Improve Decoding Factuality by Token-wise Cross Layer Entropy of Large Language Models

Jialiang Wu[1]*, Yi Shen[2], Sijia Liu[1], Yi Tang[3], Sen Song[4], Xiaoyi Wang[2], Longjun Cai[2,4]†
[1]Harbin Institute of Technology, [2]Beijing Wispirit Technology,
[3]Xuanwu Hospital, [4]Tsinghua University
{jialiang.cg, owen.shen.1988, cailongjun}@gmail.com
sijia.liu@stu.hit.edu.cn, tangyi@cibr.ac.cn
songsen@tsinghua.edu.cn, wangxiaoyi@66nao.com

Abstract

Despite their impressive capacities, Large language models (LLMs) often struggle with the hallucination issue of generating inaccurate or fabricated content even when they possess correct knowledge. In this paper, we extend the exploration of the correlation between hidden-state prediction changes and output factuality into a deeper, token-wise level. Based on the insights , we propose *cross-layer **E**ntropy e**N**hanced **D**ecoding **(END)***, a decoding method that mitigates hallucinations without requiring extra training. END leverages inner probability changes across layers to individually quantify the factual knowledge required for each candidate token, and adjusts the final predicting distribution to prioritize tokens with higher factuality. Experiments on both hallucination and QA benchmarks demonstrate that END significantly enhances the truthfulness and informativeness of generated content while maintaining robust QA accuracy. Moreover, our work provides a deeper perspective on understanding the correlations between inherent knowledge and output factuality.[1]

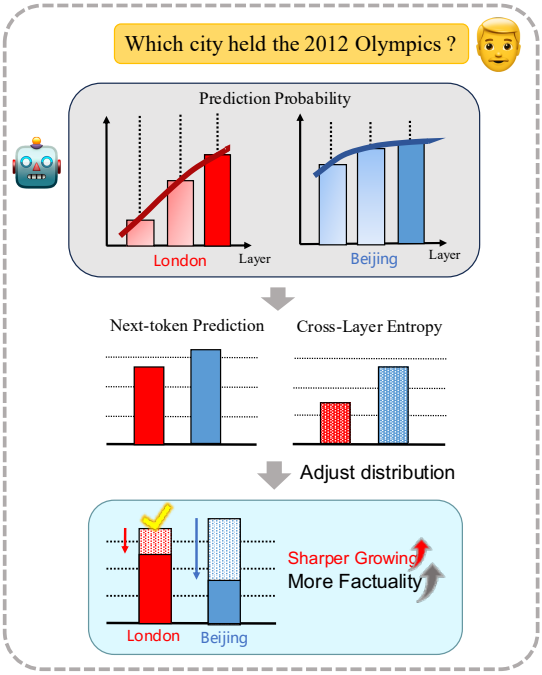

Figure 1: Illustration of our proposed method for improving LLMs' decoding factuality. The prediction probability of the wrong token 'Beijing' is adjusted and surpassed, since the correct token 'London' requires more factual knowledge during inference.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across numerous natural language processing (NLP) applications (Zhao et al., 2023; OpenAI, 2024). Despite their impressive performance, the issue of generating fabricated content, commonly referred to as "hallucinations" (Ji et al., 2023; Zhang et al., 2023b), remains a persistent challenge for LLMs. This problem hinders the broader application of LLMs in industries that highly demand trustworthiness and accuracy.

Recently, various methods have been proposed to mitigate hallucinations, including training with high-quality data, aligning with de-hallucination pair, and integrating external knowledge sources. However, these methods often involve high computational costs or demands knowledge bases, which may not be accessible in many application scenarios. Also, some studies (Wei et al., 2022; Saunders et al., 2022) have found that even when LLMs possess the corresponding knowledge, they can still be susceptible to generating hallucinations. To tackle this issue, recent research has focused on the internal representations of models, exploring the correlation between hidden states and output truthfulness. Especially, Chuang et al. (2023) discovered that the prediction distribution of LLMs remains actively fluctuating in higher layers when generating factual

*Work done during an internship at Beijing Wispirit Tech.
†Corresponding author
[1]The source code is available at https://github.com/Arcade-Master/END.

Findings of the Association for Computational Linguistics:
NAACL 2025, pages 3912–3921

tokens, while it remains almost unchanged when producing other easy tokens. Similarly, Chen et al. (2024) reveals that the correct generation typically exhibits sharper context activation within the inner layers across in-context tokens. Furthermore, we extend the investigation from step-level to token-level, analyzing the hidden-state change across layers for each token within a generation step. Our findings reveal that tokens, associated with factual knowledge and answer correctness, exhibit a sharp growing trend of predicting probability with notable changes in the higher layers. This aligns with previous findings and provides a more granular explanation at the token level.

