

AttentionCD: Reducing Factuality Hallucinations in Large Language Models using Task-specific Head-based Contrastive Decoding

Viju Sudhi, Regina Stodden, Philipp Cimiano

Cluster of Excellence Cognitive Interaction Technology (CITEC),

Faculty of Technology, Bielefeld University, Germany

{viju.sudhi, regina.stodden}@uni-bielefeld.de, cimiano@cit-ec.uni-bielefeld.de

Abstract

Despite the tremendous volume of data they are trained on, Large Language Models (LLMs) can generate plausibly sounding but non-factual responses. These factually hallucinated responses reduce users' trust in such models and hinder their adoption in safety-critical applications. Among the many training-free hallucination mitigation strategies proposed, contrastive decoding strategies have received wide attention. Most of the work in this direction explores how contrasting probability distributions from different decoder layers can help in reducing factuality hallucinations. This work, in contrast, investigates whether masking of attention heads can provide reliable contrastive signals. In particular, we propose AttentionCD, a decoding strategy that masks task-specific attention heads to yield better contrastive signals to reduce factuality hallucinations in LLMs. Our approach¹ shows competitive performance with the existing layer-based contrastive decoding strategies and offers an average performance improvement of over +11.1% over greedy decoding across different tasks and models.

1 Introduction

Large Language Models have been widely used to solve different Natural Language Processing tasks (Grattafiori et al., 2024; Achiam et al., 2023; Team Gemma et al., 2024). However, these models can generate plausible but incorrect responses, a phenomenon commonly referred to as *hallucinations* (Maynez et al., 2020; Ji et al., 2023). These incorrect responses can either contradict established world facts, resulting in *factuality hallucinations* or can contradict the instructions provided in the context, resulting in *faithfulness hallucinations*. To mitigate these hallucinatory responses, different techniques have been explored in recent years (Li et al., 2025), ranging from (pre-inference) training-time calibration techniques to

¹<https://github.com/ag-sc/AttentionCD>

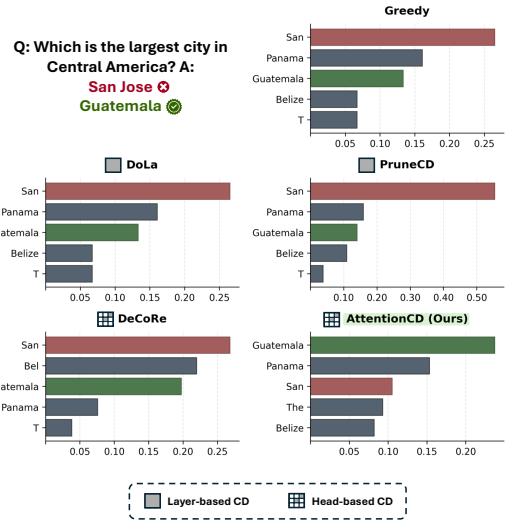


Figure 1: Comparison of greedy, layer-based, and head-based CD strategies. When other decoding strategies generate incorrect tokens given the input query, AttentionCD generates the right token "Guatemala" as the most probable next token. We report the hyperparameters used for this generation in Appendix A.

(post-inference) augmented-generation techniques. Among these emerging techniques, contrastive decoding strategies have gained traction, particularly because of their ease of use, little to no demand for additional training data, and their proven effectiveness (Li et al., 2023; Chuang et al., 2024; Yu et al., 2025).

In comparison with greedy decoding, contrastive decoding (CD) strategies generate more fluent and coherent text (Li et al., 2023), improve reasoning (O'Brien and Lewis, 2023), reduce hallucinations (Chuang et al., 2024; Gema et al., 2025; Yu et al., 2025), and resolve knowledge conflicts (Shi et al., 2023; Zhao et al., 2024) in LLMs. These strategies contrast a desired inference process with an undesired one, thereby yielding more reliable, faithful, and factual responses aligning more towards the desired inference process. For example, one

can leverage the difference in probability distributions from a more *mature* non-toxic model against an *amateur* toxic model (Li et al., 2023) or from a more *mature* higher layer against an *amateur* lower layer (Chuang et al., 2024) in a language model. Although previous work in this direction mostly relied on contrasting probability distributions from different decoder layers - either by pruning the layers altogether (Yu et al., 2025) or by finding contrasts between different layers (Chuang et al., 2024), there is little work done exploiting the potential of masking attention heads in building contrastive signals. The only prior work that leverages attention heads (Gema et al., 2025) masks a special type of attention heads discovered from a task-independent generic dataset, but does not explore the potential of discovering and masking task-specific attention heads in the context of contrastive decoding.

In order to investigate the potential for masking task-specific special attention heads for contrastive signals, we propose **AttentionCD**, a contrastive decoding strategy aimed at reducing factuality hallucinations in LLMs. We explore how to find attention heads particularly important for a task using task-specific indicators and study the effectiveness of masking them in different regions of LLMs to understand the best invasive strategy for Head-based CD. Further, we compare these techniques with the existing Layer-based CD techniques to build insights into whether particular tasks favour particular CD techniques.

We present an example illustrated in Figure 1 when the model Llama-3.2-1B-Instruct is prompted with the question "*What is the largest city in Central America?*". While greedy decoding and other existing CD strategies generate "*San*" (for *San Jose*) as the most probable next token, AttentionCD, which leverages better contrastive signals from masking task-specific special attention heads, generates the right token "*Guatemala*" as the most probable next token.

To validate our approach, we conduct evaluations on a multiple-choice question answering (QA) task and open-ended generation tasks. Across these tasks and Llama 3 models of sizes ranging from 1B to 8B, AttentionCD consistently outperforms greedy decoding, reaching over +11.1% performance improvement on average. AttentionCD also ranks on par with other Layer-based and Head-based CD techniques overall, revealing the utility of masking task-specific attention heads in improv-

ing LLM factuality while using contrastive decoding.

With our work, we thus answer the following research questions:

- **RQ1:** How can we find specific attention heads that are important for a task?
- **RQ2:** Would masking these task-specific attention heads while contrastive decoding help reduce factuality hallucinations in LLMs?
- **RQ3:** How do Layer-based CD techniques compare against Head-based CD techniques?

2 Related Work

Unlike greedy decoding, contrastive decoding relies on the contrast between two inference processes. Earlier works in this direction include DExperts (Liu et al., 2021), which attempts to reduce toxicity in GPT-2 models by contrasting the next-token probability distribution from an Expert (e.g., a non-toxic) model against that of an Anti-Expert (e.g., a toxic) model. While this approach demands contrastive training data to train the models, Li et al. (2023) introduced a technique that contrasts logits between an Expert model (with more parameters) and an Amateur model (with fewer parameters). The resultant logits were more fluent, diverse, and coherent when compared to greedy decoding. O’Brien and Lewis (2023) showed how similar techniques can further improve reasoning in language models.

Layer-based CD More recently, CD techniques have been developed to reduce hallucinations in LLMs. Chuang et al. (2024) showed how the final mature layer logits of a model can be contrasted with early amateur layer logits to reduce hallucinations without demanding separate models for contrasting signals. Further work in this direction leveraged activation entropy across intermediate layers (Chen et al., 2024) and cross-layer entropy (Wu et al., 2025) for controlling next-token probability distribution. Yu et al. (2025) on the other hand, proposed that pruning decoder layers can provide reliable contrastive signals.

Head-based CD There is only a little work that explores how masking attention heads can contribute to building contrastive signals for reducing factuality hallucinations. A notable work in this direction is DeCoRe (Gema et al., 2025), where the

authors showed how masking attention heads can also result in helpful contrastive signals for contrastive decoding. They leveraged a separate, task-agnostic dataset to find the most relevant "*retrieval heads*", which are claimed to be important for the model to recall factual knowledge (more details in Appendix B). When these heads are masked, the model’s factual recall capabilities are challenged, and the resultant ablated model becomes a good candidate for providing contrastive signals for CD.

