Attention-Guided Layer Selection for Contrastive Decoding in Large Language Models

Yusuke Sakai, Natthawut Kertkeidkachorn, Kiyoaki Shirai
Japan Advanced Institute of Science and Technology
{yusuke.sakai,natt,kshirai}@jaist.ac.jp

Abstract

Contrastive decoding methods such as DoLa improve the factuality of Large Language Models (LLMs) by contrasting the output distributions of mature and premature layers. However, DoLa's dynamic layer selection relies solely on divergences in output vocabulary distributions. In this work, we propose three attention-guided strategies—Attention-JSD, Attention-Entropy-Max, and Attention-Entropy-Min—that leverage structural information carried by internal self-attention mechanisms as a signal for layer selection. Experimental results on TruthfulQA demonstrate that our strategies, particularly Attention-JSD and Attention-Entropy-Min, consistently outperform the original DoLa. We observe significant gains on multi-answer metrics (MC2 and MC3), suggesting that attention distributions can provide a more sensitive signal for resolving factual knowledge than output vocabulary distributions.

1 Introduction

Large language models (LLMs) have achieved strong performance across a wide range of natural language processing tasks, yet they remain prone to hallucination, i.e., generating plausible but factually incorrect content (Huang et al., 2025). This issue is particularly problematic in tasks such as open-domain question answering and long-form generation, where producing unsupported claims that are not grounded in the training data poses a major obstacle to real-world deployment.

One inference-time approach to mitigating this issue is contrastive decoding (Li et al., 2023b). Contrastive decoding is a framework that does not rely solely on the output distribution of a single model; instead, it contrasts multiple distributions to relatively suppress undesirable generation tendencies and to prefer more desirable candidates. For instance, Li et al. (Li et al., 2023b) define an objective based on the difference between the log-probabilities of a large model (expert) and a small model (amateur), and introduce a plausibility constraint based on the expert model's confidence, aiming to reduce degradations in open-ended generation such as repetition and topic drift while maintaining fluency and coherence.

DoLa (Decoding by Contrasting Layers) (Chuang et al., 2024b) extends this contrastive idea within a single model by contrasting the logits of the final (mature) layer and an early (premature) layer, thereby improving factuality without additional training. Given a set of candidate premature layers, DoLa dynamically selects, at each token, the layer that maximizes the Jensen–Shannon divergence (JSD) from the final layer, and demonstrates effectiveness on benchmarks such as TruthfulQA (Lin et al., 2022) and FACTOR (Muhlgay et al., 2024). Moreover, DoLa is self-contained within a single model, requiring neither an auxiliary model nor external signals, which makes it a practical plug-in method for inference-time adoption. However, DoLa's layer selection is driven by differences in output distributions (logits) and does not directly leverage information carried by the model's internal attention mechanisms as a selection signal.

Transformers provide diverse internal signals beyond output distributions, including self-attention mechanisms and intermediate representations. Recent work has explored leveraging such signals, for example, using cross-layer entropy changes of hidden states (Wu et al., 2025), adaptively tuning contrast strength in retrieval-augmented generation based on context (Kim et al., 2024; Shi et al., 2024), and detecting hallucinations from attention patterns (Chuang et al., 2024a). Nevertheless, contrastive decoding methods still often rely on scalar statistics derived from logits or hidden states, and systematic attempts to use the structural information in attention distributions as token-level signals for layer selection remain limited.

In this study, we hypothesize that attention mechanisms can serve as a signal for identifying layers

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 351–363
November 6-10, 2026 ©2026 Association for Computational Linguistics

where factual knowledge emerges. Specifically, while keeping the DoLa framework intact, we introduce three strategies for selecting premature layers based on internal attention distributions.

Our contributions are as follows:

- We propose three attention-guided strategies—Attention-JSD, Attention-Entropy-Max, and Attention-Entropy-Min—that leverage structural information derived from self-attention mechanisms as dynamic signals for layer selection, moving beyond the conventional reliance on output vocabulary distributions.
- TruthfulQA evaluations show that our proposed methods (especially Attention-JSD and Attention-Entropy-Min) outperform existing approaches, achieving substantial gains particularly on the multiple-correct-answer metrics (MC2/MC3). In addition, on FACTOR, our methods improve over the baseline in many settings and achieve performance comparable to, or in some cases better than, DoLa.
- Through visualization of layer selection patterns and head-wise analysis, we identify that specific attention heads carry distinct signals conducive to factuality discrimination, thereby revealing potential avenues for further optimization in contrastive decoding design.

2 Related Work

Hallucinations in Large Language Models. LLM hallucinations can be broadly categorized into factuality hallucinations, which generate content that contradicts verifiable facts, and faithfulness hallucinations, which deviate from the given context or instructions. Both forms pose serious challenges across a wide range of applications, including dialogue systems, retrieval-augmented generation (RAG), and agents (Ji et al., 2023; Huang et al., 2025). To mitigate this issue, various approaches have been proposed, such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), self-consistency checks at inference time (Wang et al., 2023; Manakul et al., 2023), and multi-agent debate (Du et al., 2024). These methods typically require additional training, access to external knowledge (Lewis et al., 2020), or multiple inference passes. In contrast, this study focuses on methods that do not modify the parameters of existing models and can be realized with a single forward computation.

Layer-wise Knowledge in Transformers. Transformer-based models have been shown in many analysis studies to encode information hierarchically across layers (Tenney et al., 2019; Jawahar et al., 2019). For example, in BERT, lower layers tend to capture surface-level features and local syntactic information, while semantic and contextual information becomes increasingly salient from the middle to higher layers (Jawahar et al., 2019). It has also been observed that information corresponding to different stages of language understanding emerges progressively along the depth of the network (Tenney et al., 2019). Building on these observations, prior work has proposed methods that identify which internal computations mediate factual associations and update knowledge by editing those localized computations (Meng et al., 2022, 2023). Moreover, as an inference-time attempt to directly intervene in internal representations that arise in layers and attention mechanisms to elicit desirable properties, Inference-Time Intervention (ITI) has been proposed, which improves truthfulness by manipulating the activations of specific attention heads (Li et al., 2023a).