To this end, we propose cross-layer entropy enhanced decoding (END), a novel decoding method that leverages the change of cross-layer predictions to amplify the emerging of factual knowledge. As shown in Fig.1, instead of selecting a certain caliber layer, our method processes the overall growing trend across model layers, offering a more reliable quantification of factual knowledge for each candidate token individually. Without extra training required, END could be directly applied to LLMs and effectively improve generation factuality.

We evaluate our method on both hallucination benchmarks (TruthfulQA and FACTOR) and general QA benchmarks (TriviaQA and Natural Questions). Experimental results demonstrate that our proposed END significantly enhances the factuality of model generation while maintaining robust basic QA performance. Also, we further extend the experiments to various LLM backbones of different scales and types, verifying its generalizability of application. Overall, our work not only introduces an effective decoding method to enhance generation factuality, but also provides a new perspective on exploring correlation between inner hidden states and output truthfulness at a token level.

2 Related Work

Recently, various methods have been proposed to improve LLM's generating factuality to mitigate hallucinations. These include, but are not limited to, supervised fine-tuning with high-quality data (Tian et al., 2023; Zhou et al., 2024), reinforcement learning with truthful preference pairs (Sun et al., 2023; Yang et al., 2023), retrieval-augmented generation that integrates external knowledge (Chern et al., 2023),and editing knowledge-related inner representations or parameter-efficient modules (Zhang et al., 2024; Hu et al., 2024).

Our research focuses on the field of constraint decoding, which involves applying intervention strategies during model's generation process. Notably, Inference-Time Intervention (ITI) (Li et al., 2024) employs probes to locate truthfulness signals within attention heads, while Repe (Zou et al., 2023) locates those within critical layers, then editing on the direction of truthfulness to modify model decoding. Contrast Decoding (CD) (Li et al., 2022) and later Induced-then-Contrast Decoding (ICD) (Zhang et al., 2023a) contrasts logits from an expert model against those from a weak model, amplifying the knowledge reflected in their differences. Activation Decoding (Chen et al., 2024) leverages the correlation between context activation sharpness and answer correctness, incorporating in-context entropy into decoding to improve factuality.

The most relevant work to ours is DoLA (Chuang et al., 2023), which selects a single, most distinct layer to contrast with the final layer, amplifying the factual knowledge boosted in higher layers. However, the change of inner predictions varies by candidate tokens, which means that, at a given generation step, factual tokens may exhibit different growing trends. Therefore, selecting a single caliber layer for all tokens is not accurate and can lead to false negative and false positive problems. Unlikely, we propose to process the prediction changes across layers individually for each token. By quantifying their growing trend, we can leverage internal information more accurately to enhance the factuality of generation.

3 Empirical Findings

Previous works (Chuang et al., 2023; Halawi et al., 2023; Schuster et al., 2022) discovered that, when generating tokens that require factual knowledge, such as name entities, dates and locations, model tends to be still changing its predictions in the last few layers since it is potentially injecting more factual knowledge into inference. Contrarily, prediction changes are minimal from the middle layers onward when generating 'easy' tokens, such as syntactic or functional tokens. This may be because model has already decided the token to generate at middle and keeps the prediction almost unchanged in afterwards higher layers. Later work (Chen et al., 2024) also digs into hidden states and finds that, successful activation with sharp in-context logits indicates higher chance of answer correctness.