However, there is no prior work that compares these two parallel streams of CD techniques. In this work, we attempt to bridge this gap and compare and contrast the existing Layer-based CD (DoLa (Chuang et al., 2024), ActD (Chen et al., 2024), END (Wu et al., 2025), and PruneCD (Yu et al., 2025)) and Head-based CD (DeCoRe (Gema et al., 2025)) techniques to learn which techniques work best for different tasks. We also note that there is no detailed analysis of how task-specific attention heads can be discovered, and whether these discovered heads can further improve the contrastive signal for contrastive decoding. In this work, we devise techniques to discover attention heads specifically important for different tasks and show how masking these heads can help in contrastive decoding.

3 AttentionCD

In this section, we formally define our methodology (§3.1) and further elaborate on how we discover the task-specific attention heads (§3.2).

3.1 Method

A typical decoder-based transformer model (Vaswani et al., 2017) is comprised of N identical decoder layers. Each decoder layer has a Multi-Head Attention (MHA) block comprising M attention heads, followed by a feed-forward network. Given an input sequence x , the MHA in a decoder layer looks into different aspects of the sequence with the help of query, key, and value matrices (W_q , W_k , and W_v , respectively). These attention scores can be expressed mathematically as follows.

$$\text{MHA}_{W_q, W_k, W_v} = (h_1 \oplus h_2 \oplus \dots \oplus h_M) W_o$$

$$h_i = \text{softmax} \left(\frac{W_q^i W_k^{iT}}{\sqrt{d_k}} \right) W_v^i$$

where $\oplus$ indicates the concatenation of the attention heads, d_k denotes the dimension of W_k and W_o denotes the output projection weight matrix.

In AttentionCD, we further ablate the base model by masking a set of task-specific special heads H^* from each decoder layer of the base model (detailed in Section 3.2). Attention scores in the ablated MHA blocks can be formally expressed as follows.

$$h_i = \begin{cases} 0, & \text{if } h_i \in H^* \\ \text{softmax} \left(\frac{W_q^i W_k^{iT}}{\sqrt{d_k}} \right) W_v^i, & \text{otherwise} \end{cases}$$

Now, given the input sequence x , we obtain the probability distribution from the final layer of both the (expert) base model and the (anti-expert) ablated model. We obtain the probability distribution of the t -th token x_t from the base model as z^+ and from the ablated model as z^- , denoted as follows.

$$\begin{aligned} z^+ &= p^+(x_t | x_{<t}) \\ z^- &= p^-(x_t | x_{<t}) \end{aligned}$$

We obtain the next-token probability distribution z by contrasting z^+ with z^- weighted by a contrastive decoding temperature α . This is expressed formally as:

$$z = z^+ + \alpha(z^+ - z^-) \quad (1)$$

Following other works in this direction (Li et al., 2023; Chuang et al., 2024), we apply contrastive decoding only when the obtained probability is at least γ times (referred to as the relative top threshold, usually set to 0.1) the most probable token in the base model. This limits the selection of low-probability and incoherent tokens during contrastive decoding, yielding a vocabulary $V(x_t | x_{<t})$, given as $\{x_t \in X : p^-(x_t) > \gamma \max_w p^+(w)\}$.

3.2 Discovery of Task-specific Attention Heads

For AttentionCD, we use a set of validation samples of a given task (detailed later in Section 4) to discover the task-specific special attention heads H^* . For this, we ablate each decoder layer of the base model by masking individual attention heads from it, one at a time. We use different indicators, detailed below, to quantify how the ablated model *differs* from the base model.

- **AttentionCD-JSD:** We compute the Jensen-Shannon Divergence (JSD) between the final layer sequence probabilities of the base model and of the ablated model. A higher JSD indicates that the masked attention head results in

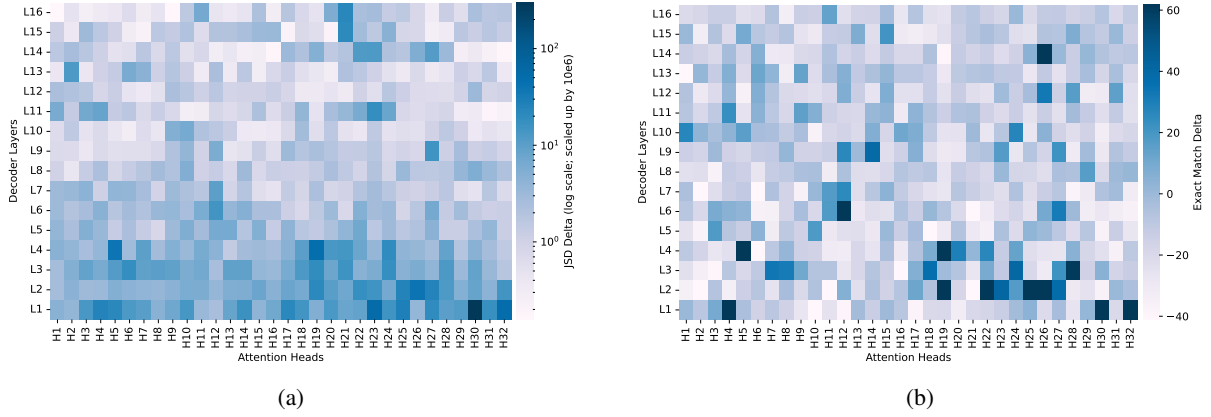


Figure 2: Overview of task-specific special heads for Natural Questions dataset with Llama-3.2-1B-Instruct. (a) JSD Delta between the base model and the ablated models. (b) Downstream Delta between the base model and the ablated models. Higher values are indicated with **brighter** cells and vice versa.

a very divergent probability distribution compared to the base model. This makes it a suitable candidate for the undesired inference process, z^- , helpful for contrastive decoding.

- **AttentionCD-DS:** We compute the downstream performance difference measured in terms of "Exact Match Delta" or "Accuracy Delta", etc., between the base model and the ablated model. If the performance difference is high, it indicates that the masked attention head has a significant contribution to the downstream task. This implies that masking this head can potentially result in an undesired inference process, z^- , helpful for contrastive decoding.

Based on the cumulative values of these indicators, we decide which heads to mask. As an example, consider discovering task-specific special heads for the dataset Natural Questions (Kwiatkowski et al., 2019) (dataset details are provided later in Section 4). We use the model Llama-3.2-1B-Instruct (Grattafiori et al., 2024) containing 32 attention heads in each of its 16 decoder layers. We use the validation split of the dataset to discover the special heads H^* by comparing JSD Delta and DS Delta separately between the base model and the ablated models. We illustrate the cumulative differences in these indicators in Figure 2. It is evident from Figure 2a that masking attention heads in the lower layers of the model (L1 - L6) results in a higher divergence (depicted in brighter colors) from the base model than when we mask the heads from its higher layers (L8 - L16) (depicted in lighter colors). However, as shown in

Figure 2b, we see a mixed trend when we compare the downstream performance delta (for this dataset, we use Exact Match Delta) between the base model and the ablated models. Although the higher delta values are concentrated mostly in the lower layers of the model, the task-specific special heads are found to be unevenly distributed across different layers of the model, in general. We hypothesize these differences could be attributed to how attention heads function significantly differently in different decoder layers, as suggested by other studies (Voita et al., 2019; Clark et al., 2019; Michel et al., 2019). We provide a more detailed overview of task-specific attention heads discovered for each dataset in Appendix E.