Contrastive Decoding. Contrastive Decoding (CD) (Li et al., 2023b) is a decoding method that uses the log-probability gap between an expert and an amateur as a contrastive objective while restricting candidates with an expert-confidence-based plausibility constraint V_{head}; DoLa (Chuang et al., 2024b) extends this contrastive idea to inter-layer contrast within a single model by contrasting the distributions of the final (mature) layer and a shallow (premature) layer, dynamically selecting at each token the premature layer that maximizes the Jensen–Shannon divergence (JSD) from the final layer. Related work includes END (Wu et al., 2025), which computes token-wise cross-layer entropy from layer-wise changes in predicted probabilities for candidate tokens and reweights the final predictive distribution accordingly, and ACD (Kim et al., 2024), which adaptively adjusts contrast strength in RAG based on, e.g., the extent to which the retrieved context reduces uncertainty (entropy). In contrast, designs that directly exploit the structural patterns of attention distributions as dynamic signals remain limited.

Attention-based Analysis. Attention mechanisms assign weights to reference tokens when updating representations at each position, and have

been used as a cue for observing layer- and head-wise behaviors. Prior work has reported that certain attention heads capture syntactic dependencies and coreference relations (Clark et al., 2019; Voita et al., 2019). It has also been suggested that some heads are involved in factual knowledge recall and attribute extraction (Geva et al., 2023), and that attention patterns can be used to detect hallucinations during generation (Chuang et al., 2024a). These findings suggest that attention distributions may serve as a signal for layer selection. However, to the best of our knowledge, attempts to directly use attention distributions (or their summaries) for token-level layer selection in dynamic decoding have not been investigated. Starting from the DoLa framework, this study is characterized by a systematic evaluation of layer selection strategies based on attention distributions and their entropy.

3 Methods

In this section, we first review the background of DoLa and then introduce our attention-based layer selection strategies. Finally, we describe the head-level analysis setup.

3.1 Background: Decoding by Contrasting Layers

Given a prefix token sequence $x_{<t}$, a Transformer-based LLM induces a next-token conditional distribution $q_l(x_t \mid x_{<t})$ from the output of each layer l.

DoLa (Chuang et al., 2024b) is an inference-time decoding method that improves factuality without additional training by contrasting a shallow premature layer with a deep mature layer (the final layer L) within the same model. It defines $\hat{P}(x_t \mid x_{<t})$ using a score $F(q_L(x_t), q_{l^*}(x_t))$ based on the ratio between the final-layer distribution and the token-wise selected premature-layer distribution.

$$\hat{P}(x_t \mid x_{<t}) = \mathrm{softmax}\big(F(q_L(x_t), q_{l^*}(x_t))\big). \tag{1}$$

$$F(q_L(x_t), q_{l^*}(x_t)) = \begin{cases} \log \dfrac{q_L(x_t)}{q_{l^*}(x_t)}, & x_t \in V_{\text{head}}(x_{<t}), \\ -\infty, & \text{otherwise.} \end{cases} \tag{2}$$

To ensure plausibility under the final layer, candidates are restricted to the token set $V_{\text{head}}(x_{<t})$. Here, $V_{\text{head}}(x_{<t})$ is defined as follows, where w ranges over the vocabulary $\mathcal{V}$:

$$V_{\text{head}}(x_{<t}) = \left\{ x_t : q_L(x_t) \geq \alpha \max_{w \in \mathcal{V}} q_L(w) \right\}. \tag{3}$$

The premature layer is then selected dynamically at each token by maximizing the Jensen–Shannon Divergence (JSD) from the final layer over a candidate set $\mathcal{C}$. Here, $\mathcal{C}$ denotes the set of candidate premature layers used for contrast; in DoLa, layers are partitioned into several buckets (contiguous ranges), one bucket is chosen via validation, and the layers within the selected bucket (including a setting that considers only even-numbered layers for efficiency) constitute $\mathcal{C}$.

$$l^* = \arg\max_{l \in \mathcal{C}} \mathrm{JSD}\left(q_L(\cdot \mid x_{<t}), q_l(\cdot \mid x_{<t})\right). \tag{4}$$

3.2 Attention-Guided Layer Selection

DoLa selects premature layers based on divergences in token (vocabulary) distributions, but this criterion does not directly capture the underlying reference structure (i.e., where the model attends when making predictions). To address this limitation, we use attention distributions as a layer-selection signal that reflects layer-wise differences in which the model attends, and introduce three attention-guided layer selection strategies while preserving the DoLa framework.

Let $A^l \in \mathbb{R}^{H \times T \times T}$ denote the self-attention tensor at layer l, where $A^l_{h,t,i}$ is the attention weight assigned by head h at query position t to position i (a past token). For query position t, we define the head-averaged attention row over past tokens $1:t$ as $a^l_t \in \mathbb{R}^t$:

$$\begin{aligned} a^l_t(i) &= \frac{1}{H} \sum_{h=1}^{H} A^l_{h,t,i}, \qquad i \in \{1, \ldots, t\}, \\ a^l_t(i) &\geq 0, \qquad \sum_{i=1}^{t} a^l_t(i) = 1. \end{aligned} \tag{5}$$

3.2.1 Attention-JSD

Since attention distributions can reflect where each layer attends when making predictions, a layer whose attention pattern differs substantially from the final layer may correspond to a different stage of computation. Under this hypothesis, Attention-JSD selects the layer whose attention distribution is most divergent from the final layer:

$$l^*_{\text{attn-jsd}} = \arg\max_{l \in \mathcal{C}} \mathrm{JSD}\left(a^L_t, a^l_t\right). \tag{6}$$