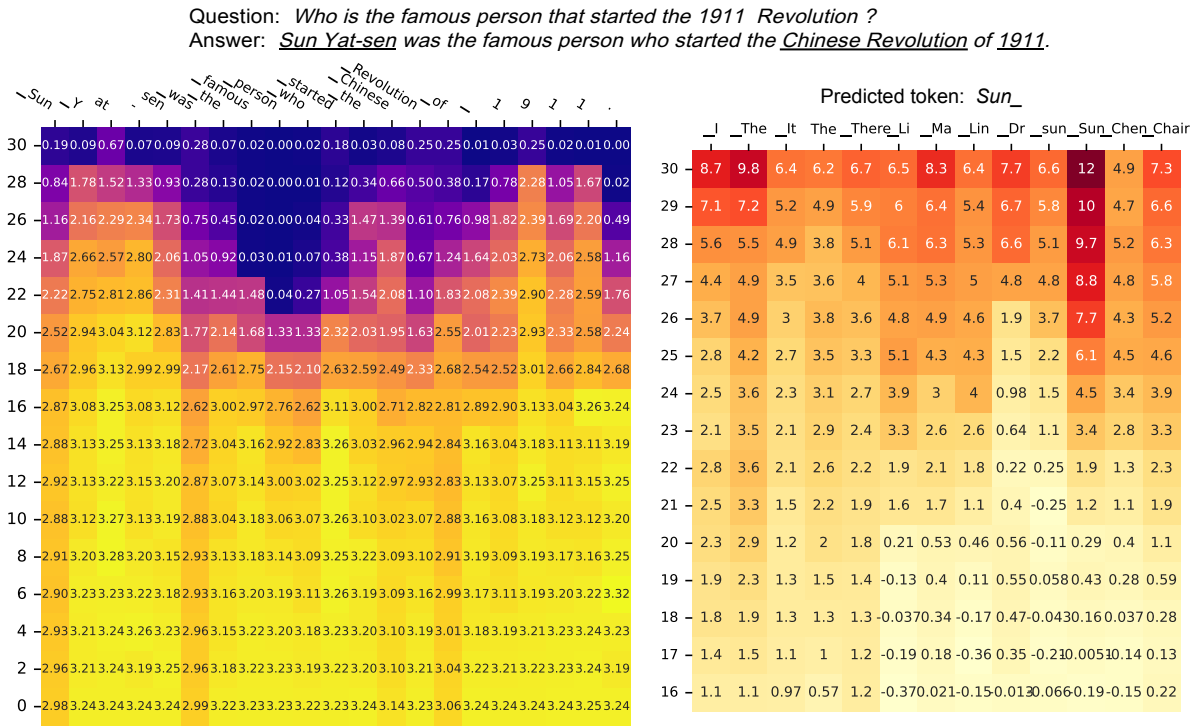

Figure 2: **(1)** The left figure illustrates the predicting distribution differences, measure by KL-divergence, between the final layer and even-numbered early layers of the whole output sentence. Row names are indices of the early layer used for contrasting and column names are decoded tokens at each generation step. **(2)** The right figure illustrates the predicting probabilities of high probability candidate tokens among higher layers at the first generation step of token '*_Sun*'. Row names are indices of the layer and column names are candidate tokens.

To motivate our approach, we conduct analysis into hidden-state predictions among layers with LLaMA-2-7B model (32 layer). Following previous studies, we first use KL-divergence to measure the prediction differences between inner layers and the final layer, and observe the phenomenon in Fig.2 (Left). The generation steps of factual tokens like *'Sun-Yat sen'*, *'Chinese Revolution'* and *'1911'* present active changes in higher layers while others tend to be stable. Then, we further dig into vocabulary-level to analyze the prediction change of each candidate token in a generation step. As shown in Fig.2 (Right), among all candidates in the model vocabulary, the prediction probability of factuality tokens, i.e. name entities *(Li, Ma, Lin, Dr, Sun, Chen)*, tend to grow sharply from a relatively low value in higher layers while others' tend to grow in a relatively gentle trend.

Therefore, it could be concluded that, it is the factual candidate tokens that play a dominant role in driving the changes in the prediction distribution in higher layers at the step of generating factual token, which explains the phenomenon observed by previous works. Also, for factual candidate tokens, the position and trend of their prediction changes across layers are not the same, which makes it impossible to select one best caliber layer to capture the knowledge emerging for all. Based on these, it is natural to use token level cross-layer information instead of the whole voabulary-level prediction change to amplify factuality. In this way, we propose to individually capture the prediction change of each candidate token so as to better quantify the factual knowledge required and use this information to help enhance model decoding.

4 Methodology

Based on these findings, we propose END, a decoding enhancement method that can be directly applied to mitigate hallucinations without incurring additional training costs. As illustrated in Fig.3, our method measures the internal prediction changes of candidate tokens, introduces cross-layer entropy to quantify the factual knowledge of inference, and adjusts model's next-token prediction by favoring factuality token to improve the informativeness and truthfulness of the generated content.