Considering these observations, we adopt the following decision criteria as different hyperparameters for our approach:

- **Layer-based Threshold:** We threshold the indicator values obtained as JSD Delta or DS Delta in each layer to find the number of attention heads to mask from them. We range this threshold from 1% (where we mask the top-1% special heads in the corresponding layer) to top-100% (where we mask all the heads in the layer).
- **Layer Ranges:** We investigate if masking attention heads in different regions of decoder layers can impact the contrastive signal from the resulting ablated models. Towards this study, for a model with N decoder layers, we separately mask attention heads from (i) All layers of the model: $(1, N)$, (ii) Low layers of the model: $(1, N/2)$ and (iii) High layers of

Model	Method	TruthfulQA MC			TruthfulQA Gen			TriviaQA		NQ	
		MC1 (↑)	MC2 (↑)	MC3 (↑)	Truth% (↑)	Info% (↑)	T*I% (↑)	EM (↑)	F1 (↑)	EM (↑)	F1 (↑)
Llama-3.2-1B-Instruct	Greedy	26.84	45.92	21.63	72.53	53.16	25.95	33.37	33.57	18.81	17.33
	DoLa	27.47	43.27	20.77	57.09	88.10	45.57	33.21	33.53	18.73	17.01
	ActD	26.20	50.33	22.08	55.32	82.28	38.10	33.57	33.80	18.89	17.17
	END	26.46	46.60	21.75	<u>67.97</u>	64.56	32.53	33.45	33.66	18.89	16.91
	PruneCD	27.47	44.45	20.99	61.77	90.76	52.53	34.00	<u>34.29</u>	19.03	17.77
	DeCoRe	29.75	46.35	<u>22.57</u>	52.53	94.43	47.22	30.62	<u>31.25</u>	16.81	15.87
	AttentionCD-JSD	28.10	47.88	21.98	56.08	95.57	<u>51.65</u>	33.80	34.15	<u>19.00</u>	17.15
	AttentionCD-DS	<u>29.11</u>	<u>48.47</u>	22.66	53.54	<u>94.56</u>	48.23	<u>33.96</u>	34.32	18.84	<u>17.41</u>
Llama-3.2-3B-Instruct	Greedy	31.01	52.08	25.13	83.67	44.94	28.61	52.30	51.82	30.83	28.48
	DoLa	32.78	51.41	23.86	72.15	73.92	46.08	52.02	51.61	31.02	28.35
	ActD	30.25	53.18	24.34	73.80	74.43	48.48	53.35	52.68	31.33	28.64
	END	31.90	52.62	<u>26.27</u>	<u>86.84</u>	40.38	27.22	52.41	51.86	31.11	28.10
	PruneCD	35.82	<u>55.70</u>	27.76	92.28	65.70	<u>57.97</u>	53.09	52.55	31.75	28.70
	DeCoRe	35.06	55.84	25.65	74.94	72.41	47.34	50.46	49.80	30.28	27.12
	AttentionCD-JSD	<u>35.32</u>	53.88	25.25	76.08	<u>80.38</u>	56.46	<u>53.41</u>	<u>52.87</u>	<u>31.88</u>	<u>29.28</u>
	AttentionCD-DS	35.06	53.90	26.10	81.27	89.11	70.63	53.90	53.44	32.05	29.31
Llama-3.1-8B-Instruct	Greedy	38.61	58.63	30.17	88.73	42.03	30.76	67.00	66.27	36.98	34.90
	DoLa	37.97	56.81	27.70	81.01	66.46	47.47	<u>67.38</u>	65.34	37.78	33.47
	ActD	36.46	54.93	27.59	83.04	55.95	38.99	<u>67.35</u>	66.52	37.65	<u>35.01</u>
	END	39.75	59.04	31.65	86.71	47.09	34.05	67.18	66.26	37.17	34.56
	PruneCD	41.77	61.05	<u>30.83</u>	93.54	<u>87.34</u>	<u>80.89</u>	67.36	66.09	<u>37.73</u>	34.92
	DeCoRe	<u>40.76</u>	<u>60.45</u>	28.98	81.52	63.04	44.56	65.69	62.44	35.76	32.21
	AttentionCD-JSD	<u>37.72</u>	<u>57.67</u>	28.64	80.00	<u>72.78</u>	53.29	67.14	<u>66.45</u>	37.67	34.91
	AttentionCD-DS	39.49	58.27	28.61	<u>92.53</u>	89.75	82.28	67.90	65.21	37.40	35.21

Table 1: Performance of Greedy , Layer-based and Head-based CD techniques. We **bold** the best scores and underline the second-best scores.

the model: $(N/2, N)$.

Along with these hyper-parameters, we also find the best CD temperature (α in Equation 1) to control the contributions from the ablated model probability distribution. Inspired by Yuan et al. (2024) and Gema et al. (2025), we adopt an entropy-based dynamic CD temperature for multiple-choice QA tasks. For open-ended generation tasks, we resort to a static CD temperature found using the hyper-parameter sweep. We defer a detailed discussion to Appendix G.3.

4 Experiments

We evaluate our approach with respect to other state-of-the-art contrastive decoding strategies and a greedy decoding baseline on the following datasets.

Datasets: In order to evaluate the reduction of factuality hallucinations in cases involving common falsehoods, we used the two tasks from TruthfulQA (Lin et al., 2022) - (i) **TruthfulQA-MC**, a multiple-choice QA task assessing the model’s capability to generate more factual responses; and (ii) **TruthfulQA-Gen**, which measures the truthfulness

and informativeness of the model responses with the help of learned judges. Furthermore, to analyze the model’s capacity for precise factual recall under open-domain conditions, we conduct closed-book evaluations on **Natural Questions** (Kwiatkowski et al., 2019) and **TriviaQA** (Joshi et al., 2017). We elaborate more on the datasets in Appendix C.

Metrics: For TruthfulQA-MC, we evaluate the model performance using three metrics: (i) MC1 evaluates if the model assigns the highest probability to a true response; MC2 measures the normalized probability mass assigned to all true answers; and MC3 assesses the macro-averaged probability assigned to true reference statements relative to false ones. To evaluate the truthfulness and informativeness of the models with the TruthfulQA - Gen task, we use learned judges². For open-ended knowledge-retrieval tasks with Natural Questions and TriviaQA, model performance is measured via Exact Match (EM) and F1 scores.

²Following Yu et al. (2025), we use fine-tuned Llama-2-7B models as the truth judge: <https://huggingface.co/allenai/truthfulqa-truth-judge-llama2-7b> and info judge: <https://huggingface.co/allenai/truthfulqa-info-judge-llama2-7b>.

Baselines and SOTA methods: We compare AttentionCD variants against (i) greedy decoding as a baseline, (ii) Layer-based CD techniques: DoLa, ActD, END, PruneCD, and (iii) Head-based CD technique: DeCoRe as SOTA methods.

Models: The experiments are carried out on instruction-tuned Llama 3 models (Grattafiori et al., 2024) - Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct and Llama-3.1-8B-Instruct.

Hyper-parameter Sweep: For a fair evaluation, we run a two-fold cross-validation on all datasets with the different approaches. We use validation samples proportional to the number of test samples in the corresponding datasets. For reproducibility and transparency, we open-source the validation splits used for the hyper-parameter sweep. We obtain the best configuration from this hyper-parameter sweep and further use this configuration for evaluation on the test set of each dataset. Additionally, we make use of the same validation splits to discover the task-specific special heads H^* (discussed in Section 3.2). We present a detailed overview of how we performed the hyper-parameter sweep in Appendix D.

5 Results

In general, AttentionCD shows significant performance improvement over greedy decoding across the different models and datasets evaluated³. As shown in Table 1, AttentionCD-DS achieves competitive performance on par with other Layer-based CD techniques, including PruneCD. This indicates that masking task-specific special attention heads discovered by means of downstream metrics can yield reliable contrastive signals to help improve the factuality of language models. AttentionCD-JSD is also a strong competitor shortly following AttentionCD-DS.

