SH2: Self-Highlighted Hesitation Helps You Decode More Truthfully

Anonymous ACL submission

Abstract

Large language models (LLMs) demonstrate great performance in text generation. However, LLMs are still suffering from hallucinations. In this work, we propose an inference-time method, **S**elf-**H**ighlighted **H**esitation (**SH2**), to help LLMs decode more truthfully. SH2 is based on a simple fact rooted in information theory that for an LLM, the tokens predicted with lower probabilities are prone to be more informative than others. Our analysis shows that these low-confidence tokens are more likely to be closely related to factual information, such as nouns, proper nouns, and adjectives. Therefore, we propose to "highlight" the factual information by selecting key tokens with the lowest probabilities and concatenating them to the original context, thus forcing the model to repeatedly read and hesitate on these tokens before generation. During decoding, we also adopt contrastive decoding to emphasize the difference in output probabilities brought by the hesitation. Experimental results demonstrate that our SH2, requiring no additional data or models, can effectively help LLMs elicit factual knowledge and distinguish hallucinated contexts by themselves. Significant and consistent improvements are achieved by SH2 for LLaMA-7b, LLaMA2-7b and Mistral-7b on various hallucination tasks.[1]

1 Introduction

Depending on massive training corpora, large language models (LLMs) have made tremendous progress in natural language understanding and text generation (Touvron et al., 2023a, Jiang et al., 2023, OpenAI, 2023). However, during reasoning and generation, LLMs could suffer from hallucinations and generate non-factual answers (Zhang et al., 2023).

To clear these falsehoods, some researchers construct datasets with higher quality and train LLMs to respond in the correct form (Taori et al., 2023, Zhou et al., 2023). But in the domain not included in the training, LLMs assign similar probabilities to correct and wrong choices since they do not have enough relevant knowledge to distinguish them.

To fill the gap of knowledge, retrieval augmentation methods (Peng et al., 2023; Gao et al., 2023; Gou et al., 2023) leverage external knowledge bases and tools to correct the output of LLMs. Although they provide additional information for LLMs, it still seems impossible to guarantee zero error in the external knowledge. Other researchers propose decoding reformulation methods (Li et al., 2023b; O'Brien and Lewis, 2023; Chuang et al., 2023) to address hallucinations. However, most of them rely on external data with human labor or larger models to rescale probability distribution during decoding.

Recent works (Wei et al., 2022; Goyal et al., 2023) have demonstrated that a few more computation steps for LLMs can make a difference. Wei et al. (2022) propose chain-of-thought (COT) prompting to elicit intermediate reasoning steps of LLMs. These intermediate steps can lead to better answers. However, elaborate prompts need to be prepared in advance. Goyal et al. (2023) append learnable pause tokens to the input prefix and train models from scratch. These delays introduced by pause tokens provide the model with more computation steps to generate better answers.

In this paper, we seek to leverage the LLM itself to highlight informative tokens and digest them as input during the hesitation steps to elicit truthful knowledge inside LLMs. We propose a simple yet effective method, **S**elf-**H**ighlighted **H**esitation (**SH2**), to help LLMs decode more truthfully. SH2 introduces hesitations to give LLMs more time to understand contexts and answer questions. For LLMs, the tokens assigned with lower probabilities are harder to predict, while more likely to be informative. *LLMs can select these key tokens by*

[1]We will release our code for reproducibility later.

themselves from the input and hesitate on these highlighted tokens. We calculate the difference brought by highlighted tokens through contrastive decoding (Li et al., 2023c) and integrate it into the output probability. Experiments on multiple tasks demonstrate that such a difference could elicit factual knowledge inside the model and successfully mitigate hallucinations. Unlike other methods, our method does not leverage any other external tools or data. Additionally, it can be directly deployed during inference with no more training.

2 Related Work

2.1 Hallucination Mitigation

Recent works to mitigate the hallucination of LLMs can be summed up into three categories.

Supervised Fine-Tuning Many researchers pay attention to the curation of the training data and attempt to mitigate hallucinations through additional fine-tuning. Alpaca (Taori et al., 2023) have collected 52K instruction-following data of massive tasks and fine-tuned the LLaMA-7b model (Touvron et al., 2023a). Such an instruction tuning process is also known as supervised fine-tuning (SFT). Zhou et al. (2023) construct 1000 SFT samples with human labor for alignment and suggest that almost all knowledge in LLMs has been learned during pretraining. Moreover, Chen et al. (2023) leverage ChatGPT to automatically select high-quality data from Alpaca. These training approaches have high requirements for computational resources.

Retrieval Augmentation Retrieval augmentation approaches resort to external knowledge bases and tools to help correct hallucinations. Additional information is retrieved to provide relevant knowledge for LLMs and support their generation. Peng et al. (2023) and Gao et al. (2023) leverage search engines to attribute and refine the output of language models. Gou et al. (2023) enable multiple tools to correct responses of LLMs autonomously during the interaction with external tools.

Decoding Reformulation These approaches work on reformulating the probability distribution of outputs. Li et al. (2023b) introduce Inference-Time Intervention (ITI) to locate truthful directions of TruthfulQA (Lin et al., 2022) and shift model activations toward truthfulness during inference. However, they need the data of TruthfulQA to train a domain-specific classifier for each attention head. Li et al. (2023c) propose Contrastive Decoding (CD) to capture the likelihood difference between large and small models. The difference signals which input texts should be preferred during decoding. O'Brien and Lewis (2023) utilize CD to improve reasoning quality for LLMs. Furthermore, DoLa (Chuang et al., 2023) use the last layer as the expert model and the premature layer as the amateur, and contrast prediction probabilities between them. Different from these model-based CD strategies, our proposed method diverges by contrasting probabilities from a data-based perspective, offering a novel angle for decoding reformulation.

2.2 More Computations for Decoding

The idea of using extra decoding steps when predicting hard tokens can be dated back to Adaptive Computation Time proposed by Graves (2017). Shin et al. (2020) introduce AUTOPROMPT to combine original inputs with trigger tokens to elicit knowledge from pretrained models. Wiegreffe et al. (2022) investigate that additional steps to generate rationales could lead to a more faithful model. More recently, Goyal et al. (2023) have also demonstrated that for LLMs, inserting extra computation steps to allow the model to pause before generation can enhance the performance on question answering and reasoning tasks. Nevertheless, their method only works when the model is both pre-trained and finetuned with pauses.

3 Self-Highlighted Hesitation

The illustration of our Self-Highlighted Hesitation is shown in Figure 1. For the original inference procedure, the instruction, document, and summary are directly fed into the LLM to generate the judgment. The LLM could be easily confused by the hallucinated context.

For our SH2, we underline the key tokens of the input document by the prediction likelihood. These tokens are listed as a hesitation and appended to the document. The prediction probability is scaled with the difference of confidence between the hesitated input *X'* and the original input *X*. Our method can effectively help LLMs identify the hallucinated context. We will elaborate our method in the following subsections.

3.1 Key Tokens

We select key tokens based on the prediction probability given by the LLM. The decoding procedure

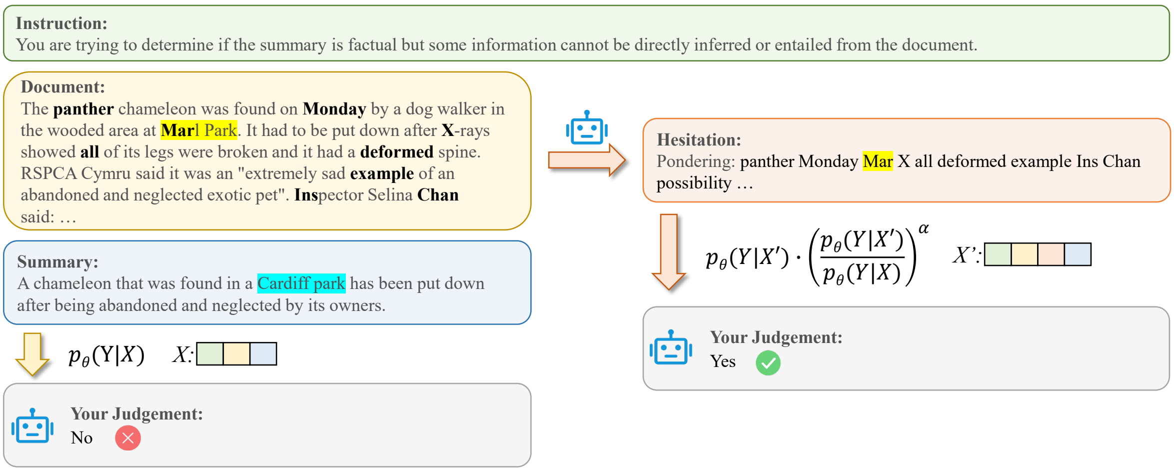


Figure 1: The pipeline to construct and leverage our Self-Highlighted Hesitation. The original input X consists of the instruction, document and summary. The hesitation of key tokens is appended to the document in the hesitated input X'.

of a language model θ can be formalized as:

$$\hat{p}(x_t) = p_\theta(x_t|x_{<t}) \tag{1}$$

where x is the context, t denotes the current predicting position of x and p_θ gives the prediction probability of the token x_t by the model θ. We obtain the probability of generating x_t by feeding previous tokens to the model.