3.2.2 Attention-Entropy-Max

The entropy of an attention distribution can be interpreted as a summary of how broadly attention is spread across past tokens. Under the hypothesis that broadly distributed attention can serve as a cue for layer selection, Attention-Entropy-Max selects the layer with the maximum attention entropy:

$$H(a_t^l) = -\sum_{i=1}^{t} a_t^l(i) \log a_t^l(i),$$
$$l^*_{\text{attn-ent-max}} = \arg\max_{l \in \mathcal{C}} H(a_t^l). \tag{7}$$

3.2.3 Attention-Entropy-Min

Conversely, low-entropy attention indicates that attention is concentrated on a small number of tokens, which may reflect strong reliance on specific evidence tokens. Under the hypothesis that such concentrated attention can serve as a cue for layer selection, Attention-Entropy-Min selects the layer with the minimum attention entropy:

$$l^*_{\text{attn-ent-min}} = \arg\min_{l \in \mathcal{C}} H(a_t^l). \tag{8}$$

3.3 Head-wise Attention Analysis

To better understand the behavior of our attention-based methods, we conduct a head-wise analysis.

In the definitions above, the attention distribution a_t^l is obtained by averaging over all H heads. In this analysis, we vary the range of heads used for averaging, and apply the resulting attention distributions to the computations of Attention-JSD and Attention-Entropy. We compare the following settings:

- **All heads**: Use all heads (default).
- **First-half**: Use only the first half of heads (1 to $H/2$).
- **Second-half**: Use only the second half of heads ($H/2 + 1$ to H), assuming H is even.

4 Experimental Setup

4.1 Models

In this study, we primarily report results on six representative models: LLaMA-7B[1], LLaMA-13B[2], LLaMA-33B[3], LLaMA-65B[4] for scaling analysis within the LLaMA family, plus Gemma-7B[5] and Mistral-7B[6] as representative non-LLaMA architectures.

Table 1: Summary of the six representative models used in the main tables. Layers and Heads denote the numbers of Transformer blocks and attention heads, respectively.

Model	Params	Layers	Heads
LLaMA-7B	7B	32	32
LLaMA-13B	13B	40	40
LLaMA-33B†	33B	60	52
LLaMA-65B†	65B	80	64
Gemma-7B	7B	28	16
Mistral-7B	7B	32	32

†Executed with int8 quantization.

Notably, LLaMA-33B and LLaMA-65B were executed with int8 quantization due to hardware memory constraints, while the other models were run in their native precision. Additional model details are provided in Appendix A.

4.2 Datasets

We use two representative benchmarks to evaluate the factuality of LLMs. For both benchmarks, we follow the same evaluation protocol as DoLa (Chuang et al., 2024b).

TruthfulQA. TruthfulQA is a benchmark designed to measure the tendency of LLMs to generate incorrect answers rooted in misconceptions or myths. We use the multiple-choice setting and report the following three metrics:

- **MC1**: Accuracy in the setting where only a single correct option exists.
- **MC2**: Normalized score in the setting where multiple correct options exist.
- **MC3**: Unnormalized score in the setting where multiple correct options exist.

FACTOR. FACTOR is a benchmark for evaluating factuality in long-form text. Each instance consists of a prefix and four candidate continuations, only one of which is factually correct. The dataset contains 4,266 four-choice questions spanning three domains: Wiki (2,994), News (1,036), and Expert (236). In this study, we evaluate the Wiki and News domains, using multiple-choice

[1] https://huggingface.co/huggyllama/llama-7b
[2] https://huggingface.co/huggyllama/llama-13b
[3] https://huggingface.co/huggyllama/llama-30b
[4] https://huggingface.co/huggyllama/llama-65b
[5] https://huggingface.co/google/gemma-7b
[6] https://huggingface.co/mistralai/Mistral-7B-v0.1

Table 2: Selected premature-layer ranges for Attention-JSD across representative models.

Model	TruthfulQA	FACTOR	Mature
LLaMA-7B	$[16, 32)$	$[2, 16)$	32
LLaMA-13B	$[20, 40)$	$[2, 20)$	40
LLaMA-33B	$[20, 40)$	$[2, 20)$	60
LLaMA-65B	$[60, 80)$	$[2, 20)$	80
Gemma-7B	$[14, 28)$	$[14, 28)$	28
Mistral-7B	$[16, 32)$	$[2, 16)$	32

scoring by comparing the log-likelihoods of candidate continuations.

4.3 Implementation Details

In DoLa, the candidate set of premature layers $\mathcal{C}$ is not taken directly from all layers; instead, layers are partitioned into several contiguous ranges (buckets), and the layers within a single bucket selected via validation are used as candidates (i.e., $\mathcal{C}$). In this section, we describe the bucket partition settings (candidate buckets).

Candidate Buckets. For the LLaMA family (7B, 13B, 33B, and 65B) and Mistral-7B, we follow the layer-selection ranges (bucket partitions) defined in the original DoLa paper (Chuang et al., 2024b). Specifically, for LLaMA-7B (32 layers) and Mistral-7B (32 layers), we use two buckets: $[0, 16)$ and $[16, 32)$. For LLaMA-13B (40 layers), we use two buckets: $[0, 20)$ and $[20, 40)$. For LLaMA-33B (60 layers), we use three buckets: $[0, 20)$, $[20, 40)$, and $[40, 60)$. For LLaMA-65B (80 layers), we use four buckets: $[0, 20)$, $[20, 40)$, $[40, 60)$, and $[60, 80)$.

In contrast, for Gemma-7B (28 layers), we partition the layers into two buckets: $[0, 14)$ and $[14, 28)$. Notably, for Gemma-7B, we exclude layer 0 from the candidate set for all methods, including the DoLa baseline (see Appendix A.1), as it primarily represents non-semantic embedding information. Furthermore, within the selected bucket, we only consider even-indexed layers as candidates to reduce computational overhead.