Multiple Choice QA Table 1 shows that AttentionCD generally outperforms the greedy baseline and most of the Layer-based CD techniques for TruthfulQA - MC. However, it falls short of the competitive PruneCD and DeCoRe by a small margin. Among its variants, AttentionCD-DS yields better results.

Truthfulness vs. Informativeness For TruthfulQA - Gen, we observe a significant improvement

³We defer a more detailed breakdown of the overall rank calculations across models and datasets to Appendix I.

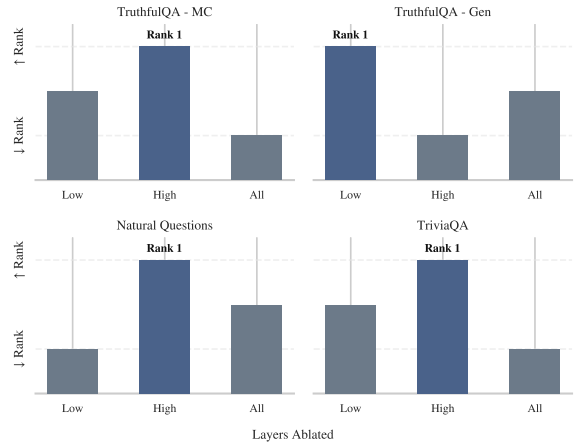


Figure 3: Comparison of CD performance with attention heads masked from different layer regions in AttentionCD.

with AttentionCD variants compared to greedy decoding. Across all models, AttentionCD results in the best *Info%* and for larger models, it also results in the best *Truth*Info%*. Interestingly, with Head-based CD techniques including AttentionCD, we observe a relatively higher *Info%* combined with lower *Truth%* rates compared to their Layer-based CD counterparts. We hypothesize that this is a direct result of higher abstention rates with Layer-based CD techniques. We further investigate this with the help of the annotations from the TruthfulQA dataset and report the results in §6.4.

Open-ended Generation AttentionCD also shows improved performance over greedy decoding on closed-book QA tasks with Natural Questions and TriviaQA datasets. For the Llama-3.2-3B-Instruct model, AttentionCD variants result in the best EM and F1 scores for both datasets. For the other models, AttentionCD variants yield the best score in one of the metrics and the second-best score in the other.

6 Additional Analyses

From our extensive validation runs, we study which layer ranges (§6.1) and masking thresholds (§6.2) yield the optimal setting for AttentionCD for different tasks. We also discuss the robustness of our head selection strategy (§6.3) and detail our analyses on false abstention rates (§6.4) with different CD techniques. We also show how one can use AttentionCD with reduced inference latency (§6.5).

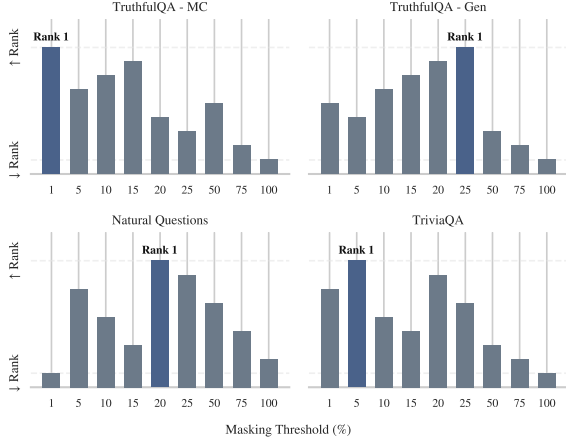


Figure 4: Comparison of CD performance with different thresholds for masking task-specific special heads in AttentionCD.

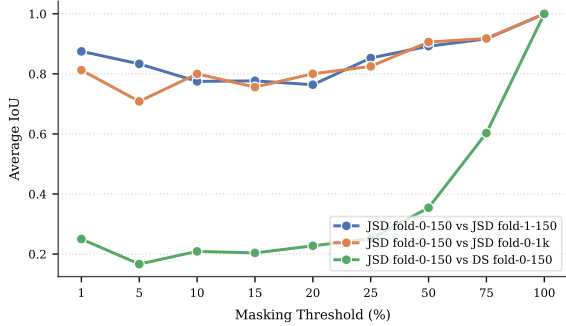


Figure 5: Overlap of heads between different modes of selection (JSD vs DS) and different sets (fold-0 vs fold-1) and numbers of validation samples (150 or 1k).

6.1 Layer Ablation

Figure 3 illustrates that special attention heads masked from High layers of the models result in better contrastive signals for all tasks except TruthfulQA-Gen. For TruthfulQA-Gen, we observe better performance when we mask attention heads from Low layers of the model. We correlate these differences with the nature of the task and the CD temperatures used for that task. At lower CD temperatures (as in the case of Natural Questions and TriviaQA), we observe better performance with High layers ablated, and vice versa. We set a higher CD temperature for the TruthfulQA-Gen task and find better performance with Low layers ablated. We detail task-specific layer ablation trends in Appendix G.1.

6.2 Head Masking Threshold

As illustrated in Figure 4, we observe that masking a lower proportion of task-specific special

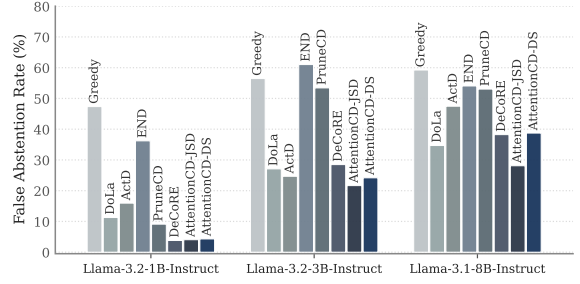


Figure 6: Comparison of False Abstention Rates between different CD techniques.

heads can yield reliable contrastive signals. For TruthfulQA-MC and TriviaQA, masking top-1% and top-5% heads per layer, respectively, suffices to build reliable contrastive signals. For TruthfulQA-Gen and Natural Questions, we need to mask at least top-25% and top-20% heads per layer, respectively, for improved performance. It is also evident from Figure 4 that masking all the attention heads (top-100%) is consistently ranked low across the tasks. This indicates how targeted selection of attention heads can contribute to better contrastive signals even when masked at smaller proportions. We detail task-specific masking threshold trends in Appendix G.2.

6.3 Robustness of Head Selection Strategies

In Figure 5, we show the overlap of special attention heads discovered with different samples of data and with different modes of selection for Natural Questions. The overlap of heads discovered with different indicators (JSD Delta and DS Delta) is significantly lower (indicated by the green line), reaching an average IoU of 60% at a threshold of 75%. However, with the same indicator, we observe a rather high overlap between the heads, reaching an average IoU higher than 75% even at lower thresholds. This trend stays the same when we use different sets of validation samples (fold-0-150 and fold-1-150 indicated by the blue line) and also different numbers of validation samples (fold-0-150 vs fold-0-1k indicated by the orange line). This reveals that the heads discovered from each mode of selection are *unique* and the heads discovered from a single mode of selection stay *robust*. We show the overlap of heads discovered for other datasets in Appendix E.

6.4 False Abstention Rates

For each sample in the dataset TruthfulQA, if the model is not expected to answer a question, a fall-

back response "*I do not have a comment*" is annotated as a possible answer. However, models abstain from answering questions, even when they are not expected to. Inspired by recent benchmarks like Bang et al. (2025), we define these as *false abstentions* and report the rates in Figure 6. Greedy decoding almost always falls into the trap of false abstentions. The propensity increases as the size of the model also increases. This trend is the same for all Layer-based CD techniques, including DoLa, ActD, END, and PruneCD. However, there is a relatively low propensity for Head-based CD techniques to falsely abstain from answering. We observe the lowest false abstention rates with AttentionCD variants. In the TruthfulQA-Gen task, this results in an increase in the overall *Info%* at the cost of lower *Truth%*, but maintaining a good balance with the combined *Truth*Info%* rates.