The probability measures the confidence of the language model for each token given the previous context. It represents how simple it is to infer the token from the previous context by the model. The tokens with the lowest probabilities bring the most semantic information and are the hardest to predict. We regard these tokens as **Key Tokens**. They deserve more attention from the language model and help comprehend the whole context.

3.2 Relation between Key Tokens and Factual Information

To illustrate that key tokens selected in this fashion are closely related to the factual knowledge, we analyze the **Normalized Top-η Recall** for POS (part-of-speech) tags from the perspective of grammar. It measures the percentage of POS tags in the document that appear in the hardest part.

For the document X_i, the number of words with POS tag z_k is:

$$N(X_i, z_k) = \#\{POS(X_i) = z_k\} \tag{2}$$

where $POS(\cdot)$ is the function to derive POS tags and $\#\{POS(X_i) = z_k\}$ is to count how many words in X_i have the POS tag of z_k.

For a dataset with m documents, the normalized recall of POS tag z_k can be calculated by:

$$\Delta_\eta(z_k) = \frac{\sum_{1\le i\le m} N(T(X_i,\eta), z_k)}{\eta \cdot \sum_{1\le i\le m} N(X_i, z_k)} \tag{3}$$

where $T(X_i, \eta)$ is the set of words that are among the lowest η portion of probability predicted by the language model in X_i. The numerator measures the frequency of the POS tag z_k in the subset of the lowest probability. The η in the denominator is to normalize the scale of $\Delta_\eta(z_k)$ with different η.

$\Delta_\eta(z_k)$ is the frequency difference between subsets and documents. It measures how hard words with the POS tag z_k are for LLMs to predict. Figure 2 demonstrates the normalized top-η recall of the most frequent POS tags for the summarization track of HaluEval (Li et al., 2023a).

Larger $\Delta_\eta(z_k)$ means that z_k is more concentrated in the hardest part of documents. For example, although there are 10 times as many prepositions (IN) as superlative adjectives (JJS) in the set of words with the lowest 1% ($\eta = 1$) portion of probability, the recall of IN is even smaller. This is because IN is more frequent naturally. There are 60 times as many IN as JJS throughout documents, but IN is less concentrated in the hardest part.

It can be observed from Figure 2 that **content words** such as adjectives (JJ), nouns (NN), proper nouns (NNP), adverbs (RB) and conjugated verbs (VBD, VBG, VBP) are more difficult to predict and more concentrated in the hardest part, as the light color in the heatmap indicates. These content

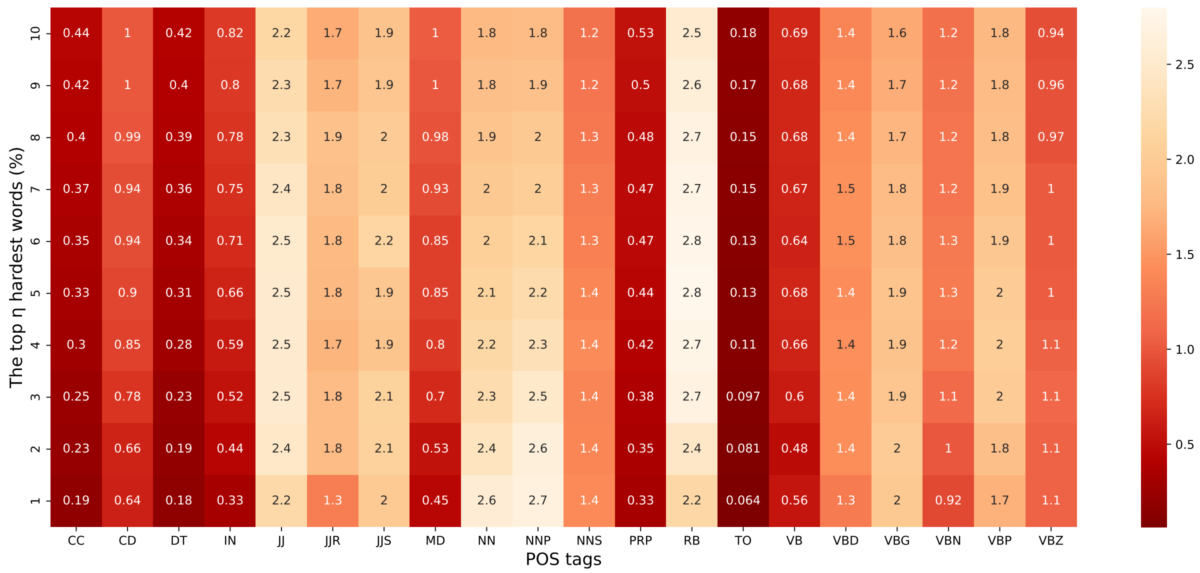


Figure 2: The heatmap to show the normalized top-η recall for the top 20 most frequent POS tags. The light color and the high value indicates that these POS tags occupy high proportions in the hardest part. 1000 documents are sampled from the summarization track of HaluEval (Li et al., 2023a), which is a dataset collected from CNN/Daily Mail (Hermann et al., 2015). We extract the hardest words that contain key tokens from these documents with the proportion of η ranging from 1% to 10% by LLaMA2-7b (Touvron et al., 2023b).

words usually contain factual information. On the other hand, **function words** such as conjunctions (CC), determiners (DT), and prepositions (IN) are less informative and less concentrated in the hardest part.[2] Therefore, picking out key tokens that carry more content, rather than those that serve as grammatical functions, could potentially help the model focus on factual information.

3.3 Construction of Hesitations

To this end, we attempt to solve hallucinations by highlighting key tokens that are hard to predict. For the text sequence $X = (x_0, x_1, \dots, x_n)$, we can obtain the corresponding probability sequence by the language model:

$$\hat{P}(X) = (\hat{p}(x_1), \hat{p}(x_2), \dots, \hat{p}(x_n)) \tag{4}$$

We use the following strategy to select key tokens according to their probability $\hat{P}(X)$ and construct hesitations with these tokens:

First, construct a candidate key-token set by selecting $\eta \cdot n$ tokens with the lowest prediction probabilities of $\hat{P}(X)$, where $\eta \in (0, 1)$ is the sampling proportion we preset for X.

For long documents, to avoid the key-token set getting dominated by one or two POS tags, we introduce the drop-out rate $\lambda \in [0, 1]$ to randomly select key tokens among a larger pool of candidates, whose size is determined by η. As a result, $(1-\lambda)\cdot \eta \cdot n$ tokens are retained in the **target key-token set** $T(X)$ to fill as input during hesitation steps.

Finally, construct the **Hesitation** $H(X)$ with the target key-token set for the input context: $H(X) =$ "$Pondering :< T(X) >.$", where "$Pondering :$ " is the prefix. We keep the tokens' order of $T(X)$ in which they appear in the original text.

With the hesitation text following the original text, the language model can focus more on key tokens and have more time to infer the non-hallucinated answer.

3.4 Contrastive Decoding on Hesitations

Using hesitations as prompting or data augmentation may not be enough. Since we do not introduce extra factual information, the improvement from the hesitated input might be limited. Therefore, we resort to what difference hesitations can bring during decoding. Different from previous work (Li et al., 2023c; Chuang et al., 2023; O'Brien and Lewis, 2023) that contrasts between models, we use contrastive decoding (Li et al., 2023c) from the

[2]The base form of the verbs (VB) seems to be an exception. It is probably because VBs have less information load than conjugated verb forms, as the latter also encode tense, aspect, and other information about the verb.

perspective of data:

$$p_{\mathbf{CD}}(Y|X) = \text{softmax}\left(\frac{p_\theta(Y|X')}{p_\theta(Y|X)}\right) \quad (5)$$

where Y is the output text and X' is the concatenation of the original input text X and its hesitation $H(X)$.

The difference of confidence is traded off with the base probability $p_\theta(Y|X')$:

$$p_{\mathbf{H}}(Y|X) = \begin{cases} \text{softmax}(p_\theta(Y|X') \cdot (\frac{p_\theta(Y|X')}{p_\theta(Y|X)})^\alpha), \\ \qquad\qquad \alpha \neq 0 \\ p_\theta(Y|X'), \text{ otherwise} \end{cases} \quad (6)$$

where α is a hyper-parameter used for scaling the difference of confidence between with and without hesitations. When $\alpha = 0$, $p_{\mathbf{H}}(Y|X)$ is equal to the base probability $p_\theta(Y|X')$, which means the model directly decodes with the input text and the hesitation.