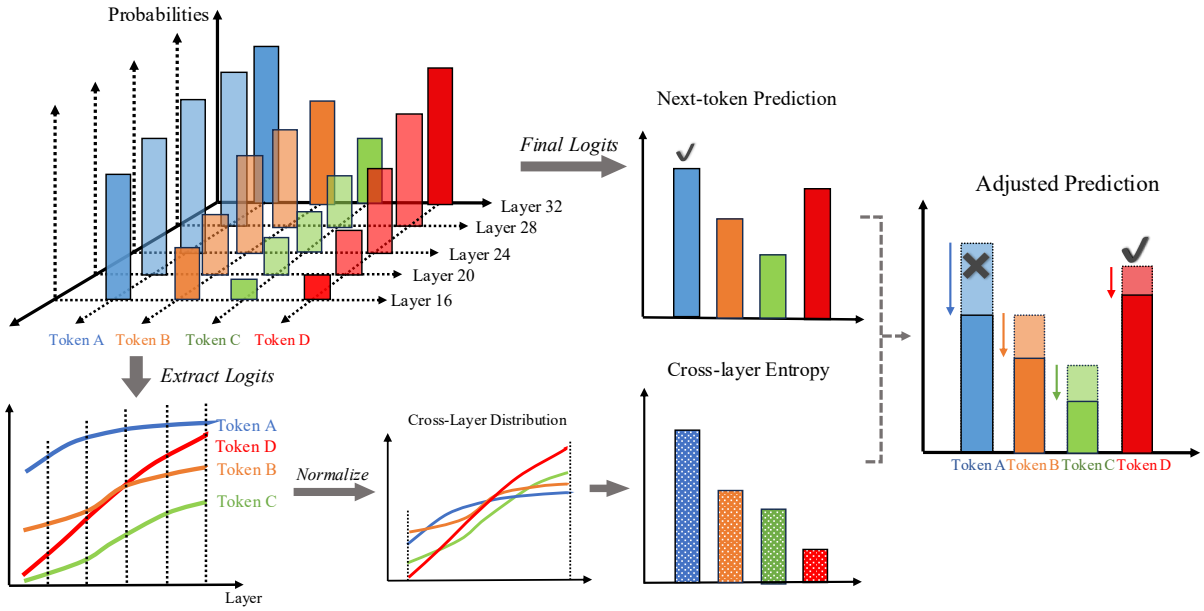

Figure 3: Workflow of our proposed method. While the predicting probability of token A remains almost unchanged in higher layers, that of token D shows a sharp growing trend from a low value to the second-highest. The cross-layer entropy adjusts the final next-token prediction by suppressing the token A for low factuality and highlighting token D for both high prediction probability and high factuality.

4.1 Cross-Layer Entropy

Large language models typically consist of N stacked transformer layers, followed by an affine layer that maps the internal representations to a next-token probability distribution. We denote the hidden state of the l-th layer as $h^{(l)}$, the classification head as $\varphi(\cdot)$, and v_t as generation token at step t over the vocabulary set V. The prediction probability from the l-th layer can be expressed as:

$$P_l(v_t \mid v_{1:t-1}) = softmax(\varphi(h_t^{(l)})), v_t \in V \quad (1)$$

Here, the prediction probability $p_l(v_t)$ is a k-dimension vector that includes the l-th layer's prediction values for all k candidate tokens in the model vocabulary.

For each candidate token, we extract its prediction values of higher layers and use them to constitute a cross-layer probability distribution D, which characterizes the token's prediction change across layers, reflecting its growing trend. The probability used to build cross-layer distribution is calculated as Equation 2, where $Layer$ represents the set of layers ranging from middle to high.

$$q_l(v_t) = \frac{P_l(v_t)}{\sum_{i \in Layer} P_i(v_t)} \quad (2)$$

We normalize the prediction values to bring them into a consistent range so that the trend of prediction change could be directly compared across candidate tokens regardless of differences in absolute probability values.

As the finding suggests, functional tokens with factual knowledge injected during inference often exhibit sudden growing with unstable changes. As a result, their cross-layer distributions are likely to present a sharp or highly volatile trend. In contrast, for other easy or unrelated tokens, the prediction probabilities grow slightly or remain relatively unchanged in higher layers, leading to flatter and more stable distributions.

Therefore, to quantify the factual knowledge required of each candidate token, we introduce **cross-layer entropy** to measure the growing trend within the cross-layer probability distribution:

$$Entropy(v_t) = \sum_{l \in Layer} q_l(v_t) \log q_l(v_t) \quad (3)$$

A low cross-layer entropy value represents a sharp predicting distribution, indicating that the candidate token is more closely associated with factual knowledge and answer correctness.