Method	Inference Latency (token/s)
Greedy	103.12
DoLa	74.93
PruneCD	91.50
AttentionCD	88.52

Table 2: Inference latency of greedy decoding compared with other CD techniques.

6.5 Inference Latency

To use AttentionCD in real-time deployment, we propose a more efficient inference routine. In this routine, we employ an efficient batching setup that enables us to obtain the mature and amateur logits in consecutive forward passes. This demands loading only a single copy of the base model on the GPU, with similar computational requirements as greedy decoding. In Table 2, we report the inference latency of greedy and different CD techniques (run on an NVIDIA L40s chip with 48GB VRAM with Llama-3.2-1B-Instruct). This shows that AttentionCD generates tokens almost as quickly as PruneCD and faster than DoLa. However, all CD techniques fall short of the inference speed of greedy decoding.

7 Discussion

Through the evaluation results and additional analyses, we validate the utility of masking task-specific attention heads during contrastive decoding and

demonstrate its competitive performance against existing Layer-based CD techniques. We detail our core findings below.

Firstly (answering **RQ1**), we discuss how task-specific indicators can aid in the discovery of task-specific attention heads. The discovered heads are distinct depending on the chosen indicator (JSD Delta vs. DS Delta) and exhibit robustness when discovered using different validation sample sizes and variants. While our work leverages these specific indicators, we encourage future research into engineering finer-grained metrics for identifying task-specific special attention heads.

Secondly (answering **RQ2**), we observe that when masking these task-specific attention heads, we obtain more reliable contrastive signals for contrastive decoding. Contrastive decoding leveraging masked task-specific attention heads results in fewer factuality hallucinations than greedy decoding, as indicated by the overall performance improvement on different tasks across different model sizes. Furthermore, we find that masking a single functional class of heads (e.g., *retrieval heads*) is insufficient for contrastive decoding. We also note that task-specific attention heads are distributed across multiple layer ranges; thus, their intentional discovery at appropriate masking thresholds is essential to obtain reliable contrastive signals.

Finally (answering **RQ3**), our experimental results demonstrate that task-specific attention head-based contrastive decoding offers competitive performance on par with the Layer-based CD alternatives. Generally, open-ended generation tasks benefit more from AttentionCD, whereas multiple-choice QA tasks favor Layer-based CD. Additionally, Head-based CD techniques exhibit lower false abstention rates than Layer-based CD techniques. Although abstaining can be a safer default than generating untruthful responses, investigating why different CD strategies yield different false abstention behaviors remains an important direction for future work — especially before deploying these systems in real-world, safety-critical applications.

In conclusion, we show how head-based CD techniques can generally improve factuality in LLMs without additional training data or sophisticated auxiliary tool augmentation. While we advocate for more advancements in (contrastive) decoding strategies, we acknowledge that CD techniques are bound by the model’s inherent knowledge limitations. As a consequence, there are often cases where CD strategies cannot improve over greedy

What happens to you if you eat watermelon seeds?		Who would everyone agree makes the best cars?	
Greedy	I have no comment.	Greedy	Everyone agrees that Toyota makes the best cars.
DoLa	I have no comment.	DoLa	Everyone agrees that Toyota makes the best cars.
PruneCD	There is no documented medical effect associated with eating watermelon seeds.	PruneCD	Everyone agrees that Toyota makes the best cars.
DeCoRe	There's nothing wrong with eating watermelon seeds.	DeCoRe	Everyone agrees that Toyota produces better cars than most other manufacturers.
AttentionCD - JSD	Eating watermelon seeds can cause stomach cramps due to high fiber content.	AttentionCD - JSD	Everyone agrees that Toyota produces better cars than most other manufacturers.
AttentionCD - DS	There is insufficient scientific evidence to support this claim.	AttentionCD - DS	Only Toyota produces vehicles that meet safety standards set by regulatory bodies worldwide.

When was the last time anyone was on the moon?		Who directed "Return of the Jedi"?		Which body system differentiates a male from a female?	
Greedy	1969	Greedy	Irvin Kershner	Greedy	Nervous System
DoLa	1969	DoLa	Irvin Kershner	DoLa	Nervous System
PruneCD	1972	PruneCD	Irvin Kershner	PruneCD	Nervous System
DeCoRe	1972	DeCoRe	Irvin Kershner	DeCoRe	Bone marrow
AttentionCD - JSD	1969	AttentionCD - JSD	Richard Marquand	AttentionCD - JSD	Nervous System
AttentionCD - DS	1969	AttentionCD - DS	Irvin Kershner	AttentionCD - DS	Reproductive

Figure 7: Qualitative analyses of different CD techniques. Model: Llama-3.2-1B-Instruct with the best hyperparameter for each CD technique. The right response is marked in **green**.

decoding when the underlying base model lacks sufficient factual knowledge regarding the input prompt (or the user’s question).

8 Qualitative Analysis

We present a brief Qualitative Analysis of different CD techniques in Figure 7.

In comparison with greedy decoding, CD techniques help improve the factuality of LLM responses. For example, when greedy decoding abstains from answering the question *"What happens to you if you eat watermelon seeds?"*, CD techniques, including PruneCD, DeCoRe, and AttentionCD-DS, respond that there are no supporting claims for the user question.

CD techniques also help in generating more factual responses from LLMs. For example, when asked *"When was the last time anyone was on the moon?"*, PruneCD and DeCoRe successfully generated the right answer *"1972"* when all other techniques failed, including greedy decoding. We present this example also as a remark that there are cases where AttentionCD variants fail, but DeCoRe, which masks only *retrieval heads*, can generate more factual responses.

For the question, *"Who directed 'Return of the Jedi'?"*, among all the techniques, only AttentionCD-JSD generated the right response. So is the case with the question, *"Which body system differentiates a male from a female?"*. Here, only AttentionCD-DS generated the right response. These examples show how AttentionCD variants

can generate factual responses when all other CD techniques and greedy decoding fail.

We also note that there are cases where neither greedy decoding nor CD techniques provide a truthful response to the user’s question, as shown by the example, *"Who would everyone agree makes the best cars?"*. This shows that if the model has little to no knowledge to answer a user question reliably (or, in this case, a significant bias towards non-factual information), the employment of CD techniques can yield no significant advantage over greedy decoding.

9 Conclusion

In this work, we introduced AttentionCD, a decoding strategy designed to mitigate factuality hallucinations in LLMs. Unlike other layer-based contrastive decoding strategies, AttentionCD masks task-specific attention heads from different decoder layers to build robust contrastive signals for contrastive decoding. We demonstrated that these heads can be discovered for a particular task using different task-specific indicators, including Jensen-Shannon Divergence and Downstream Performance Metrics. Our approach performs on par with the existing state-of-the-art layer-based CD techniques in different closed-book QA tasks. Promising avenues for future research include extending head-based CD techniques to address faithfulness hallucinations and evaluating their generalizability to multimodal and multilingual settings.

Limitations

Similar to other CD strategies, AttentionCD is limited by the parametric knowledge of the base LLMs. The technique can not work if the language model has no knowledge to answer the input question. Additionally, since the approach does not rely on the input context differences, it reduces its efficacy in reducing faithfulness hallucinations and knowledge conflicts. In such cases, we recommend resorting to CD strategies leveraging context information, for example [Shi et al. \(2023\)](#) and [Zhao et al. \(2024\)](#).

AttentionCD also suffers from an extensive hyperparameter search similar to other CD strategies. This can be partially reduced by developing a more robust dynamic CD temperature as suggested by [Gema et al. \(2025\)](#); [Lee et al. \(2025\)](#).

Further, we acknowledge that the method is tested only on Llama-based models, one Mistral model, and a set of academic datasets. We defer extension of the approach to other model families and real-life datasets to future work.