Bucket Selection. For TruthfulQA, following DoLa (Chuang et al., 2024b), we perform 2-fold validation and choose the bucket that yields the highest MC3 score. For FACTOR, we treat the Wiki and News subsets as two independent folds and select the optimal bucket for each. Table 2 reports the buckets selected by Attention-JSD for the representative models.

5 Results

This section reports results for the six representative models: LLaMA-7B, 13B, 33B, 65B, Gemma-7B, and Mistral-7B. Results for additional models are provided in Appendix D.

5.1 TruthfulQA

Table 3 reports the results of the six representative models on TruthfulQA. Here, the Baseline refers to evaluating each model with standard autoregressive decoding. The three proposed attention-distribution-based strategies consistently outperformed the Baseline across all evaluated models. Notably, the attention-based methods (Attention-JSD and Attention-Entropy-Max) contributed substantially to improving MC2 and MC3, which consider multiple acceptable answers.

Specifically, for LLaMA-7B, Attn-Ent-Min achieved the highest performance in MC2 (62.8%) and MC3 (36.5%), significantly surpassing DoLa. In LLaMA-13B, Attn-JSD yielded the best results across all metrics. For LLaMA-33B and LLaMA-65B, while DoLa maintained the highest MC1, our attention-based methods outperformed DoLa in both MC2 and MC3. This suggests that internal attention distributions provide a more sensitive signal than output vocabulary distributions for extracting factual knowledge, regardless of model scale.

A similar trend was observed in Gemma-7B and Mistral-7B, where the proposed methods consistently maintained or improved upon DoLa's scores in the MC2 and MC3 metrics. Overall, these results confirm that utilizing structural information from attention distributions, such as attention concentration (Min-Entropy) and divergence (JSD), serves as a robust mechanism for dynamically selecting contrastive layers to enhance factual accuracy.

5.2 FACTOR

Table 4 presents the evaluation results on the FACTOR benchmark. Overall, our proposed attention-guided strategies consistently improved factual accuracy over the Baseline. Furthermore, our methods demonstrated performance comparable to DoLa, with several configurations showing additional improvements in identifying factually correct continuations.

Specifically, on the LLaMA-7B Wiki subset, Attn-Ent-Max achieved an accuracy of 62.6%, outperforming DoLa (62.1%). For the Gemma-7B Wiki subset, Attn-JSD reached 64.1%, the highest

Table 3: Results on TruthfulQA for the six representative models (MC1 / MC2 / MC3, %). **Bold** indicates the best result for each model.

Model	Baseline	DoLa	Attn-JSD	Attn-Ent-Max	Attn-Ent-Min
LLaMA-7B	25.3 / 40.4 / 20.6	**34.6** / 56.5 / 30.5	34.1 / 56.2 / 29.3	34.0 / 55.4 / 28.8	33.4 / **62.8** / **36.5**
LLaMA-13B	28.3 / 42.7 / 22.4	30.5 / 59.0 / 32.9	**32.2** / **61.5** / **36.0**	32.1 / 60.4 / 35.4	30.2 / 60.7 / 34.2
LLaMA-33B	30.4 / 47.1 / 24.6	**32.0** / 57.7 / 33.0	30.8 / 57.5 / 33.1	29.0 / **57.8** / **33.3**	30.4 / 57.6 / 32.8
LLaMA-65B	30.8 / 46.5 / 24.8	**33.3** / 58.7 / 32.7	32.3 / 59.0 / 33.8	31.0 / **59.7** / **34.6**	32.4 / 59.2 / 33.9
Gemma-7B	31.5 / 47.5 / 24.6	36.8 / 57.7 / 31.1	36.8 / 57.7 / **31.2**	36.5 / 57.6 / 30.9	**37.2** / **58.2** / 31.1
Mistral-7B	30.8 / 48.0 / 25.6	**35.1** / 56.3 / 29.9	35.0 / **56.4** / **30.1**	35.0 / 56.3 / 29.9	35.0 / 56.3 / 29.9

among all compared methods, surpassing DoLa's 63.3%. In the Mistral-7B News subset, Attn-Ent-Max yielded the best performance with 76.3% accuracy.

While attention-guided methods did not surpass DoLa in certain cases, such as the Gemma-7B News subset, our results across most configurations indicate that attention-based signals—specifically attention entropy and JSD—serve as a powerful and competitive alternative to logit-based signals for identifying layers rich in factual knowledge.

6 Analysis

6.1 Layer Selection Patterns

Figure 1 illustrates the layer selection distributions for each method on LLaMA-7B. Here, we focus on methods that exhibit noticeable differences in performance (DoLa, Attn-JSD, and Attn-Ent-Min). Both DoLa and Attn-JSD exhibit a strong tendency to concentrate their selections on the shallowest available layers within the candidate bucket. On TruthfulQA, both methods select layer 16 in over 80% of cases.

In contrast, Attn-Ent-Min demonstrates a markedly different behavior, with selections distributed across a wider range of relatively deeper layers. On TruthfulQA, selections are spread across layers 22–30, and on FACTOR Wiki, they are distributed throughout the middle layers (8–14). This pattern indicates that the layer with the minimum attention entropy varies dynamically for each token, suggesting that Attn-Ent-Min adaptively selects premature layers based on the specific input context.

This qualitative difference in layer selection patterns suggests a potential contribution to the improvements observed in the MC2 and MC3 metrics, which evaluate the model's performance across multiple correct answer options.