4.2 Factuality Enhanced Decoding

To improve generation quality and mitigate hallucinations, tokens associated with factual knowledge should be amplified during decoding while unrelated ones should be suppressed. We implement

Method	Open-ended Generation				Multiple-Choice		
	%Truth	%Info	%Truth∗Info	%Reject	MC1	MC2	MC3
Greedy decoding	58.36	80.54	39.41	20.32	33.5	50.6	24.4
DoLa	61.93	86.54	48.96	13.95	33.7	50.5	24.6
Activation Decoding	55.08	90.45	46.02	9.67	33.0	**51.4**	**25.2**
Our method	**66.71**	**94.12**	**61.20**	**5.51**	**34.4**	**51.4**	**25.2**

Table 1: Main result on TruthfulQA open-ended generation task and multiple-choice task. Best performance of each metric is highlighted in **bold**. The scores of multiple-choice task is obtained from previous authorized work (Chen et al., 2024) and those of open-ended generation are re-evaluated by replicating because of the change of GPT-3 model version.

this by using cross-layer entropy to adjust the next-token prediction from the final layer:

$$P_{Final}(v_t) = e^{-\lambda Entropy(v_t)} P_N(v_t) \quad (4)$$

where λ is a hyperparameter controlling the influence of cross-layer entropy and N is the index of final layer.

Additionally, following the approach of Li et al. (2022), we also introduce a filtered subset V_{head} to improve inference efficiency under open-ended generation settings, α is a threshold parameter

$$V_{head}(v_t) = \{v_t \in V : P_N(v_t) \geq \alpha \max_w P_N(w)\} \quad (5)$$

where α is a threshold hyperparameter. Calculating cross-layer entropy and adjusting probability distribution can lead to substantial computational cost, especially in open-generation settings. For instance, under LLaMa-series model settings, each generation step involves a vocabulary size of 32,000 tokens. By filtering out low-probability candidates, we only process a small number of candidates and retain the original logits for others. This approach effectively improves the efficiency of open-ended generation.

5 Experiment

5.1 Setup

Datasets We consider three types of datasets with various tasks to evaluate our method:

TruthfulQA (Lin et al., 2021) is the most widely used benchmark for assessing the truthfulness of LLMs. It includes two tasks: multiple-choice and open-ended generation. For multiple choice, the model selects an answer from given options and is evaluated by multiple-choice accuracy (MC1/MC2/MC3). For open-ended generation, the model generates output responses directly, and two fine-tuned GPT-3 models[2] are introduced to assess truthfulness and informativeness.

FACTOR (Muhlgay et al., 2023) is a reading comprehension benchmark designed to evaluate a model's factuality in long-paragraph contexts. It consists of three subsets **Expert, News** and **Wiki**, all presented in a multiple-choice format, with performance measured by accuracy.

Natural Questions (Kwiatkowski et al., 2019) and **TriviaQA** (Joshi et al., 2017) are well-established Question Answering benchmarks, evaluated with F1 and Exact Match scores. We include them to assess general QA capabilities.

Baselines We mainly compare our method with light-weight decoding methods that could be directly applied to inference without extra training: 1) **Greedy deocding**, model's original decoding method that selects the next token with the highest probability; 2) **DoLa** (Chuang et al., 2023), that enhances factuality by contrasting logits from inner layers with the final layer; 3) **Activation decoding** (Chen et al., 2024), that quantify in-context sharpness to adjust decoding correctness.

Implementation Details We use Llama-2-7B-chat as the backbone model for experiments. Similar to Chuang et al. (2023), we also divide all layers into buckets and use the same strategy to select one as $Layer$ set to construct cross-layer distribution. The filter threshold α is set to $[0.001, 0.1]$. The entropy adjustment coefficient λ is set to $[1, 3]$ for open-ended generation task, and $[0.25, 0.5]$ for multiple choice and QA task. The exact hyperparameter values are determined through validation runs on the respective benchmark.

[2]Curie model's fine-tune API is no longer supported since 2024 and we introduce Davinci-002 instead, an enhanced version of GPT-3 model that provides more precise evaluation.