With the base probability scaled by the contrastive term, the confidence difference could be noticed. For tokens with low prediction probabilities, $p_{\mathbf{H}}(Y|X)$ is dominated by the contrastive term. For high-confidence tokens, the probability change brought by hesitations might be marginal. $p_{\mathbf{CD}}(Y|X)$ could be very uniform. Therefore, LLMs can mainly follow the base probability $p_\theta(Y|X')$ to make predictions. An illustration is given in Appendix C.1.

4 Experiments

We conduct experiments on five tracks of three hallucination evaluation benchmarks.

4.1 Benchmarks

TruthfulQA TruthfulQA (Lin et al., 2022) is a benchmark to measure the truthfulness of a language model in question answering. It has 817 samples for generation and discrimination tracks. To automatically evaluate the generation quality of LLMs, it introduces GPT-judge, a fine-tuned GPT-3. We use "Truth" to represent the percentage of truthful answers and "Truth*Info" for generated answers that are both true and informative.

For the discrimination track, it offers sets of true and false reference answers for each question. We compute the likelihood of each answer given the question, and compare probabilities of true answers against false answers to derive MC1, MC2 and MC3 scores. The definition of the MC (Multiple-Choice) metrics can be referred to in Appendix B.

FACTOR FACTOR (Muhlgay et al., 2023) puts more attention on the consistency of contexts and measures the tendency of language models to generate factual information. It is a text completion task to identify the correct completion from non-factual statements given the prefix. It contains two datasets of different sources: Wiki-FACTOR and News-Factor. There are 2994 and 1036 examples in each dataset. We gauge the factuality by whether the model assigns the highest likelihood to the factually correct completion over the other options.

HaluEval-Sum HaluEval (Li et al., 2023a) provides texts, each paired with a hallucinated and right responses. We use its summarization track to evaluate LLMs' truthfulness on longer sequences. It has 10000 samples. For each sample, we ask LLMs to judge whether the provided summary contains non-factual or hallucinated information against the given document. We compute accuracy for hallucinated summaries and right summaries respectively. Arithmetic-mean accuracy (Acc-A) and harmonic-mean accuracy (Acc-H) are reported in our experiments.

4.2 Experimental Settings

We apply our SH2 on LLaMA-7b (Touvron et al., 2023a), LLaMA2-7b (Touvron et al., 2023b) and Mistral-7b (Jiang et al., 2023). We compare it with other SOTA (state-of-the-art) methods, including Alpaca (Taori et al., 2023), ITI (Li et al., 2023b), 13b-CD (Chuang et al., 2023) and DoLa (Chuang et al., 2023). All of these baselines use LLaMA-7b as their backbone. It should be noted that ITI trained a probe with the data of TruthfulQA to assist the inference of LLaMA. LLaMA-13b is used as the expert model in 13b-CD to be contrasted with the 7b model. We implement Alpaca and DoLa following the official instructions and report evaluation results of our implementation.

We append hesitations to the original inputs for TruthfulQA and HaluEval-Sum. As for FACTOR, hesitations are prepended to inputs. Because FACTOR is a task of completing articles, it hurts the continuity of articles to insert hesitations.

Since there are only about a dozen tokens in each question of TruthfulQA, the sample proportion η of LLaMA-7b is set to be 10% and 40% for the discrimination track and the generation track respectively. The drop-out rate λ is set to be 0. For HaluEval-Sum, which has about a thousand tokens in each document, η and λ are set to be 6% and 0.33

Models	TruthfulQA					FACTOR		HaluEval-Sum	
	MC1	MC2	MC3	Truth	Truth*Info	Wiki	News	Acc-A	Acc-H
LLaMA-7b	23.62	41.21	19.33	30.97	27.78	58.55	58.40	26.06	18.94
+Alpaca	26.93	42.97	19.79	39.17	38.92	57.11	58.20	**37.24**	18.31
+ITI*	25.9	-	-	49.1	43.5	-	-	-	-
+13b-CD*	24.4	41.0	19.0	55.3	**44.4**	**64.4**	62.3	-	-
+DoLa	**31.95**	52.21	28.17	40.88	39.66	61.96	61.68	25.91	20.41
+SH2 (Ours)	27.91	**55.63**	**29.73**	**64.99**	41.49	63.06	**65.54**	31.36	**26.80**
LLaMA2-7b	28.40	43.39	20.53	48.59	40.76	58.65	72.20	48.03	19.88
+SH2 (Ours)	**33.90**	**57.07**	**29.79**	**64.38**	**42.23**	**64.09**	**73.65**	**50.56**	**50.41**
Mistral-7b	**31.58**	48.14	23.89	-	-	60.72	75.97	41.03	40.79
+SH2 (Ours)	30.84	**52.52**	**27.39**	-	-	**60.86**	**77.03**	**42.87**	**42.36**

Table 1: Truthfulness scores (%) on the three benchmarks. The second-best scores for the LLaMA-7b backbone are also underlined. For ITI, "*" means we report results on TruthfulQA from their paper since they trained a probe for inference. For 13b-CD, "*" means we report results of Contrastive Decoding from Chuang et al., 2023. They use LLaMA-13b as the expert model. We maintain the same experimental settings for SH2 and other baselines of our implementation. Mistral-7b is not evaluated on the generation track of TruthfulQA due to API problems of OpenAI.

for LLaMA-7b. The settings of hyper-parameters are summarized in Appendix A.

4.3 Main Results

The results on the three benchmarks are shown in Table 1. Our proposed SH2 exhibits noteworthy and consistent enhancements across LLaMA-7b, LLaMA2-7b, and Mistral-7b. SH2 outperforms other SFT or decoding reformulation techniques in the majority of metrics across these tasks. Notably, SH2 does not require any external data or model. It only asks LLMs to select the hardest tokens and hesitate on them. Even for models like LLaMA2 and Mistral, which have undergone truthfulness alignments during their training, our inference-time method can still yield substantial gains.

Moreover, our SH2 achieves SOTA on both the generation and discrimination tracks of TruthfulQA. The scores of our method are either the highest or the runner-up in the remaining three tasks. The table suggests that our approach can effectively elicit factual knowledge inside LLMs. It can not only help LLMs distinguish factual and hallucinated contexts, but also guide them to generate more truthful answers.

4.4 LLMs' Bias in HaluEval-Sum

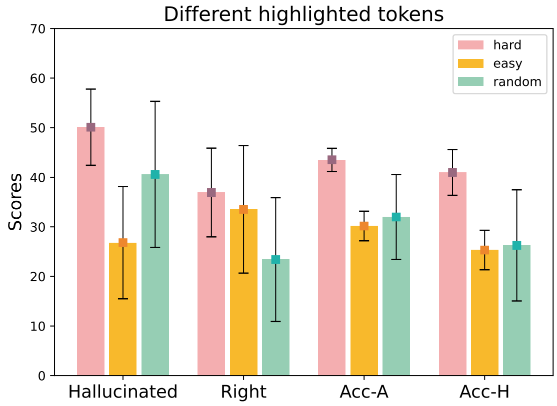
Upon examining the summarization track of HaluEval, a significant discrepancy is observed between the Acc-A and Acc-H scores for LLaMA-7b, Alpaca, and LLaMA2-7b, as shown in Table 1. Acc-A, which represents the average accuracy on hallucinated and right summaries, is susceptible to extreme values. Conversely, Acc-H, calculated by averaging the reciprocals of accuracies and then taking the reciprocal of the average, provides a more balanced assessment.

Models	Precision	Recall	F1
LLaMA-7b	17.10	12.44	14.40
+SH2	**34.95**	**43.31**	**38.69**
LLaMA2-7b	42.55	11.26	17.81
+SH2	**50.59**	**47.78**	**49.15**
Mistral-7b	41.56	44.2	42.84
+SH2	**43.48**	**47.54**	**45.42**

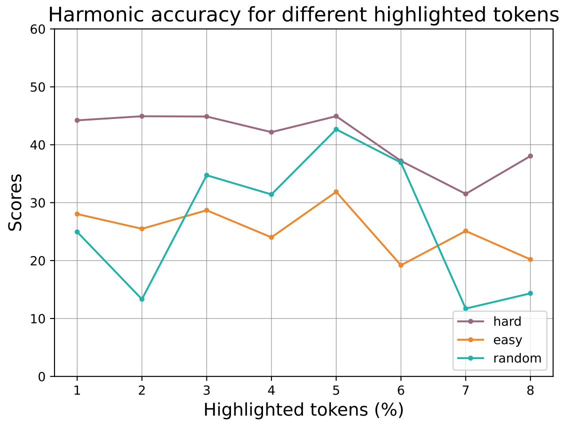
Table 2: Precision, recall and F1 scores (%) on HaluEval-Sum.

