham - which is due to open in December 2026.In June, the Department for Transport (DfT) said First Trenitalia West Coast, MTR West Coast Partnership and West Coast Partnership had all been shortlisted to operate the service on the line.
Gold Summary	A memorial service has been held for 60,000 people whose remains are due to be exhumed in London as part of the 7bn HS2 high-speed rail project.
Vanilla beam search	Hundreds of people have attended a service to mark the centenary of the birth of the first black boxer.
Pinocchio	A service has been holds at a Camden church for those buried in a former burial ground to be removed for the HS2 railway line project.
FactEdit	Hundreds of people have attended a service to mark the centenary of the birth of the first black boxer.
Ours (Saliency V1)	A memorial service has been held for people buried in a former cemetery in south-east Britain.
Ours (Saliency V2)	A memorial service has been held in the grounds of a Victorian burial ground in south-east London.

Table 10: Examples demonstrations, from XSum dataset, showing how the proposed method avoids hallucinations