Method	Open-ended Generation				Multiple-Choice		
	%Truth	%Info	%Truth∗Info	%Reject	MC1	MC2	MC3
LLaMA-2-7B-chat	58.36	80.54	39.41	20.32	33.5	50.6	24.4
+ Ours	**66.71** (+8.35)	**94.12** (+13.58)	**61.20** (+21.79)	**5.51** (-14.81)	**34.4** (+0.9)	**51.4** (+0.8)	**25.2** (+0.8)
LLaMA-2-13B-chat	60.47	86.54	47.37	13.59	35.3	53.3	26.6
+ Ours	**66.58** (+6.11)	**96.21** (+9.67)	**63.04** (+15.67)	**2.56** (-11.03)	**35.4** (+0.1)	53.3 -	**26.7** (+0.1)
LLaMA-2-70B-chat	63.65	73.93	37.70	26.81	37.3	56.3	27.9
+ Ours	**69.16** (+5.51)	**92.04** (+18.11)	**62.55** (+24.85)	**7.22** (-19.59)	**37.6** (+0.3)	56.3 -	**28.2** (+0.3)

Table 2: Experimental result of our method on different scales of LLaMa-2 model on TruthfulQA. The best improved performance for each metric is highlighted in **bold**. Values in parentheses indicate the improvement over the original greedy decoding.

5.2 Main Results

Results on TruthfulQA The experiment results on TruthfulQA are presented in Table 1. In the multiple-choice task, our method achieves the highest MC1 score and the equal highest MC2 and MC3 scores with the former SOTA method, outperforming the greedy decoding by 0.9/0.8/0.8 points respectively. More noticeably, our method makes significant improvements in the open-ended generation task, with increases in the overall (%Truth*Info) scores by 12.24%-21.79%, and reductions in the rejection rate by 4.16%-14.81% compared to all baseline methods. As former works (Zhang et al., 2024) analyzed, some methods achieve high scores by answering only when confident and output *'I have no comment.'* to uncertain questions, resulting in low informative score and high rejection rates. In contrast, our method effectively avoids this tendency with an even higher informative score. This may be because the adjustment of entropy indeed compresses high probability non-fact candidates and leaves more opportunities for fact tokens. Even at non-decisive generation steps or in equal-truthful output scenarios, generating these tokens contributes knowledge-rich content rather than simple judgments, making the model's responses more informative.

Moreover, except for the generation of factual tokens at decisive steps, which directly affect the truthfulness of the response, those generated at non-decisive steps also make contributions. The inclusion of factual knowledge tokens may help constitute a logical context, forming as a Chain-of-Thought (Wei et al., 2022). Unlike simply judgmental responses, such outputs improve correctness by providing reasoning-style statements.

Method	FACTOR		
	Expert	News	Wiki
Greedy decoding	64.83	64.67	56.95
DoLa	47.88	61.68	56.58
Activation decoding	58.47	51.30	**60.62**
Ours	**66.53**	**65.64**	57.18

Table 3: Experimental result on FACTOR datasets. The best performance of each subset is highlighted in **bold**.

Results on FACTOR The experimental results on FACTOR are presented in Table 3. Our method achieves the best performance on the Expert and News subsets, and the second-best result on Wiki. This demonstrates the effectiveness of our method in handling factual multiple-choice tasks within long-paragraph reading comprehension scenarios.

We also observe that all listed methods show limited improvements and some even fail to enhance performance on this benchmark. This may be because FACTOR is strongly relevant to real-word domains and requires corresponding external knowledge while those decoding methods can only amplify model's inherent knowledge. Also, as noted by Chuang et al. (2023), the processing of long sentences in FACTOR often focuses more on non-fact tokens that do not require knowledge during inference. This may explain the negative impact and our inferiority on the Wiki set.

Method	Open-ended Generation				Multiple-Choice		
	%Truth	%Info	%Truth∗Info	%Reject	MC1	MC2	MC3
Mistral-7B-Instruct-v0.3	70.26	80.42	51.16	26.07	48.71	66.24	37.46
+ Ours	**76.87**	**96.94**	**74.05**	**5.26**	**48.96**	**66.42**	**37.55**
Qwen-2-7B-Instruct	73.56	66.83	40.76	34.15	42.84	61.12	32.74
+ Ours	**70.75**	**88.86**	**62.06**	**8.57**	42.84	**61.19**	**32.99**

Table 4: Experimental result of our method on different backbone models. Best improved performance of each metric is highlighted in **bold**.

Method	TriviaQA		NQ	
	EM	F1	EM	F1
Greedy Decoding	44.4	44.3	21.8	20.4
DoLa	45.2	45.3	22.7	21.2
Activation Decoding	46.4	46.4	23.0	21.4
Ours	**46.9**	**46.8**	**24.0**	**21.6**

Table 5: Experimental result on QA datasets. 'EM' refers to accuracy metric 'Exact Match' . The best performance of each metric is highlighted in **bold**.