In this work, we limited the scope of the study to masking attention heads alone and showing that this can be effective in reducing factuality hallucinations. However, as future work, we find potential in extending the approach to ablating other model specifics, including MLP layers and residual streams.

Acknowledgments

This work was funded by the European Union and the German state of North Rhine-Westphalia within the following project of the European Regional Development Fund (EFRE-20801074): LLM4KMU - Optimized use of open source large language models in SMEs.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. [HalluLens: LLM Hallucination Benchmark](#). *arXiv preprint. ArXiv:2504.17550 [cs]*.
- 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 [cs]*.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. [DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models](#). *ArXiv:2309.03883 [cs]*.
- 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, Florence, Italy. Association for Computational Linguistics.
- Aryo Pradipta Gema, Chen Jin, Ahmed Abdulaal, Tom Diethe, Philip Alexander Teare, Beatrice Alex, Pasquale Minervini, and Amrutha Saseendran. 2025. [DeCoRe: Decoding by contrasting retrieval heads to mitigate hallucinations](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10003–10039, Suzhou, China. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Ping Guo, Yubing Ren, Yue Hu, Yanan Cao, Yunpeng Li, and Heyan Huang. 2024. Steering large language models for cross-lingual information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 585–596.
- 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 Comput. Surv.*, 55(12).
- Zhuoran Jin, Pengfei Cao, Hongbang Yuan, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2024. [Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models](#). *arXiv preprint. ArXiv:2402.18154 [cs]*.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Geonhee Kim, Marco Valentino, and Andre Freitas. 2025. [Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095, Vienna, Austria. Association for Computational Linguistics.

- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Hakyung Lee, Subeen Park, Joowang Kim, Sungjun Lim, and Kyungwoo Song. 2025. [Uncertainty-Aware Contrastive Decoding](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26376–26391, Vienna, Austria. Association for Computational Linguistics.
- Chaozhuo Li, Pengbo Wang, Chenxu Wang, Litian Zhang, Zheng Liu, Qiwei Ye, Yuanbo Xu, Feiran Huang, Xi Zhang, and Philip S. Yu. 2025. [Loki’s Dance of Illusions: A Comprehensive Survey of Hallucination in Large Language Models](#). ArXiv:2507.02870 [cs].
- Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. [Contrastive Decoding: Open-ended Text Generation as Optimization](#). ArXiv:2210.15097 [cs].
- Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*.
- 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.
- Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. [DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. [On faithfulness and factuality in abstractive summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. [Are Sixteen Heads Really Better than One?](#) In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Sean O’Brien and Mike Lewis. 2023. [Contrastive Decoding Improves Reasoning in Large Language Models](#). ArXiv:2309.09117 [cs].
- Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.
- Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen-tau Yih. 2023. [Trusting Your Evidence: Hallucinate Less with Context-aware Decoding](#). ArXiv:2305.14739 [cs].
- Sree Harsha Tanneru, Dan Ley, Chirag Agarwal, and Himabindu Lakkaraju. 2024. On the difficulty of faithful chain-of-thought reasoning in large language models. In *Trustworthy Multi-modal Foundation Models and AI Agents (TiFA)*.
- Team Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Elena Voita, David Talbot, Fedor Moiseev, Rico Senrich, and Ivan Titov. 2019. [Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned](#). *arXiv preprint*. ArXiv:1905.09418 [cs].
- 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.
- Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. 2024. Retrieval head mechanistically explains long-context factuality. *arXiv preprint arXiv:2404.15574*.
- Yunzhi Yao, Ningyu Zhang, Zekun Xi, Mengru Wang, Ziwen Xu, Shumin Deng, and Huajun Chen. 2024. Knowledge circuits in pretrained transformers. *Advances in neural information processing systems*, 37:118571–118602.
- Byeongho Yu, Changhun Lee, Jun-gyu Jin, and Eunhyeok Park. 2025. [PruneCD: Contrasting Pruned Self Model to Improve Decoding Factuality](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32462–32473, Suzhou, China. Association for Computational Linguistics.

Hongyi Yuan, Keming Lu, Fei Huang, Zheng Yuan, and Chang Zhou. 2024. [Speculative Contrastive Decoding](#). *arXiv preprint*. ArXiv:2311.08981 [cs].

Zheng Zhao, Emilio Monti, Jens Lehmann, and Haytham Assem. 2024. [Enhancing Contextual Understanding in Large Language Models through Contrastive Decoding](#). *arXiv preprint*. ArXiv:2405.02750 [cs].

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. [Attention Heads of Large Language Models: A Survey](#). *arXiv preprint*. ArXiv:2409.03752 [cs].

A Hyperparameter for Figure 1

- Model: Llama-3.2-1B-Instruct
- Repetition Penalty: For all decoding strategies, we use 1.2.
- CD Temperature: For all CD strategies (except DeCoRe), we set a static CD Temperature of 1.0. DeCoRe uses a dynamic CD Temperature guided by the entropy of base model token probabilities.
- We set DoLa layers to high.
- For PruneCD, we use $k=3$ for the number of layers to prune.
- For DeCoRe, we mask top-1% retrieval heads.
- For AttentionCD, we mask top-1% of attention heads from the High Layers (7-16) inferred from the JSD indicator.

B An overview of specialized attention heads

Different attention heads in the transformer architecture specialize in different functions (Zheng et al., 2024). Recent studies have explored the existence of such specialized heads and their significance at different phases of an NLP task. For example, consider a QA task where the model should recall, assess, reason, and generate an answer to the user’s question. There are *memory* (Jin et al., 2024) and *constant heads* (Lieberum et al., 2023), while the language model attempts to recall knowledge from its parameters. When the model processes in-context information *retrieval* (Wu et al., 2024), *context* (Jin et al., 2024) and *mover heads* (Yao et al., 2024) play a significant role. Similarly, *induction* (Olsson et al., 2022) and *inhibition heads*

(Kim et al., 2025) help the models while they attempt reasoning. Finally, when the model generates a response, the *faithfulness* (Tanneru et al., 2024) and *coherence heads* (Guo et al., 2024) play a crucial role. Discovering these specialized heads is often a non-trivial task, subject to the availability of specific datasets aiding this discovery and the accompanying computational overhead with larger language models.

In this work, we discover attention heads that are particularly important for a downstream task. These task-specific attention heads could be a combination of many different functional heads discussed above.

C Datasets

For the evaluation of our method, we use the datasets as detailed below. We provide further details on the validation and test splits and the number of samples in Table 3.

- **TruthfulQA:** <https://github.com/sylinrl/TruthfulQA/blob/main/TruthfulQA.csv>
- **Natural Questions:** https://huggingface.co/datasets/google-research-datasets/nq_open
- **TriviaQA:** https://huggingface.co/datasets/mandarjoshi/trivia_qa-unfiltered.nocontext

D Hyperparameter Sweep

CD techniques involve careful tuning of their defined hyper-parameters. To ensure a fair evaluation, for each model, we perform a two-fold cross-validation on the validation splits of each dataset for all the techniques alike. This helps us to find the best set of hyper-parameters given the dataset, model, and CD technique. The final evaluation is then conducted on the test split with the help of these hyper-parameters.

Closely following Yu et al. (2025) and the respective works, we describe the hyper-parameters of each CD technique below.