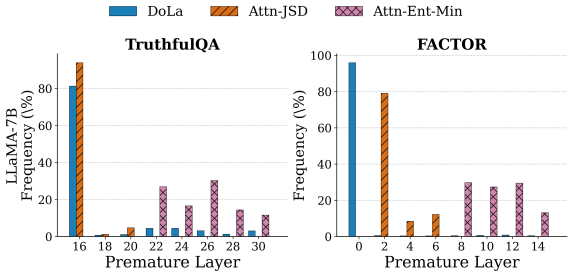


Figure 1: Layer selection distributions for LLaMA-7B on TruthfulQA (left) and FACTOR Wiki (right).

6.2 Head-wise Analysis

To investigate which attention heads contribute to layer selection in attention-based methods, we conducted a detailed head-wise analysis on TruthfulQA.

Coarse-grained Analysis. Table 5 presents the performance of Attn-Ent-Min across different head configurations. We observe that the optimal head group for layer selection varies by model. For LLaMA-7B, the First-half configuration achieved the best performance across all metrics, surpassing both the All-heads average and the Second-half group.

Conversely, for Gemma-7B, utilizing all heads resulted in the highest performance. In the case of Mistral-7B, the scores remained identical across all head configurations.

These results suggest that the distribution of layer-selection signals within attention entropy depends on the model architecture. While informative signals may be concentrated in specific head subsets for some models, they appear to be more dispersed or uniformly distributed in others.

Single-Head Evaluation. To further investigate the characteristics of the signals within individual attention heads, we analyzed their layer-selection tendencies when using each head alone. Specifically, we compared each head's selected layers against the all-heads average selection (All heads

Table 4: Results on FACTOR for the six representative models (Accuracy, %). **Bold** indicates the best result per model and split. Underline indicates a result worse than the Baseline.

Model	Split	Baseline	DoLa	Attn-JSD	Attn-Ent-Max	Attn-Ent-Min
LLaMA-7B	Wiki	58.3	62.1	62.4	**62.6**	62.0
	News	58.2	61.7	61.5	**62.0**	**62.0**
LLaMA-13B	Wiki	62.5	66.2	66.3	**66.5**	66.2
	News	60.7	62.5	62.7	**63.4**	63.0
LLaMA-33B	Wiki	68.3	**69.0**	68.6	**69.0**	68.8
	News	62.5	63.5	65.3	65.2	**65.5**
LLaMA-65B	Wiki	72.1	70.4	70.9	70.8	71.1
	News	62.8	63.5	65.2	**65.3**	65.2
Gemma-7B	Wiki	60.5	63.3	**64.1**	63.3	63.2
	News	74.0	**74.8**	73.3	73.6	72.7
Mistral-7B	Wiki	60.6	**64.6**	64.5	64.4	64.4
	News	75.9	75.9	75.3	**76.3**	76.0

Table 5: Head-wise analysis for Attn-Ent-Min on TruthfulQA. Best result per model is in **bold**.

Model	Heads	MC1	MC2	MC3
LLaMA-7B	All	**33.4**	62.8	36.5
	First-half	**33.4**	**62.9**	**37.3**
	Second-half	32.1	62.2	36.3
Gemma-7B	All	**37.2**	**58.2**	**31.1**
	First-half	36.2	57.5	30.5
	Second-half	36.8	58.0	31.0
Mistral-7B	All	**35.0**	**56.3**	**29.9**
	First-half	**35.0**	**56.3**	**29.9**
	Second-half	**35.0**	**56.3**	**29.9**

Table 6: Single-head evaluation for Attn-Ent-Min on TruthfulQA (LLaMA-7B). Best result per metric is **bold**.

Configuration	MC1	MC2	MC3
DoLa	34.6	56.5	30.5
All heads avg. (32)	32.4	62.7	36.8
Head 6	**35.3**	**64.9**	**38.4**
Heads 6, 24, 26	34.8	63.9	38.1

avg.) and focused on heads that most strongly deviate from this aggregate behavior. We selected Heads 6, 24, and 26 as the top three heads with the lowest agreement with the all-heads average selection. Table 6 presents their evaluation results on LLaMA-7B.

The results show that decoding guided solely by Head 6 achieved an MC3 score of 38.4%, significantly exceeding both DoLa (30.5%) and the all-heads average (36.8%). While the all-heads average selection provides a stable and robust baseline, combining these top three heads also maintained high performance, consistently surpassing both DoLa and the All heads avg. configuration.

These findings demonstrate that beyond simple averaging, factuality can be further improved by isolating and utilizing specific internal signals based on the quality and characteristics of the information they carry.

6.3 Head Stability Analysis

This section provides detailed results of the head stability analysis. We investigated whether the factuality-improving effects of specific attention heads are localized phenomena dependent on certain question sets or if they represent statistically stable characteristics across different contexts.

We generated five random subsets of 100 questions each from TruthfulQA and evaluated all 32 heads individually using the Attn-Ent-Min strategy on LLaMA-7B.

Table 7 lists the top-3 performing heads for each random subset based on the MC3 score.

As a result, Head 6 achieved the highest MC3 score in three out of the five subsets. While the performance ranking fluctuates depending on the subset, these results suggest a strong tendency for specific internal signals to be frequently associated with higher factual accuracy. This indicates that it may be possible to pre-identify or estimate superior configurations for layer selection—such as specific attention heads—to further optimize decoding performance.

Table 7: Top-3 performing heads for Attn-Ent-Min on each random subset (LLaMA-7B, MC3).

Subset	Rank 1	Rank 2	Rank 3
Subset 1	**Head 6**	Head 1	Head 22
Subset 2	Head 1	Head 15	Head 10
Subset 3	Head 26	**Head 6**	Head 5
Subset 4	**Head 6**	Head 22	Head 26
Subset 5	**Head 6**	Head 26	Head 22

6.4 Score Distribution Analysis

On LLaMA-33B, DoLa achieved the highest MC1 score (31.9%), while Attention-Entropy-Max showed a lower MC1 (29.0%) but slightly outperformed DoLa in MC2 and MC3 (MC2: 57.8% vs 57.7%, MC3: 33.3% vs 33.0%). To understand this difference, we analyzed the score distributions of individual samples when using DoLa and attention-based methods.