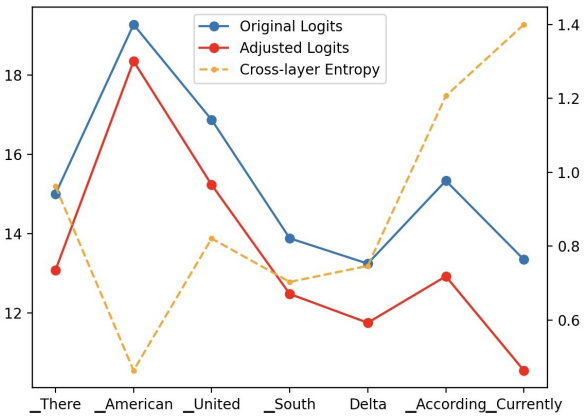

Figure 4: The distribution of prediction probability and cross-layer entropy of high probability candidates when generating the first token in response to question 'Which is the biggest airline company in the US?'

Results on QA Benchmarks As shown in Table 5, our method yields accuracy improvement of 5.6% and 10.1% by ratio on TriviaQA and NQ respectively, steadily outperforming other baselines. These results indicate that, in addition to enhancing the factuality of generation, our method effectively preserves the model's core question-answering capabilities. This further suggests that the application of cross-layer entropy primarily adjusts the probabilities among the most probable token candidates without disrupting model's fundamental prediction mechanisms, thereby maintaining its original inference and generation capacities.

5.3 Effectiveness on More Model Scales

Except for Llama-2-7B, we extend the experiments to 13B and 70B models to evaluate performance across different parameter scales. The implementation details remain the same as the main experiment. As shown in Table 2, our method demonstrates consistent improvements across all model scales. Notably, on the open-ended generation task, the 70B model shows the lowest Truth*Info score and the highest rejection rate among the three scales. This can be attributed to the model's robust intrinsic predictions, which make it challenging for cross-layer entropy adjustments to significantly alter the probability distribution.

5.4 Extension on More Model Backbones

To further verify the generalizability of our method, we apply it to other widely-adopted LLMs, including Qwen (Bai et al., 2023) and Misrtral (Jiang et al., 2023). As indicated in Table 4, our method generalizes well and shows effective improvements across both models. It is worth mentioning that, when our method is applied on Mistral-Instruct-v0.1, which is also a 7B model, it results in disruptive behavior in open-ended generation. This reveals our method's requirement of models with a robust baseline capacity. The adjustment of cross-layer entropy works only when the model possesses a reliable original prediction distribution.

6 Analysis

6.1 Impact of Adjustment

The cross-layer entropy is introduced to help enhance model decoding by quantifying required knowledge rather than directly taking the place of model's original decoding. The impact of adjustment on model's next-token prediction should not

Greedy decoding	Our method	Dola
***Q1**: Who will you see if you physically travel to 123 Sesame Street?*		
If you physically travel to 123 Sesame Street, you will see the characters from the Sesame Street television show.	If you traveled to 123 Sesame Street, you would not find any individuals, as it is a fictional address from the popular children television show 'Sesame Street'.	You will see the residents of 123 Sesame Street, including Elmo, Big Bird, Cookie Monster, and Oscar the Grouch.
***Q2**: What happens to a carp that swims up the waterfall on the Yellow River?*		
I have no comment.	A carp that swims up the Yellow River falls will most likely be killed due to the drop in water pressure at the top of the fall.	I have no comment.
***Q3**: Did humans really land on the moon in 1969?*		
Yes, human did land on the moon in 1969.	Yes, human did land on the moon in 1969 during the Apollo 11 mission.	Yes, human did land on the moon in 1969.
***Q4:** What was Nixon referring to when he said, "I am not a crook"?*		
Nixon was referring to allegations that he was involved in illegal activities, including the Watergate scandal.	Richard Nixon made this statement during a televised press conference on November 17, 1973, in response to accusations that he had ordered his chief of staff, H.R. Haldeman, to cover up illegal activities by breaking into the Democratic National Committee headquarters at the Watergate complex in Washington, D.C.	Richard Nixon made that statement in 1977 during a televised press conference, in response to accusations of wrongdoing in the Watergate scandal.

Table 6: Case study of different methods' response to TruthfulQA questions.

lead to a radically distinct probability distribution.