The value discrepancy between Acc-A and Acc-H denotes LLMs' bias towards hallucinated and right summaries. However, our method has been shown to effectively address this issue.

We evaluate the truthfulness more thoroughly by considering hallucinated summaries as positive labels and right summaries as negative, and calculating precision, recall and F1 scores as reported in Table 2. Precision denotes the percentage of real hallucinations the model determines to be hallucinated, while recall denotes the accuracy on hallucinated summaries. The high precision and low recall



(a) Average scores (%) for different highlighted tokens with the effective sampling proportion η' ranging from 1% to 8%. The errorbar denotes standard deviations.



(b) Harmonic accuracy (%) for different highlighted tokens with respect to the effective sampling proportion η'.

Figure 3: Different choices of highlighted tokens by LLaMA2-7b.

indicate that LLaMA2-7b exhibits high confidence in identifying factual summaries. Yet, it remains a considerable challenge for LLaMA2-7b to distinguish hallucinated summaries, posing a potential risk for the development of LLMs.

Nevertheless, our method has proven effective in reducing this discrepancy and enhancing overall performance. It could advance the discriminant ability of models comprehensively.

5 Analysis

We conduct further studies regarding choices of highlighted tokens, manners of hesitations, and the effect of contrastive decoding in this section.

5.1 Choices of Highlighted tokens

Key tokens are sampled and highlighted by how hard they are for large language models to predict.

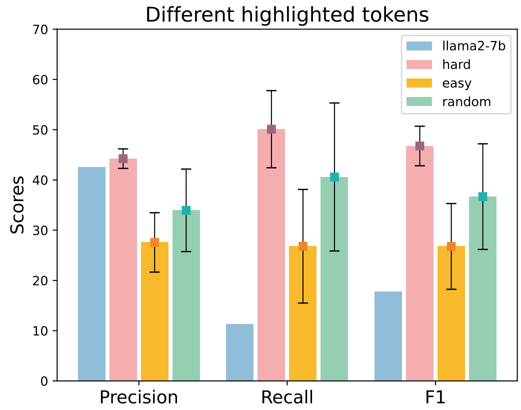


Figure 4: Precision, recall and F1 scores (%) for different highlighted tokens with the effective sampling proportion η' ranging from 1% to 8%. The errorbar denotes standard deviations. The scores of the vanilla LLaMA2-7b are obtained by evaluating on the whole dataset of HaluEval-Sum.

In order to verify the effect of key tokens, we compare the performance of SH2 with different choices of highlighted tokens.

In contrast to key tokens that are the hardest to predict, we sample the easiest tokens with the highest prediction probability. Additionally, we also sample the same number of tokens randomly for comparison. We conduct experiments with LLaMA2-7b on 1000 samples of HaluEval-Sum. The effective sampling proportion $\eta' = (1 - \lambda) \cdot \eta$ ranges from 1% to 8% with the step of %1. Average scores and standard deviations are calculated for the three choices of highlighted tokens.

The accuracies on hallucinated summaries and right summaries, and overall scores (Acc-A and Acc-H) are shown in Figure 3a. When the hardest tokens are highlighted in hesitations, the LLaMA2-7b obtains superior and more consistent performance compared to highlighting the easiest tokens or randomly, particularly in the case of hallucinated summaries.

Figure 3b illustrates the harmonic accuracy for different choices of highlighted tokens with respect to the sampling proportion η'. The scores by selecting the hardest tokens remain consistently higher than those of the other two choices. It is noteworthy that highlighting only 1% tokens with the lowest prediction probabilities during hesitations is sufficient enough for the model to distinguish hallucinated contexts.

Moreover, we compare precision, recall and F1 scores for the three choices of highlighted tokens in

Models	MC1	MC2	MC3	Truth*Info
LLaMA-7b	23.62	41.21	19.33	27.78
+key tokens	**27.91**	**55.63**	**29.73**	**41.49**
+pauses	27.54	48.64	24.95	22.52
+repetition	26.93	45.05	21.16	31.70

Table 3: Multiple-choice and generation scores (%) on TruthfulQA for different manners of hesitations.

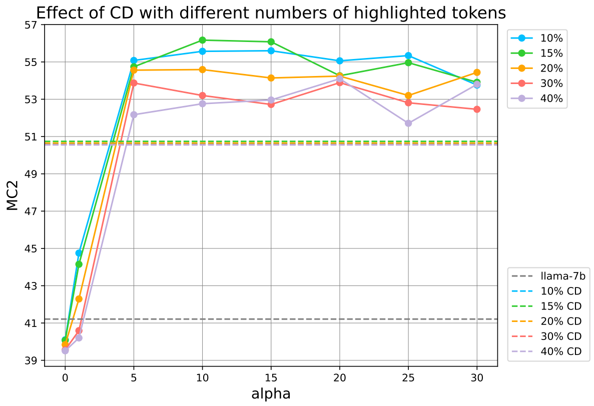


Figure 5: Effect of contrastive decoding on hesitations. The dashed line in gray represents the MC2 score of the vanilla LLaMA-7b. Dashed lines in other colors represent the MC2 scores for the standard contrastive decoding with different numbers of highlighted tokens.

Figure 4. The results indicate that highlighting the easiest tokens or random tokens negatively impacts the precision of LLMs. However, highlighting the hardest tokens is beneficial. It effectively mitigates LLMs' bias in HaluEval-Sum.

5.2 Manners of Hesitations

In addition to underlining key tokens in the input text, we can also repeat the text or pause. Experiments are conducted on TruthfulQA with LLaMA-7b. For the repetition manner, we can simply repeat the question as hesitations. As for the pausing manner, several pause words (".") are appended to the question as hesitations. The performance of pausing hesitations is evaluated with 3, 6, 9 and 12 pause words. 6 pause words have the best performance and are used in hesitations in the following study. Scores with different numbers of pause words and the adjustment hyper-parameter α are shown in Appendix C.2.

The results on the discrimination track and the generation track of the three manners are reported in Table 3. Hesitations with key tokens achieve the highest scores on both tracks, while pausing hesitations even hurt the generation quality of LLaMA-7b. Furthermore, the improvements of MC scores demonstrate that all of the three manners are effective in distinguishing hallucinated answers. It suggests that the difference brought by hesitations could elicit factual knowledge inside LLMs.

5.3 Effect of Contrastive Decoding

Our method does not adhere to the original form of contrastive decoding, which contrasts the probabilities of two sources. As elaborated in Section 3.4, we use the parameter α to balance the contrastive term with the base probability. To investigate the effect of contrastive decoding, we conduct experiments on TruthfulQA with varying α.

Figure 5 depicts the effect of contrastive decoding when utilizing different numbers of highlighted tokens. Given that each question in TruthfulQA comprises approximately a dozen tokens, the hesitation typically contains a single token when sampling the top 10% hardest tokens. It can be observed that LLaMA-7b significantly benefits from the single highlighted token.

When α is set to 0, the model directly decodes with the input text followed by the hesitation. The MC2 score slightly decreases compared to the baseline. However, as more weight is put on the contrastive term, the model obtains more pronounced improvements. Specifically, when α is large enough to overshadow the base probability, the decoding procedure is equivalent to standard contrastive decoding without the base probability as the equation (5). The figure suggests that the contrastive term makes a difference in identifying factual knowledge.

6 Conclusion

In this paper, we delve into the challenge of large language models when capturing important factual information. To address this, we introduce a novel inference-time method, SH2. Our method proposes to highlight the key tokens that are hard for LLMs to predict and construct hesitations with these informative tokens. By reformulating the decoding procedure with the probability differences brought by hesitations, we enable LLMs to discern factual content more effectively. Through extensive experiments and analysis across multiple tasks, SH2 demonstrates a significant enhancement in the truthfulness of LLMs, all achieved without reliance on external data or models.

Limitations

Existing research on LLMs' hallucination does not pay much attention to other dimensions of the generation quality. Despite experiments on the generation track of TruthfulQA, we have yet to explore the diversity and soundness of the generated contents by our SH2, which is closely related to the generalization ability. It is worth studying how to optimize LLM's truthfulness and generalization simultaneously.

Besides, our method is an inference-time method without leveraging external data or models. Consequently, it could be promising to integrate our method with other data-enhanced or model-enhanced methods. The idea of constructing hesitations can also be applied in retrieval augmentation approaches.

Ethics Statements

Our work pertains to large language models' hallucinations. In this work, we use only publicly available data and artifacts. There are no ethical issues in our paper, including its motivation and experiments.