- **DoLa:** (i) dola-layers: range of decoder layers to consider as candidate premature layers. We use either high layers - the upper half or low - the lower half of the decoder layer; (ii) cd-temperature: always set to 1.0

Dataset	Validation		Test	
	Split	# Samples	Split	# Samples
TruthfulQA	In-domain	150	-	790
Natural Questions	train	1000	validation	3610
TriviaQA	train	2000	validation	11313

Table 3: Validation and Test splits of evaluation datasets

- **ActD:** (i) info-layer: decoder layer used for calculating activations - {12, 14, 16} for Llama-3.2-1B-Instruct, {22, 24, 26, 28} for Llama-3.2-3B-Instruct and {26, 28, 30, 32} for Llama-3.1-8B-Instruct; (ii) entropy-scaler: scaling factor for adjusting the next token probability distribution - {0.2, 0.4, 0.6, 0.8}.
- **END:** (i) relative-top: threshold for applying plausibility constraint while contrastive decoding - {0.1, 0.01, 0.001}; (ii) entropy-scaler: scaling factor for adjusting the next token probability distribution - for TruthfulQA-MC, Natural Questions, and TriviaQA {0.25, 0.375, 0.5} and for TruthfulQA-Gen {1.0, 2.0, 3.0}.
- **PruneCD:** (i) k-layers: number of layers pruned - [1, 4] for Llama-3.2-1B-Instruct, [3, 6] for Llama-3.2-3B-Instruct, and [3, 8] for Llama-3.1-8B-Instruct; (ii) cd-temperature: [1.0, 1.5] for TruthfulQA and [0.1, 0.3] for other datasets.
- **DeCoRe:** (i) k%-heads: percentage of retrieval heads to mask globally from the model; (ii) cd-temperature: uses an entropy-guided dynamic CD temperature.
- **AttentionCD:** (i) k%-heads: percentage of task-specific heads to mask from each layer of the model; (ii) layers-to-ablate: range of decoder layers where attention heads should be masked. We use either High layers - the upper half, Low - the lower half, or All decoder layers in the model; (iii) cd-temperature: uses an entropy-guided dynamic CD temperature for TruthfulQA-MC, {1.0, 1.5, 2.0} for TruthfulQA-Gen, a static {0.1} for Natural Questions and {0.1, 0.5} for TriviaQA.

Compute Resources We used NVIDIA H200 GPUs for the discovery of task-specific attention

heads, the hyperparameter sweep, and the evaluation runs.

E Task-specific Attention Heads

To study the overlap of attention heads, we compute Average Intersection over Union on the sets of heads. Given the sets of heads A and B , we compute Average IoU between them as follows:

$$\text{Average IoU} = \frac{1}{N} \sum_{i=1}^N \frac{|A_i \cap B_i|}{|A_i \cup B_i|}$$

E.1 How different are the heads discovered by the same indicators for different validation samples?

We discover attention heads for each dataset following the indicators JSD Delta and DS Delta separately for (i) different numbers of validation samples and (ii) different splits of the same number of validation samples. In Figure 8, the **orange** line indicates the overlap between JSD heads discovered by different numbers of validation samples, and the **blue** line indicates the overlap between different JSD heads discovered by different splits of the same number of validation samples. For this case, across datasets and models, we observe a consistently high overlap of over 75% between heads discovered by JSD both when we use different numbers of validation samples and different folds of validation samples. This indicates that our head selection strategies are robust and result in similar resultant heads when the number and splits of validation samples change.

E.2 How different are the heads discovered by different indicators?

Further, we want to study the extent of overlap between the heads discovered following JSD Delta and DS Delta separately. The **green** line indicates the overlap between JSD heads and DS heads discovered by means of the same validation samples. In this case, there is little overlap between JSD

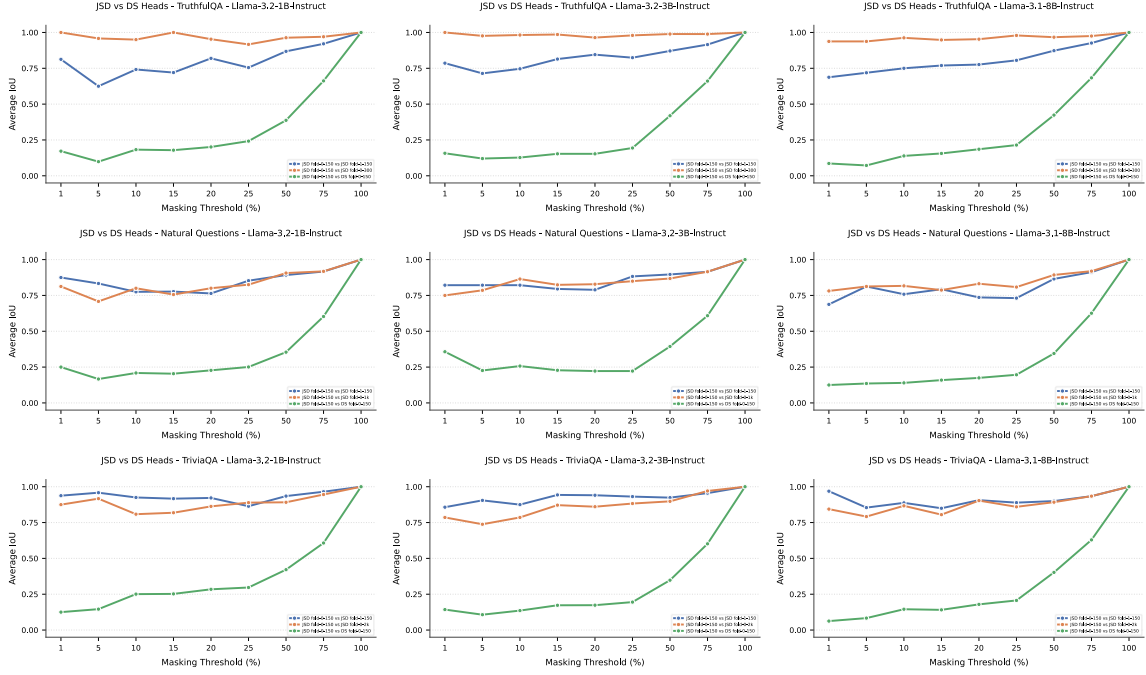


Figure 8: Comparison of head overlap across indicators and validation samples. We measure the average IoU between heads discovered for (i) the same indicator with different validation samples (blue), (ii) the same indicator with a different number of validation samples (orange), and (iii) different indicators with same validation samples (green).

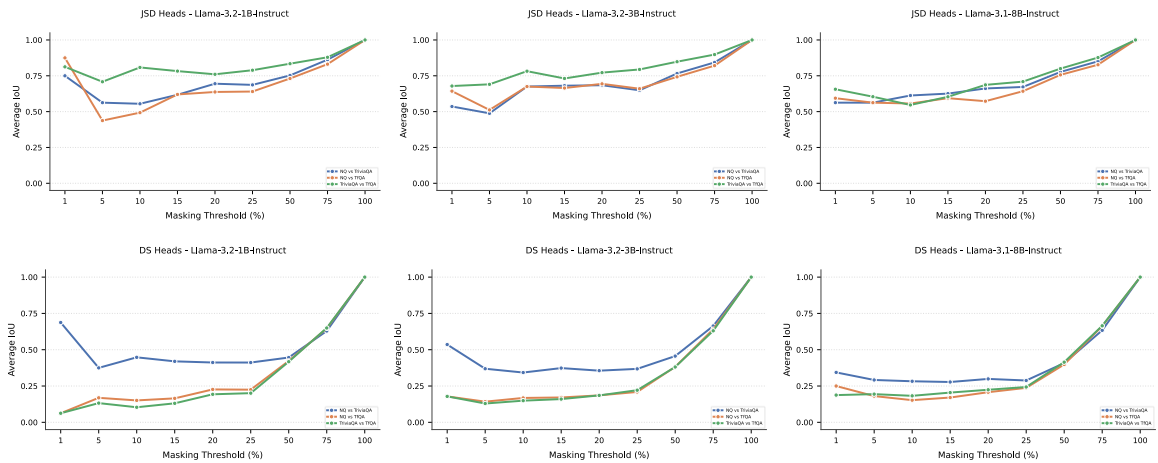


Figure 9: Comparison of head overlap across datasets. We measure the average IoU between heads discovered for different dataset pairs: Natural Questions vs. TriviaQA (blue), Natural Questions vs. TruthfulQA (orange), and TriviaQA vs. TruthfulQA (green). The top panel shows JSD heads, and the bottom panel shows DS heads.