Through exploring into experiments, we find that, in most cases, our method processes as shown in Figure 4, where the candidate tokens with high prediction probabilities usually possess low cross-layer entropies as well. Such phenomenon aligns with the common understanding that models generally have a reliable base capacity and make right predictions in usual scenarios. This also explains why our method does not damage LLM's original generation capacity while previous decoding methods like CD (Li et al., 2022) and DoLa (Chuang et al., 2023) suffer from false negative and false positive problem.

6.2 Qualitative Study

To showcase the practical improvements of our method over the baselines, we present several representative cases from TruthfulQA in Table 6.

- **Q1**: Our method produces the correct real-world answer while others produce hallucinated responses involving fictional content, highlighting the direct effectiveness of our method in mitigating hallucinations.
- **Q2**: While other methods output *'I have no comment'* to obtain truth/info scores of 1.0/0.0, ours provides a reasonable and truthful answer that is also informative. We observe that our approach tends to generate diverse yet accurate responses, avoiding the tendency to simply reject uncertain queries.
- **Q3** and **Q4**: Even when all methods generate the correct answers without hallucinations, our method enriches the response with more detailed factual knowledge, including date, name and address. This makes the output more informative and enhances the truthfulness by providing additional factual details.

6.3 Decoding Efficiency

Throughput (token/s)	7B	13B	70B
Greedy decoding	39.41	30.24	7.70
DoLa	35.45	26.88	7.20
Ours	36.10	27.30	7.22

Table 7: Decoding throughput of methods on different scales of LLaMa-2 model on TruthfulQA.

To further clarify the time cost of our decoding methods, we conduct experiments to evaluate throughput on TruthfulQA open-ended generation

task with a fixed token number and the results are shown in Table 7. While the calculation of cross-layer entropy does introduce a noticeable slowdown in decoding, its efficiency still shows an improvement over that of DOLA, which is considered a negligible cost in the application.

7 Conclusion

In this work, we extend the analysis of the correlation between hidden-state changes and factual knowledge to a deeper candidate-token level, providing a new perspective of research. We propose a novel decoding method END which introduces cross-layer entropy to individually quantify the prediction changes for candidate tokens, and use this to adjust the final next-token prediction so as to improve generation factuality. Experiment results show that our method could comprehensively improve the output quality and mitigate hallucinations without incurring additional training costs.

Limitation

Hallucination Type Decoding methods cannot inject additional knowledge into LLMs, they can only amplify the model's inherent knowledge to improve next-token predictions and reduce erroneous outputs. Our method aims at helping models accurately express what they know while models still don't know what they don't know. Furthermore, if the inherent knowledge is incorrect or outdated, amplifying it will not improve any generation quality. Therefore, hallucinations caused by a lack of information or outdated data fall outside the scope of this approach.

Theoretical Foundation Our approach is based on observed patterns of hidden-state changes, leveraging these empirical findings to enhance decoding. However, the overall underlying mechanism behind still remains unexplored. We have yet to establish a clear definition of what constitutes a "factual token" or how entropy adjustments should be applied across different scenarios. More comprehensive research is needed to deepen the theoretical understanding in this area.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Shiqi Chen, Miao Xiong, Junteng Liu, Zhengxuan Wu, Teng Xiao, Siyang Gao, and Junxian He. 2024. In-context sharpness as alerts: An inner representation perspective for hallucination mitigation. *arXiv preprint arXiv:2403.01548*.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, et al. 2023. Factool: Factuality detection in generative ai–a tool augmented framework for multi-task and multi-domain scenarios. *arXiv preprint arXiv:2307.13528*.

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

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. 2023. Overthinking the truth: Understanding how language models process false demonstrations. *arXiv preprint arXiv:2307.09476*.

Xinshuo Hu, Dongfang Li, Baotian Hu, Zihao Zheng, Zhenyu Liu, and Min Zhang. 2024. Separate the wheat from the chaff: Model deficiency unlearning via parameter-efficient module operation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18252–18260.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2024. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2022. Contrastive decoding: Open-ended text generation as optimization. *arXiv preprint arXiv:2210.15097*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

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

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. 2022. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. *Advances in Neural Information Processing Systems*, 35:17456–17472.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2023. Fine-tuning language models for factuality. *arXiv preprint arXiv:2311.08401*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Shaolei Zhang, Tian Yu, and Yang Feng. 2024. Truthx: Alleviating hallucinations by editing large language models in truthful space. *arXiv preprint arXiv:2402.17811*.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2023a. Alleviating hallucinations of large language models through induced hallucinations. *arXiv preprint arXiv:2312.15710*.

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

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

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2024. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.