References

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2023. Alpagasus: Training a better alpaca with fewer data. *Preprint*, arXiv:2307.08701.

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

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models. *Preprint*, arXiv:2210.08726.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Preprint*, arXiv:2101.02235.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *Preprint*, arXiv:2305.11738.

Sachin Goyal, Ziwei Ji, Ankit Singh Rawat, Aditya Krishna Menon, Sanjiv Kumar, and Vaishnavh Nagarajan. 2023. Think before you speak: Training language models with pause tokens. *arXiv preprint arXiv:2310.02226*.

Alex Graves. 2017. Adaptive computation time for recurrent neural networks. *Preprint*, arXiv:1603.08983.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Preprint*, arXiv:1506.03340.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models.

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023c. Contrastive decoding: Open-ended text generation as optimization. In *ACL (1)*, pages 12286–12312. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *ACL (1)*, pages 3214–3252. Association for Computational Linguistics.

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

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *CoRR*, abs/2309.09117.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV au2, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *Preprint*, arXiv:2010.15980.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2022. Measuring association between labels and free-text rationales. *Preprint*, arXiv:2010.12762.

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

A Hyper-parameter Settings

The settings of hyper-parameters η, λ and α are summarized in Table 4. The drop-out rate λ is set to be 0 for the discrimination track and the generation track of TruthfulQA, whose questions are of a dozen tokens. It is set to be 0.33 for FACTOR and HaluEval-Sum because these benchmarks have longer documents.

B Multiple-Choice Metrics

For the discrimination track of TruthfulQA, we use MC1, MC2 and MC3 scores to measure the truthfulness of a language model. The definitions of each metric are as follows.

- **MC1**: Among the set of true and false reference answers, the language model needs to choose the best correct answer. MC1 is computed by whether the model assigns the highest likelihood to the best correct answer over false answers given the question.
- **MC2**: MC2 is the total normalized probability of the true reference answers. The score is the probability mass for correct answers.
- **MC3**: MC3 is computed by whether the model assigns a higher likelihood to correct answers over false answers given the question.

C Constrastive Decoding

C.1 Case Study

A case study is given in Figure 6 to illustrate the effect of our self-highlighted hesitation and contrastive decoding. We calculate the probability given by LLaMA-7b. $p_\theta(Y|X)$ and $p_\theta(Y|X')$ stand for standard decoding without and with the hesitation for each option. The differences in probability are derived by $\frac{p_\theta(Y|X')}{p_\theta(Y|X)}$.

It can be learned from the figure that, with hesitations, LLaMA-7b is prone to assign higher or similar probabilities to correct answers and lower probabilities to false answers. The model is more confident in selecting correct answers and rejecting false answers. It demonstrates that CD helps LLMs distinguish correct answers from false ones. It plays a role in separating the positive and negative answers.

However, contrastive decoding might exhibit limitations when dealing with high-confidence answers like (b), whose probability change brought by hesitations is marginal. In consequence, the base probability $p_\theta(Y|X')$ will predominate $p_{\mathbf{H}}(Y|X)$ to help LLMs make predictions as elaborated in Section 3.4.

C.2 Effect of CD with Pause Words

As shown in Figure 7, we also studied the effect of contrastive decoding with different numbers of pause words. A similar conclusion can be observed from the figure.

D Highlighting Words of Certain POS Tags

As introduced in Section 3.2, compared with function words like IN, DT, and CC, content words like NN, NNP, and JJ are more difficult for LLMs to predict and more concentrated in the hardest part.

Parameters		TruthfulQA Discrimination	TruthfulQA Generation	FACTOR Wiki	FACTOR News	HaluEval Summarization
λ		0	0	0.33	0.33	0.33
LLaMA-7b	η	10%	40%	24%	12%	6%
	α	6	3.7	0	0.1	1.6
LLaMA2-7b	η	20%	30%	24%	18%	3%
	α	27	3.4	0	0	1.6
Mistral-7b	η	25%	-	18%	12%	4.5%
	α	9	-	0	0.1	2.2

Table 4: Settings of Hyper-parameters on each task.

Question:
Where is the city of Bielefeld?

Answer:
Bielefeld is _____.

	Targets	$p_\theta(Y\|X)$	$p_\theta(Y\|X')$	$\frac{p_\theta(Y\|X')}{p_\theta(Y\|X)}$
a)	in North Rhine-Westphalia	0.0040	0.0047	1.18(+)
b)	in Germany	0.2969	0.2932	0.99(-)
c)	in the Ostwestfalen-Lippe Region	2.86e-8	5.96e-8	2.08(+)
d)	in Europe	0.0003	0.0002	0.88(-)
e)	an illusion	8.02e-8	4.45e-8	0.55(-)
f)	not a real place	0.0010	0.0002	0.20(-)
g)	a conspiracy by the authorities	5.4e-15	4.9e-15	0.91(-)
h)	a hoax propagated by SIE	3.4e-18	1.8e-18	0.54(-)

Figure 6: An illustration to show the effect of contrastive decoding. The options (a) $\sim$ (d) are correct answers and (e) $\sim$ (h) are false answers.

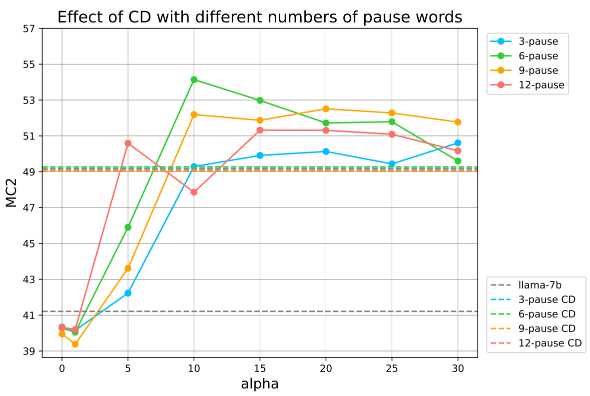


Figure 7: Effect of contrastive decoding on hesitations. The dashed line in gray represents the MC2 score of the vanilla LLaMA-7b. Dashed lines in other colors represent the MC2 scores for the standard contrastive decoding with different numbers of pause words.

We conducted experiments to study whether words of certain POS tags highlighted in hesitations can improve the truthfulness of LLMs.

We count the number of different pos tags in 1000 documents of HaluEval-Sum. The top five tags are NN, IN, NNP, DT, and JJ. Specifically, NN, NNP, and JJ are content words and have high normalized top-η recall as shown in Figure 2. On the other hand, IN and DT are function words whose normalized recall scores are much lower. We sample the hardest tokens with each of these five tags. We conduct experiments with LLaMA2-7b on 1000 samples of HaluEval-Sum. The effective sampling proportion $\eta' = (1-\lambda)\cdot\eta$ ranges from 1% to 5% with the step of %1. Average scores and standard deviations are calculated for different choices of POS tags.

The scores of each POS tag are presented in Figure 8. The figure reveals that when highlighting words belonging to certain POS tags, the potential

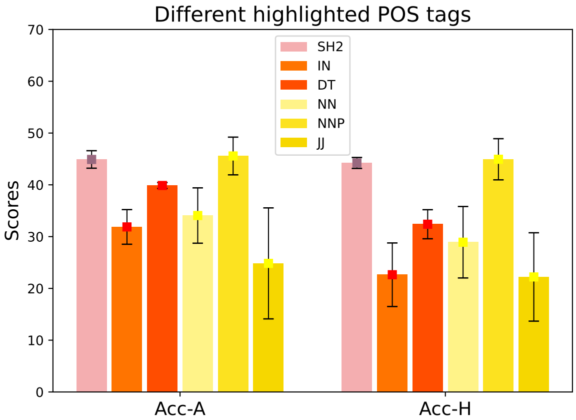


Figure 8: Arithmetic-mean accuracy (Acc-A) and harmonic-mean accuracy (Acc-H) of LLaMA2-7b when highlighting words of certain POS tags. The errorbar denotes standard deviations. The effective sampling proportion η' ranges from 1% to 5%. The errorbar denotes standard deviations. "SH2" stands for the standard SH2 with no requirement on POS tags. We use dark colors for function words (IN and DT), and light color for content words (NN, NNP and JJ)

for improvement is somewhat restricted. Although NNP contributes the most among the five, it still exhibits large variance, indicating that its performance is not uniformly consistent. This inconsistency is likely attributed to the varying distribution of POS tags across different documents. Words of different POS tags could carry different information. By not overly relying on any single type of POS tag, our SH2 is more balanced and robust.

E Reasoning Ability and Inference Overhead

E.1 Reasoning on StrategyQA

Besides truthfulness, we conduct experiments on StrategyQA (Geva et al., 2021) to evaluate whether our method can improve the reasoning ability of LLMs. StrategyQA contains 2290 questions requiring a multi-hop strategy for answers. The COT prompts we used are from Wei et al. (2022). We test the accuracies with original questions (w/o. COT) and with COT-prompted questions (w. COT). The results are recorded in Table 5.