We first classified samples into four categories based on MC1 correctness (Table 8). Among the three attention-based methods, Attn-Ent-Max shows the least overlap with DoLa: only 150 samples are correct under both methods, while 111 are correct only with DoLa and 87 only with Attn-Ent-Max. This indicates that the two methods tend to correctly answer different samples. In contrast, Attn-JSD and Attn-Ent-Min show greater overlap with DoLa (191–194 samples correct under both), and their MC1 scores are closer to DoLa's.

We then analyzed the overall score distribution characteristics of each method (Table 9). For top margin (the gap between the highest correct and incorrect scores), DoLa achieves the largest value (14.8), while attention-based methods show smaller values (8.5–10.3). Notably, Attn-Ent-Max has the smallest top margin (8.5), indicating the weakest discriminability for identifying the top correct answer. Conversely, for true std (standard deviation of correct option scores), DoLa shows the largest value (31.3), while Attn-Ent-Max shows the smallest (17.5). This suggests that DoLa generates "sharp" distributions concentrating confidence on specific correct answers, whereas Attn-Ent-Max generates "flat" distributions spreading scores more evenly across all correct options.

These results reveal that DoLa's sharp layer selection is effective for identifying a single best answer (MC1), while attention-based methods with flatter selection patterns are advantageous for comprehensively evaluating multiple correct answers (MC2/MC3), demonstrating complementary characteristics between the two approaches.

Table 8: Sample-level correctness patterns (MC1) on TruthfulQA for LLaMA-33B.

Comparison	Both corr.	DoLa only	Attn only	Both wrong
DoLa vs Attn-JSD	194	67	58	498
DoLa vs Attn-Ent-Max	150	111	87	469
DoLa vs Attn-Ent-Min	191	70	57	499

Table 9: Overall score distribution statistics (LLaMA-33B). Top margin = max(correct) − max(incorrect). True std = standard deviation among correct option scores.

Method	Top Margin	True Std
DoLa	14.8	31.3
Attn-JSD	10.3	23.6
Attn-Ent-Max	8.5	17.5
Attn-Ent-Min	10.2	23.1

7 Conclusion

In this work, we proposed three strategies that leverage self-attention mechanisms rather than conventional vocabulary distributions as dynamic signals for layer selection in contrastive decoding. Our evaluation on TruthfulQA and FACTOR demonstrates that the proposed methods achieve factuality improvements equal to or greater than the existing DoLa method across multiple models.

In particular, our score distribution analysis suggests that DoLa tends to produce sharper distributions that concentrate confidence on a specific correct answer, which is effective for MC1, whereas attention-based methods produce flatter distributions that spread scores across all correct options, which is effective for MC2/MC3.

Furthermore, our head-wise analysis confirms that some attention heads exhibit stronger factuality signals than the all-head average. This suggests that further improvements in factuality and efficiency may be achieved by context-dependent head selection or by optimizing the set of heads used. In addition, our method requires no additional training and keeps the inference-time computational overhead within a practical range.

As future work, we will evaluate our approach in free-form generation settings and explore automated, context-dependent selection of layers and heads.

Limitations

This work has several limitations that define important directions for future research.

First, our evaluation is primarily limited to multiple-choice benchmarks. A crucial next step is to verify how attention-guided layer selection functions in free-form generation tasks, where sequential token dependencies may differ from likelihood based ranking.

Second, the selection of the candidate premature-layer bucket still requires prior validation. While attention-based signals provide sensitive cues, developing algorithms that can fully automate bucket selection or generalize across different model depths remains a necessary improvement.

Third, the underlying mechanism of why specific attention heads excel at factual knowledge recall remains to be fully elucidated. While the statistical stability observed in this study suggests that certain internal representations mediate truthfulness, identifying the theoretical basis for this localization is an essential task for future investigation.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24K20834.

References

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024a. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1419–1436.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024b. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 11733–11763.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

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.

Youna Kim, Hyuhng Joon Kim, Cheonbok Park, Choonghyun Park, Hyunsoo Cho, Junyeob Kim, Kang Min Yoo, Sang-goo Lee, and Taeuk Kim. 2024. Adaptive contrastive decoding in retrieval-augmented generation for handling noisy contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2421–2431.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2023. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*.

Dor Muhlgay, Ori Ram, Inbal Magar, Yoav Levine, Nir Ratner, Yonatan Belinkov, Omri Abend, Kevin Leyton-Brown, Amnon Shashua, and Yoav Shoham. 2024. Generating benchmarks for factuality evaluation of language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 49–66, St. Julian's, Malta. Association for Computational Linguistics.

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

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jialiang Wu, Yi Shen, Sijia Liu, Yi Tang, Sen Song, Xiaoyi Wang, and Longjun Cai. 2025. Improve decoding factuality by token-wise cross layer entropy of large language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3912–3921, Albuquerque, New Mexico. Association for Computational Linguistics.

A Additional Model Details

In addition to the six representative models reported in the main text, we also ran experiments on four additional models. In this appendix, we additionally report results for these models (LLaMA-3.1-8B[7], LLaMA-3.2-3B[8], Gemma-2-2B[9], and Phi-2[10]). Their specifications are summarized in Table 10.

Table 10: Additional models used in this study.

Model	Params	Layers	Heads
LLaMA-3.1-8B	8B	32	32
LLaMA-3.2-3B	3B	28	24
Gemma-2-2B	2.6B	26	8
Phi-2	2.7B	32	32

A.1 Handling of Layer 0 in Models with Tied Embeddings

The DoLa paper notes that when the input embeddings and the LM head share weights, the resulting operation becomes close to an identity mapping; therefore, in such cases, layer 0 is excluded from the candidate set. In our implementation, we consult the Hugging Face configuration (e.g., `tie_word_embeddings`) and exclude layer 0 when `tie_word_embeddings` is true (i.e., when weight tying is enabled), applying this rule to models such as Gemma variants and Llama 3.2-3B.