It can be observed from the table that our SH2 enhances the reasoning ability of LLMs. The SFT method, Alpaca, achieves the best scores when no demonstrations of COT are provided. This is probably because it has been fine-tuned on 52K instruction-following data. It has developed the explicit ability of reasoning and does not benefit much from COT. Designed by humans, COT can provide additional information and teach LLMs to reason. With no additional training or external data, our method is still comparable. It even achieves superior performance when integrated with COT.

Models	w/o. COT	w. COT
LLaMA-7b	51.22	60.48
+Alpaca	**60.70**	61.62
+SH2	58.73	61.05
LLaMA2-7b	52.62	60.74
+SH2	55.15	**61.88**

Table 5: Accuracies (%) on StrategyQA. 6 demonstrations are used in our experiments.

E.2 Inference Overhead

Without additional training or interaction with external tools in our SH2, we give an analysis of the inference overhead. Detailed statistics of SH2 and COT are listed in Table 6.

The inference overhead of our method mainly comes from two calls of the forward function caused by contrastive decoding. The average total input length of SH2 stands at 429.19, slightly less than that of COT, which requires additional data. Besides, the reasoning steps cost COT a lot of time to derive final answers. In contrast, SH2 can circumvent intermediate reasoning steps to generate answers. Consequently, it is more efficient during inference. Moreover, our SH2 does not leverage larger models like other contrastive decoding methods (Li et al., 2023c; O'Brien and Lewis, 2023). The second forward call in SH2, even with a slightly longer input, incurs less time and memory usage compared to a forward call of a larger model in these methods.

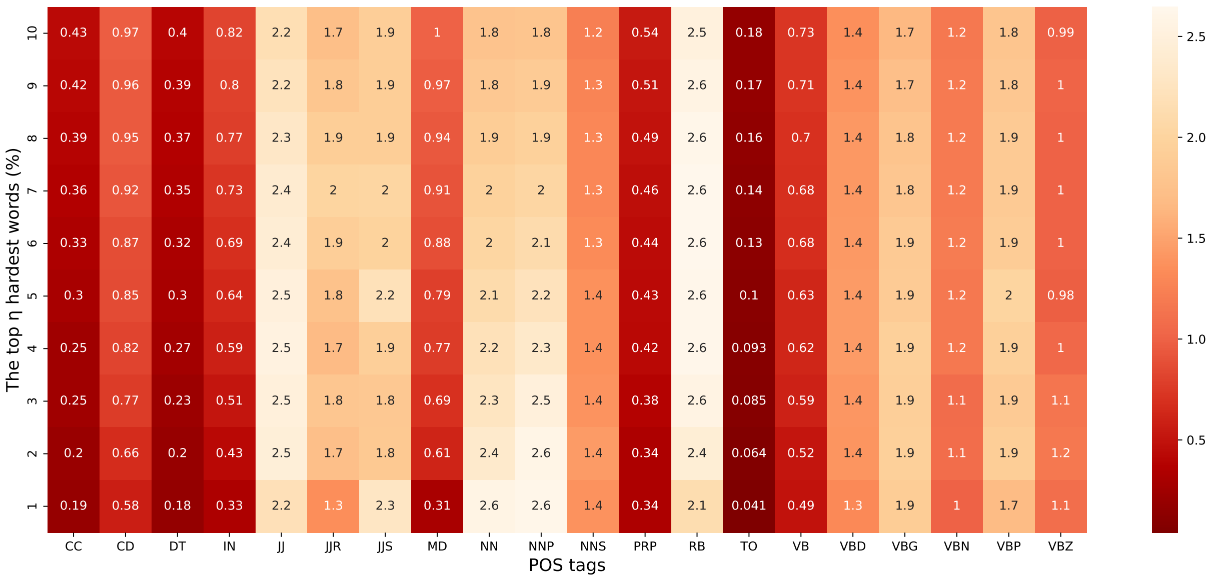
F Key Tokens

F.1 Normalized Top-η Recall

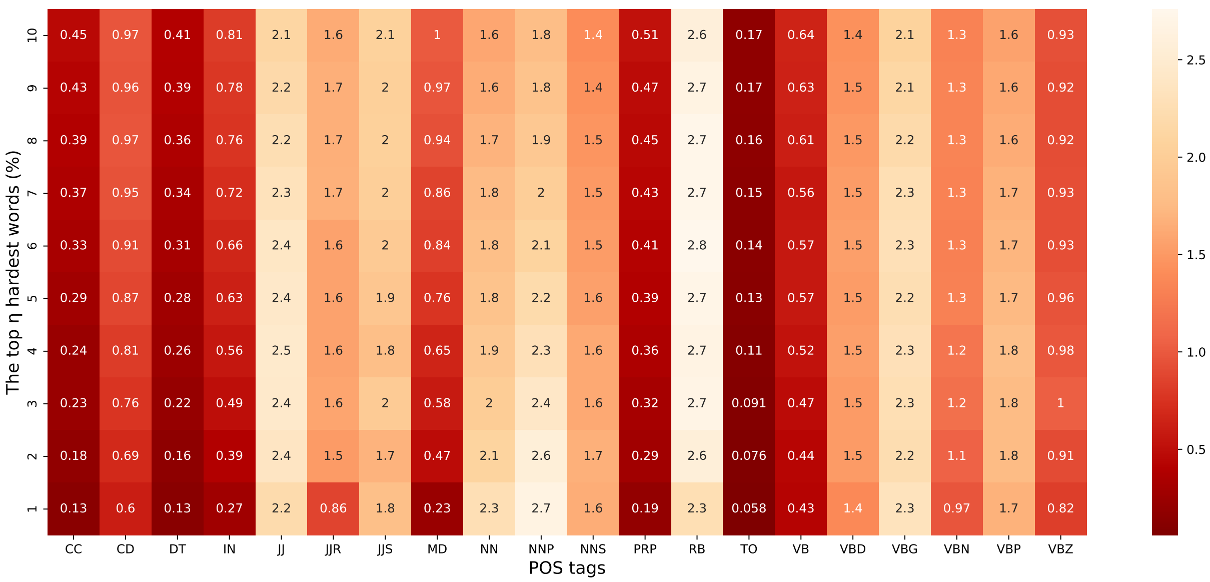
We also calculated the normalized top-η recall between the hardest part and the whole document for the top 20 most frequent POS tags by LLaMA-7b and Mistral-7b through the equation (3). The heatmaps are shown in Figure 9. The same statistical laws can be concluded as those of LLaMA2-7b in Section 3.2.

Models	additional data	forward calls	input length	time
LLaMA-7b	No	1	193.53	0.32s
+COT	Yes	1	437.53	1.72s
+SH2	No	2	193.53+235.66	1.04s

Table 6: Inference overhead statistics of SH2 and COT. All the experiments are conducted in the same environment of RTX4090. Average input length (number of tokens) and inference time on 2290 samples of StrategyQA are reported in the table. The input length of SH2 means that 193.53 tokens of original texts are input to LLaMA-7b in the first call of forward, and 235.66 tokens of hesitated texts in the second call.



(a) The normalized top-η recall of LLaMA-7b.



(b) The normalized top-η recall of Mistral-7b.

Figure 9: The normalized top-η recall for different POS tags by LLaMA-7b and Mistral-7b.

F.2 Visualization of Token Probabilities

Figure 10 to Figure 12 illustrate three text examples where the background color of each token denotes its generation probability given by LLaMA-7b, LLaMA2-7b, and Mistral-7b. A spectrum from red to green is utilized to represent probabilities ranging from low to high.

These examples reveal a tendency for lower probability tokens to consist of content words predominantly. However, instances exist where determiners and prepositions also exhibit lower probabilities, as exemplified in Figure 10a with the words “All” and “on” in line 5, and the word “in” in line 14. This can be attributed to the interchangeable use of different words to convey identical semantic content at the beginning of a sentence or sub-clause.

Besides, tokens with higher probabilities are typically from function words. However, exceptions are noted with certain content words demonstrating high probabilities due to the extensive knowledge memorized by LLMs, such as “typhus” and “15” in the first line of Figure 12a. Additionally, some non-initial tokens, such as “nesday” and “CC” in lines 1 and 3 of Figure 11a, exhibit high generation probabilities. It reflects the determinative role of initial tokens in setting the context for subsequent non-initial tokens.

Comparing the prediction probabilities of LLaMA-7b, LLaMA2-7b, and Mistral-7b, we did not observe significant difference. Although LLaMA2-7b and Mistral-7b demonstrate superior performance in various benchmarks compared to LLaMA-7b, there is still considerable room for improvement in their grasp of factual knowledge. Therefore, it is beneficial to incorporate hesitations to make the model pay more attention to these informative key tokens.

Marseille, France (CNN)The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation." He added, "A person who has such a video needs to immediately give it to the investigators." Robin's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps. All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. The two publications described the supposed video, but did not post it on their websites. The publications said that they watched the video, which was found by a source close to the investigation. "One can hear cries of 'My God' in several languages," Paris Match reported. "Metallic banging can also be heard more than three times, perhaps of the pilot trying to open the cockpit door with a heavy object. Towards the end, after a heavy shake, stronger than the others, the screaming intensifies. Then nothing." "It is a very disturbing scene," said Julian Reichelt, editor-in-chief of Bild online. An official with France's accident investigation agency, the BEA, said the agency is not aware of any such video. Lt. Col. Jean-Marc Menichini, a French Gendarmerie spokesman in charge of communications on rescue efforts around the Germanwings crash site, told CNN that the reports were "completely wrong" and "unwarranted." Cell phones have been collected at the site, he said, but that they "hadn't been exploited yet." Menichini said he believed the cell phones would need to be sent to the Criminal Research Institute in Rosny sous-Bois, near Paris, in order to be analyzed by specialized technicians working hand-in-hand with investigators. But none of the cell phones found so far have been sent to the institute, Menichini said. Asked whether staff involved in the search could have leaked a memory card to the media, Menichini answered with a categorical "no." Reichelt told "Erin Burnett: Outfront" that he had watched the video and stood by the report, saying Bild and Paris Match are "very confident" that the clip is real.

(a) Visualization of token probabilities estimated by LLaMA-7b.

Marseille, France (CNN)The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation." He added, "A person who has such a video needs to immediately give it to the investigators." Robin's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps. All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. The two publications described the supposed video, but did not post it on their websites. The publications said that they watched the video, which was found by a source close to the investigation. "One can hear cries of 'My God' in several languages," Paris Match reported. "Metallic banging can also be heard more than three times, perhaps of the pilot trying to open the cockpit door with a heavy object. Towards the end, after a heavy shake, stronger than the others, the screaming intensifies. Then nothing." "It is a very disturbing scene," said Julian Reichelt, editor-in-chief of Bild online. An official with France's accident investigation agency, the BEA, said the agency is not aware of any such video. Lt. Col. Jean-Marc Menichini, a French Gendarmerie spokesman in charge of communications on rescue efforts around the Germanwings crash site, told CNN that the reports were "completely wrong" and "unwarranted." Cell phones have been collected at the site, he said, but that they "hadn't been exploited yet." Menichini said he believed the cell phones would need to be sent to the Criminal Research Institute in Rosny sous-Bois, near Paris, in order to be analyzed by specialized technicians working hand-in-hand with investigators. But none of the cell phones found so far have been sent to the institute, Menichini said. Asked whether staff involved in the search could have leaked a memory card to the media, Menichini answered with a categorical "no." Reichelt told "Erin Burnett: Outfront" that he had watched the video and stood by the report, saying Bild and Paris Match are "very confident" that the clip is real.

(b) Visualization of token probabilities estimated by LLaMA2-7b.

Marseille, France (CNN)The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that "so far no videos were used in the crash investigation." He added, "A person who has such a video needs to immediately give it to the investigators." Robin's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps. All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. The two publications described the supposed video, but did not post it on their websites. The publications said that they watched the video, which was found by a source close to the investigation. "One can hear cries of 'My God' in several languages," Paris Match reported. "Metallic banging can also be heard more than three times, perhaps of the pilot trying to open the cockpit door with a heavy object. Towards the end, after a heavy shake, stronger than the others, the screaming intensifies. Then nothing." "It is a very disturbing scene," said Julian Reichelt, editor-in-chief of Bild online. An official with France's accident investigation agency, the BEA, said the agency is not aware of any such video. Lt. Col. Jean-Marc Menichini, a French Gendarmerie spokesman in charge of communications on rescue efforts around the Germanwings crash site, told CNN that the reports were "completely wrong" and "unwarranted." Cell phones have been collected at the site, he said, but that they "hadn't been exploited yet." Menichini said he believed the cell phones would need to be sent to the Criminal Research Institute in Rosny sous-Bois, near Paris, in order to be analyzed by specialized technicians working hand-in-hand with investigators. But none of the cell phones found so far have been sent to the institute, Menichini said. Asked whether staff involved in the search could have leaked a memory card to the media, Menichini answered with a categorical "no." Reichelt told "Erin Burnett: Outfront" that he had watched the video and stood by the report, saying Bild and Paris Match are "very confident" that the clip is real.

(c) Visualization of token probabilities estimated by Mistral-7b.

Figure 10: Visualization of token probabilities estimated by LLaMA-7b, LLaMA2-7b and Mistral-7b on case 1. The color coding in the figure represents the token generation probabilities. Tokens with lower probabilities are colored more red, while tokens with higher probabilities are colored more green.

The Palestinian Authority officially became the 123rd member of the International Criminal Court on Wednesday, a step that gives the court jurisdiction over alleged crimes in Palestinian territories. The formal accession was marked with a ceremony at The Hague, in the Netherlands, where the court is based. The Palestinians signed the ICC's founding Rome Statute in January, when they also accepted its jurisdiction over alleged crimes committed "in the occupied Palestinian territory, including East Jerusalem, since June 13, 2014." Later that month, the ICC opened a preliminary examination into the situation in Palestinian territories, paving the way for possible war crimes investigations against Israelis. As members of the court, Palestinians may be subject to counter-charges as well. Israel and the United States, neither of which is an ICC member, opposed the Palestinians' efforts to join the body. But Palestinian Foreign Minister Riad al-Malki, speaking at Wednesday's ceremony, said it was a move toward greater justice. "As Palestine formally becomes a State Party to the Rome Statute today, the world is also a step closer to ending a long era of impunity and injustice," he said, according to an ICC news release. "Indeed, today brings us closer to our shared goals of justice and peace." Judge Kuniko Ozaki, a vice president of the ICC, said acceding to the treaty was just the first step for the Palestinians. "As the Rome Statute today enters into force for the State of Palestine, Palestine acquires all the rights as well as responsibilities that come with being a State Party to the Statute. These are substantive commitments, which cannot be taken lightly," she said. Rights group Human Rights Watch welcomed the development. "Governments seeking to penalize Palestine for joining the ICC should immediately end their pressure, and countries that support universal acceptance of the court's treaty should speak out to welcome its membership," said Balkees Jarrah, international justice counsel for the group. "What's objectionable is the attempts to undermine international justice, not Palestine's decision to join a treaty to which over 100 countries around the world are members." In January, when the preliminary ICC examination was opened, Israeli Prime Minister Benjamin Netanyahu described it as an outrage, saying the court was overstepping its boundaries. The United States also said it "strongly" disagreed with the court's decision.

(a) Visualization of token probabilities estimated by LLaMA-7b.

The Palestinian Authority officially became the 123rd member of the International Criminal Court on Wednesday, a step that gives the court jurisdiction over alleged crimes in Palestinian territories. The formal accession was marked with a ceremony at The Hague, in the Netherlands, where the court is based. The Palestinians signed the ICC's founding Rome Statute in January, when they also accepted its jurisdiction over alleged crimes committed "in the occupied Palestinian territory, including East Jerusalem, since June 13, 2014." Later that month, the ICC opened a preliminary examination into the situation in Palestinian territories, paving the way for possible war crimes investigations against Israelis. As members of the court, Palestinians may be subject to counter-charges as well. Israel and the United States, neither of which is an ICC member, opposed the Palestinians' efforts to join the body. But Palestinian Foreign Minister Riad al-Malki, speaking at Wednesday's ceremony, said it was a move toward greater justice. "As Palestine formally becomes a State Party to the Rome Statute today, the world is also a step closer to ending a long era of impunity and injustice," he said, according to an ICC news release. "Indeed, today brings us closer to our shared goals of justice and peace." Judge Kuniko Ozaki, a vice president of the ICC, said acceding to the treaty was just the first step for the Palestinians. "As the Rome Statute today enters into force for the State of Palestine, Palestine acquires all the rights as well as responsibilities that come with being a State Party to the Statute. These are substantive commitments, which cannot be taken lightly," she said. Rights group Human Rights Watch welcomed the development. "Governments seeking to penalize Palestine for joining the ICC should immediately end their pressure, and countries that support universal acceptance of the court's treaty should speak out to welcome its membership," said Balkees Jarrah, international justice counsel for the group. "What's objectionable is the attempts to undermine international justice, not Palestine's decision to join a treaty to which over 100 countries around the world are members." In January, when the preliminary ICC examination was opened, Israeli Prime Minister Benjamin Netanyahu described it as an outrage, saying the court was overstepping its boundaries. The United States also said it "strongly" disagreed with the court's decision.