B Selected Premature-Layer Ranges by Method

The range of candidate premature layers for dynamic selection is determined via validation for each combination of model and dataset. Tables 11 through 14 list the selected buckets for all evaluated models across the four contrastive decoding methods.

C Efficiency Analysis

Table 15 shows the inference efficiency of each method on LLaMA-7B. We sampled 100 questions from TruthfulQA and measured the average processing time per sample.

All methods complete within a single forward pass, with overheads remaining within 1.15–1.27× relative to the Baseline. DoLa incurs a 1.24× cost due to per-candidate-layer logits and JSD computations. Among attention-based methods, Attn-Ent-Max/Min is lightweight (as low as 1.15×) because it only requires entropy computation, whereas Attn-JSD is the most expensive (1.27×) due to JSD computation between the final-layer attention distribution and each candidate layer. Overall, these costs remain practical given the factuality improvements.

[7] https://huggingface.co/meta-llama/Llama-3.1-8B

[8] https://huggingface.co/meta-llama/Llama-3.2-3B

[9] https://huggingface.co/google/gemma-2-2b

[10] https://huggingface.co/microsoft/phi-2

Table 11: Selected premature-layer ranges for DoLa.

Model	TruthfulQA	FACTOR	Mature
LLaMA-7B	[16, 32)	[0, 16)	32
LLaMA-13B	[20, 40)	[0, 20)	40
LLaMA-33B	[40, 60)	[0, 20)	60
LLaMA-65B	[60, 80)	[0, 20)	80
Gemma-7B	[14, 28)	[14, 28)	28
Mistral-7B	[0, 16)	[0, 16)	32
LLaMA-3.1-8B	[0, 16)	[16, 32)	32
LLaMA-3.2-3B	[2, 14)	[2, 14)	28
Gemma-2-2B	[2, 14)	[2, 14)	26
Phi-2	[16, 32)	[0, 16)	32

Table 12: Selected premature-layer ranges for Attention-JSD.

Model	TruthfulQA	FACTOR	Mature
LLaMA-7B	[16, 32)	[2, 16)	32
LLaMA-13B	[20, 40)	[2, 20)	40
LLaMA-33B	[20, 40)	[2, 20)	60
LLaMA-65B	[60, 80)	[2, 20)	80
Gemma-7B	[14, 28)	[14, 28)	28
Mistral-7B	[16, 32)	[2, 16)	32
LLaMA-3.1-8B	[16, 32)	[2, 16)	32
LLaMA-3.2-3B	[2, 14)	[2, 14)	28
Gemma-2-2B	[2, 14)	[2, 14)	26
Phi-2	[2, 16)	[2, 16)	32

D Full Experimental Results

This section reports full experimental results for all models evaluated in this study.

D.1 TruthfulQA: Full Results

Table 16 shows the results for all models on TruthfulQA. In newer models such as LLaMA-3.1-8B and LLaMA-3.2-3B, our proposed strategies (specifically Attn-JSD and Attn-Ent-Min) demonstrate stable improvements over the Baseline. Notably, on Phi-2, Attn-Ent-Min achieved an MC3 score of 34.8%, substantially outperforming DoLa (30.3%).

Table 13: Selected premature-layer ranges for Attention-Entropy-Max.

Model	TruthfulQA	FACTOR	Mature
LLaMA-7B	$[16, 32)$	$[2, 16)$	32
LLaMA-13B	$[20, 40)$	$[2, 20)$	40
LLaMA-33B	$[40, 60)$	$[2, 20)$	60
LLaMA-65B	$[60, 80)$	$[2, 20)$	80
Gemma-7B	$[14, 28)$	$[14, 28)$	28
Mistral-7B	$[2, 16)$	$[16, 32)$	32
LLaMA-3.1-8B	$[2, 16)$	$[2, 16)$	32
LLaMA-3.2-3B	$[2, 14)$	$[2, 14)$	28
Gemma-2-2B	$[14, 26)$	$[14, 26)$	26
Phi-2	$[16, 32)$	$[2, 16)$	32

Table 14: Selected premature-layer ranges for Attention-Entropy-Min.

Model	TruthfulQA	FACTOR	Mature
LLaMA-7B	$[16, 32)$	$[2, 16)$	32
LLaMA-13B	$[20, 40)$	$[2, 20)$	40
LLaMA-33B	$[20, 40)$	$[2, 20)$	60
LLaMA-65B	$[60, 80)$	$[2, 20)$	80
Gemma-7B	$[2, 14)$	$[2, 14)$	28
Mistral-7B	$[2, 16)$	$[16, 32)$	32
LLaMA-3.1-8B	$[2, 16)$	$[2, 16)$	32
LLaMA-3.2-3B	$[14, 28)$	$[2, 14)$	28
Gemma-2-2B	$[14, 26)$	$[2, 14)$	26
Phi-2	$[16, 32)$	$[2, 16)$	32

Table 15: Inference efficiency on TruthfulQA (LLaMA-7B, 100 samples).

Method	Time (s/sample)	Relative
Baseline	0.84 ± 0.32	$1.00\times$
DoLa	1.03 ± 0.44	$1.24\times$
Attn-JSD	1.06 ± 0.41	$1.27\times$
Attn-Ent-Max	0.96 ± 0.38	$1.15\times$
Attn-Ent-Min	1.00 ± 0.34	$1.20\times$

D.2 FACTOR: Full Results

Table 17 summarizes the full FACTOR results. While contrastive decoding generally contributes to factuality in both Wiki and News subsets, we observed certain cases, such as Gemma-2-2B, where contrastive methods led to a performance drop relative to the Baseline.