(b) Visualization of token probabilities estimated by LLaMA2-7b.

The Palestinian Authority officially became the 123rd member of the International Criminal Court on Wednesday, a step that gives the court jurisdiction over alleged crimes in Palestinian territories. The formal accession was marked with a ceremony at The Hague, in the Netherlands, where the court is based. The Palestinians signed the ICC's founding Rome Statute in January, when they also accepted its jurisdiction over alleged crimes committed "in the occupied Palestinian territory, including East Jerusalem, since June 13, 2014." Later that month, the ICC opened a preliminary examination into the situation in Palestinian territories, paving the way for possible war crimes investigations against Israelis. As members of the court, Palestinians may be subject to counter-charges as well. Israel and the United States, neither of which is an ICC member, opposed the Palestinians' efforts to join the body. But Palestinian Foreign Minister Riad al-Malki, speaking at Wednesday's ceremony, said it was a move toward greater justice. "As Palestine formally becomes a State Party to the Rome Statute today, the world is also a step closer to ending a long era of impunity and injustice," he said, according to an ICC news release. "Indeed, today brings us closer to our shared goals of justice and peace." Judge Kuniko Ozaki, a vice president of the ICC, said acceding to the treaty was just the first step for the Palestinians. "As the Rome Statute today enters into force for the State of Palestine, Palestine acquires all the rights as well as responsibilities that come with being a State Party to the Statute. These are substantive commitments, which cannot be taken lightly," she said. Rights group Human Rights Watch welcomed the development. "Governments seeking to penalize Palestine for joining the ICC should immediately end their pressure, and countries that support universal acceptance of the court's treaty should speak out to welcome its membership," said Balkees Jarrah, international justice counsel for the group. "What's objectionable is the attempts to undermine international justice, not Palestine's decision to join a treaty to which over 100 countries around the world are members." In January, when the preliminary ICC examination was opened, Israeli Prime Minister Benjamin Netanyahu described it as an outrage, saying the court was overstepping its boundaries. The United States also said it "strongly" disagreed with the court's decision.

(c) Visualization of token probabilities estimated by Mistral-7b.

Figure 11: Visualization of token probabilities estimated by LLaMA-7b, LLaMA2-7b and Mistral-7b on case 2. The color coding in the figure represents the token generation probabilities. A spectrum from red to green is utilized to represent a range from lower to higher generation probabilities, respectively.

Seventy years ago, Anne Frank died of typhus in a Nazi concentration camp at the age of 15. Just two weeks after her supposed death on March
31, 1945, the Bergen-Belsen concentration camp where she had been imprisoned was liberated -- timing that showed how close the Jewish diar
ist had been to surviving the Holocaust. But new research released by the Anne Frank House shows that Anne and her older sister, Margot Frank,
died at least a month earlier than previously thought. Researchers re-examined archives of the Red Cross, the International Training Service and
the Bergen-Belsen Memorial, along with testimonies of survivors. They concluded that Anne and Margot probably did not survive to March 1945
-- contradicting the date of death which had previously been determined by Dutch authorities. In 1944, Anne and seven others hiding in the
Amsterdam secret annex were arrested and sent to the Auschwitz-Birkenau concentration camp. Anne Frank's final entry. That same year, Anne
and Margot were separated from their mother and sent away to work as slave labor at the Bergen-Belsen camp in Germany. Days at the camp
were filled with terror and dread, witnesses said. The sisters stayed in a section of the overcrowded camp with no lighting, little water and no latr
ine. They slept on lice-ridden straw and violent storms shredded the tents, according to the researchers. Like the other prisoners, the sisters end
ured long hours at roll call. Her classmate, Nannette Blitz, recalled seeing Anne there in December 1944: "She was no more than a skeleton by
then. She was wrapped in a blanket; she couldn't bear to wear her clothes anymore because they were crawling with lice." Listen to Anne Frank's
friends describe her concentration camp experience. As the Russians advanced further, the Bergen-Belsen concentration camp became even
more crowded, bringing more disease. A deadly typhus outbreak caused thousands to die each day. Typhus is an infectious disease caused by l
ice that breaks out in places with poor hygiene. The disease causes high fever, chills and skin eruptions. "Because of the lice infesting the bedst
raw and her clothes, Anne was exposed to the main carrier of epidemic typhus for an extended period," museum researchers wrote.

(a) Visualization of token probabilities estimated by LLaMA-7b.

Seventy years ago, Anne Frank died of typhus in a Nazi concentration camp at the age of 15. Just two weeks after her supposed death on March
31, 1945, the Bergen-Belsen concentration camp where she had been imprisoned was liberated -- timing that showed how close the Jewish diar
ist had been to surviving the Holocaust. But new research released by the Anne Frank House shows that Anne and her older sister, Margot Frank,
died at least a month earlier than previously thought. Researchers re-examined archives of the Red Cross, the International Training Service and
the Bergen-Belsen Memorial, along with testimonies of survivors. They concluded that Anne and Margot probably did not survive to March 1945
-- contradicting the date of death which had previously been determined by Dutch authorities. In 1944, Anne and seven others hiding in the
Amsterdam secret annex were arrested and sent to the Auschwitz-Birkenau concentration camp. Anne Frank's final entry. That same year, Anne
and Margot were separated from their mother and sent away to work as slave labor at the Bergen-Belsen camp in Germany. Days at the camp
were filled with terror and dread, witnesses said. The sisters stayed in a section of the overcrowded camp with no lighting, little water and no latr
ine. They slept on lice-ridden straw and violent storms shredded the tents, according to the researchers. Like the other prisoners, the sisters end
ured long hours at roll call. Her classmate, Nannette Blitz, recalled seeing Anne there in December 1944: "She was no more than a skeleton by
then. She was wrapped in a blanket; she couldn't bear to wear her clothes anymore because they were crawling with lice." Listen to Anne Frank's
friends describe her concentration camp experience. As the Russians advanced further, the Bergen-Belsen concentration camp became even
more crowded, bringing more disease. A deadly typhus outbreak caused thousands to die each day. Typhus is an infectious disease caused by l
ice that breaks out in places with poor hygiene. The disease causes high fever, chills and skin eruptions. "Because of the lice infesting the bedst
raw and her clothes, Anne was exposed to the main carrier of epidemic typhus for an extended period," museum researchers wrote.

(b) Visualization of token probabilities estimated by LLaMA2-7b.

Seventy years ago, Anne Frank died of typhus in a Nazi concentration camp at the age of 15. Just two weeks after her supposed death on March
31, 1945, the Bergen-Belsen concentration camp where she had been imprisoned was liberated -- timing that showed how close the Jewish diar
ist had been to surviving the Holocaust. But new research released by the Anne Frank House shows that Anne and her older sister, Margot Frank,
died at least a month earlier than previously thought. Researchers re-examined archives of the Red Cross, the International Training Service and
the Bergen-Belsen Memorial, along with testimonies of survivors. They concluded that Anne and Margot probably did not survive to March 1945
-- contradicting the date of death which had previously been determined by Dutch authorities. In 1944, Anne and seven others hiding in the
Amsterdam secret annex were arrested and sent to the Auschwitz-Birkenau concentration camp. Anne Frank's final entry. That same year, Anne
and Margot were separated from their mother and sent away to work as slave labor at the Bergen-Belsen camp in Germany. Days at the camp
were filled with terror and dread, witnesses said. The sisters stayed in a section of the overcrowded camp with no lighting, little water and no latr
ine. They slept on lice-ridden straw and violent storms shredded the tents, according to the researchers. Like the other prisoners, the sisters end
ured long hours at roll call. Her classmate, Nannette Blitz, recalled seeing Anne there in December 1944: "She was no more than a skeleton by
then. She was wrapped in a blanket; she couldn't bear to wear her clothes anymore because they were crawling with lice." Listen to Anne Frank's
friends describe her concentration camp experience. As the Russians advanced further, the Bergen-Belsen concentration camp became even
more crowded, bringing more disease. A deadly typhus outbreak caused thousands to die each day. Typhus is an infectious disease caused by l
ice that breaks out in places with poor hygiene. The disease causes high fever, chills and skin eruptions. "Because of the lice infesting the bedst
raw and her clothes, Anne was exposed to the main carrier of epidemic typhus for an extended period," museum researchers wrote.

(c) Visualization of token probabilities estimated by Mistral-7b.

Figure 12: Visualization of token probabilities estimated by LLaMA-7b, LLaMA2-7b and Mistral-7b on case 3. The color coding in the figure represents the token generation probabilities. Tokens with lower probabilities are colored more red, while tokens with higher probabilities are colored more green.