E Head-wise Analysis: Full Results

To supplement the analysis in the main text, we provide detailed performance metrics for different head configurations under the Attn-JSD and Attn-Ent-Max strategies. Tables 18 and 19 report these results for LLaMA-7B, Gemma-7B, and Mistral-7B.

Table 16: Results on TruthfulQA for all models (MC1 / MC2 / MC3, %). Bold indicates best per model. Underline indicates worse than Baseline.

Model	Baseline	DoLa	Attn-JSD	Attn-Ent-Max	Attn-Ent-Min
LLaMA-7B	25.3 / 40.4 / 20.6	**34.6** / 56.5 / 30.5	34.1 / 56.2 / 29.3	34.0 / 55.4 / 28.8	33.4 / **62.8** / **36.5**
LLaMA-13B	28.3 / 42.7 / 22.4	30.5 / 59.0 / 32.9	**32.2** / **61.5** / **36.0**	32.1 / 60.4 / 35.4	30.2 / 60.7 / 34.2
LLaMA-33B	30.4 / 47.1 / 24.6	**32.0** / 57.7 / 33.0	30.8 / 57.5 / 33.1	29.0 / **57.8** / **33.3**	30.4 / 57.6 / 32.8
LLaMA-65B	30.8 / 46.5 / 24.8	**33.3** / 58.7 / 32.7	32.3 / 59.0 / 33.8	31.0 / **59.7** / **34.6**	32.4 / 59.2 / 33.9
Gemma-7B	31.5 / 47.5 / 24.6	36.8 / 57.7 / 31.1	36.8 / 57.7 / **31.2**	36.5 / 57.6 / 30.9	**37.2** / **58.2** / 31.1
Mistral-7B	30.8 / 48.0 / 25.6	**35.1** / 56.3 / 29.9	35.0 / **56.4** / **30.1**	35.0 / 56.3 / 29.9	35.0 / 56.3 / 29.9
LLaMA-3.1-8B	31.6 / 49.1 / 26.3	**35.6** / 57.4 / **30.6**	**35.6** / **57.5** / **30.6**	35.5 / 57.4 / **30.6**	35.5 / 57.4 / 30.5
LLaMA-3.2-3B	27.5 / 44.5 / 23.9	**34.3** / 55.9 / 29.7	33.8 / 55.9 / 29.5	34.2 / 55.9 / 29.6	34.0 / **56.0** / **29.8**
Gemma-2-2B	27.1 / 42.6 / 22.0	**34.3** / 3.0 / **31.1**	30.8 / 46.9 / 27.8	30.0 / **54.3** / 28.0	28.3 / 53.3 / 30.8
Phi-2	28.5 / 42.8 / 21.9	33.4 / 55.5 / 30.3	**35.0** / 55.1 / 28.2	33.4 / 53.3 / 28.7	29.3 / **59.5** / **34.8**

Table 17: Results on FACTOR for all models (Accuracy, %). **Bold** indicates best per model. Underline indicates worse than Baseline.

Model	Split	Baseline	DoLa	Attn-JSD	Attn-Ent-Max	Attn-Ent-Min
LLaMA-7B	Wiki	58.3	62.1	62.4	**62.6**	62.0
	News	58.2	61.7	61.5	**62.0**	**62.0**
LLaMA-13B	Wiki	62.5	66.2	66.3	**66.5**	66.2
	News	60.7	62.5	62.7	**63.4**	63.0
LLaMA-33B	Wiki	68.3	**69.0**	68.6	**69.0**	68.8
	News	62.5	63.5	65.3	65.2	**65.5**
LLaMA-65B	Wiki	**72.1**	70.4	70.9	70.8	71.1
	News	62.8	63.5	65.2	**65.3**	65.2
Gemma-7B	Wiki	60.5	63.3	**64.1**	63.3	63.2
	News	74.0	**74.8**	73.3	73.6	72.7
Mistral-7B	Wiki	60.6	**64.6**	64.5	64.4	64.4
	News	75.9	75.9	75.3	**76.3**	76.0
LLaMA-3.1-8B	Wiki	63.9	**67.4**	**67.4**	67.3	67.1
	News	75.6	75.9	74.6	74.8	74.6
LLaMA-3.2-3B	Wiki	56.5	**61.6**	61.0	61.5	60.9
	News	68.9	69.4	**69.5**	69.4	**69.5**
Gemma-2-2B	Wiki	**52.4**	40.4	40.0	40.0	45.3
	News	**67.7**	46.5	44.0	45.2	52.7
Phi-2	Wiki	54.6	58.5	**58.8**	58.6	55.1
	News	57.8	**61.7**	61.4	60.3	58.6

Table 18: Head-wise analysis for Attn-JSD on TruthfulQA. Best result per model is in **bold**.

Model	Heads	MC1	MC2	MC3
LLaMA-7B	All	34.2	56.2	29.3
	First-half	**34.4**	55.4	28.9
	Second-half	34.1	**57.2**	**30.9**
Gemma-7B	All	**36.8**	**57.7**	**31.2**
	First-half	36.0	57.0	31.0
	Second-half	35.0	57.0	29.8
Mistral-7B	All	**35.0**	56.4	**30.1**
	First-half	**35.0**	**56.6**	29.9
	Second-half	34.8	56.2	29.7

Table 19: Head-wise analysis for Attn-Ent-Max on TruthfulQA. Best result per model is in **bold**.

Model	Heads	MC1	MC2	MC3
LLaMA-7B	All	**34.0**	**55.4**	**28.8**
	First-half	33.9	54.7	28.4
	Second-half	33.9	54.7	28.4
Gemma-7B	All	36.5	**57.6**	30.9
	First-half	**36.6**	**57.6**	**31.0**
	Second-half	35.6	57.5	30.4
Mistral-7B	All	**35.0**	**56.3**	**29.9**
	First-half	**35.0**	**56.3**	**29.9**
	Second-half	**35.0**	**56.3**	**29.9**