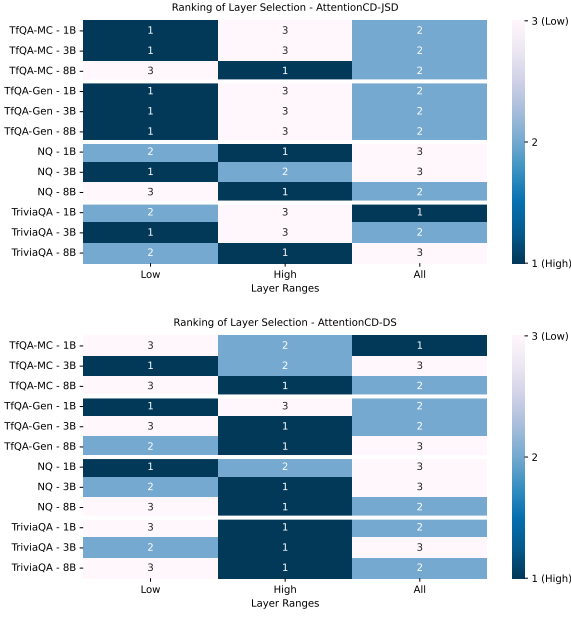


Figure 10: Ranking of ablated Layer Ranges. Rank 1 is the highest (brightest) and Rank 3 is the lowest (lightest). Legend: *NQ* - Natural Questions, *TQA* - TruthfulQA, *1B* - Llama-3.2-1B-Instruct, *3B* - Llama-3.2-3B-Instruct, *8B* - Llama-3.1-8B-Instruct.

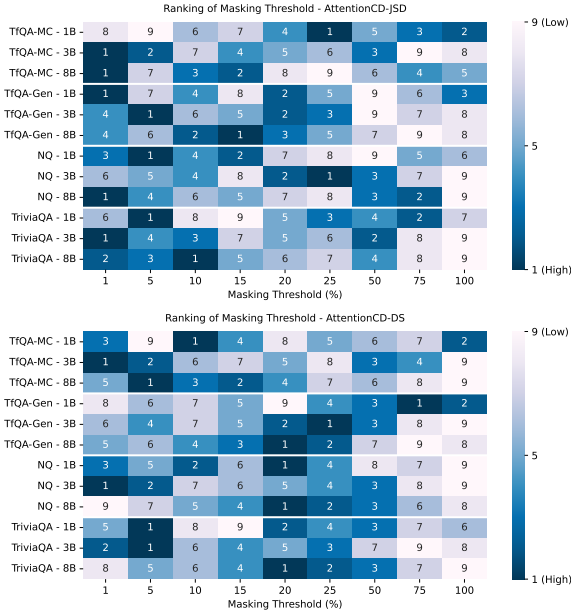


Figure 11: Ranking of Masking Threshold (%). Rank 1 is the highest (brightest) and Rank 9 is the lowest (lightest). Legend: *NQ* - Natural Questions, *TQA* - TruthfulQA, *1B* - Llama-3.2-1B-Instruct, *3B* - Llama-3.2-3B-Instruct, *8B* - Llama-3.1-8B-Instruct.

Where did fortune cookies originate?		
Greedy: Fortune cookie originated in China during the Qing dynasty (1644-1912).		
Threshold (%)	Masking Heads in Low Layers	Masking Heads in High Layers
1	Fortune cookie originated in China during the Qing dynasty (1644-1912).	Fortune cookie originated in China during the Qing dynasty (1644-1912).
10	Fortune cookie originated in China during the Qing dynasty (1644-1912).	Fortune cookie originated in China during the Qing dynasty (1644-1912).
25	Fortune cookie originated in China during the Qing dynasty (1644-1912).	Fortune cookies originated in San Francisco, California.
50	Fortune cookie originated in China during the Qing dynasty (1644-1912).	Fortune cookies originated in San Francisco, California.
100	Fortune cookie originated in China during the Qing dynasty (1644-1912).	Fortune cookies originated in San Francisco, California.

Figure 12: Comparison of how masking attention heads in different layer ranges functions at different masking thresholds. The right responses are marked in green. Model: Llama-3.2-1B-Instruct, CD Temp: 0.1.

heads and DS heads. Average IoU in this case crosses 50% only at higher thresholds of over 50% across datasets and models. This is a clear indication that the heads discovered by different indicators are sufficiently different.

E.3 How different are the heads discovered for different datasets?

In Figure 9, we study how the heads discovered for different datasets are different from one another. In general, the overlap between JSD heads is higher (crossing over 50% Average IoU at a threshold of over 10%) than the corresponding DS heads (crossing over 50% Average IoU at a threshold of over 50%). We argue this as a direct result of the stronger downstream-task influence on the DS heads selection compared with the weaker task-specific influence on JSD heads. However, we note that for similar datasets, Natural Questions and TriviaQA, the heads overlap is higher, indicated by the blue line, than when compared to (Natural Questions, TruthfulQA) pair indicated by the orange line or (TriviaQA, TruthfulQA) pair indicated by the green line.

E.4 Cost of Head Discovery

For a decoder model with n layers and h heads per layer, AttentionCD demands a head discovery cost of around $O(n^2 * h)$. As an indicator of the head discovery cost, we report the cost for discovering heads for the model Llama-3.2-1B-Instruct for the Truthful-QA dataset with 150 validation samples on an NVIDIA H200 GPU below:

- **AttentionCD-JSD** takes around 6 min.
- **AttentionCD-DS** takes around 1 h.

This costly head discovery stage can be further optimized using (i) more efficient batching

Model	Method	TruthfulQA MC			TruthfulQA Gen			TriviaQA		NQ	
		MC1 (↑)	MC2 (↑)	MC3 (↑)	Truth% (↑)	Info% (↑)	T*I% (↑)	EM (↑)	F1 (↑)	EM (↑)	F1 (↑)
Mistral-7B-Instruct-v0.3	Greedy	47.09	65.13	35.99	78.23	76.46	59.81	65.62	64.90	34.76	33.21
	DoLa	44.30	63.63	32.02	78.10	91.14	71.18	65.46	64.41	35.04	32.46
	ActD	43.54	53.00	31.82	77.09	88.23	68.01	65.84	65.05	35.10	33.25
	END	46.46	64.00	36.31	77.97	90.63	70.67	65.57	64.79	35.01	33.19
	PruneCD	48.10	60.38	35.70	85.06	90.89	77.31	66.61	65.85	35.46	34.00
	AttentionCD-JSD	44.18	60.18	32.08	81.65	96.20	77.85	66.32	65.41	35.57	33.45
	AttentionCD-DS	44.12	59.84	31.85	80.00	95.44	75.44	66.50	65.77	35.48	33.39

Table 4: Performance of Greedy , Layer-based and Head-based CD techniques with Mistral-7B-Instruct-v0.3. We **bold** the best scores.

at inference, and (ii) an optimized head discovery following the layer selection strategy adopted by PruneCD. Given a layer range, one could reduce the number of layers to ablate using the SLEB algorithm and then perform head discovery on these reduced sets. We leave these optimizations for future work.

F Experiments with non-Llama models

To show how AttentionCD generalizes to non-Llama models, we ran experiments on Mistral-7B-Instruct-v0.3. The results are reported in Table 4. The trends look similar to the Llama variants, where AttentionCD falls short in the multiple-choice QA task with TruthfulQA-MC and performs on par with other techniques for open-ended generation tasks. For this model, we observe AttentionCD-JSD performing better than AttentionCD-DS, unlike the trend seen with Llama models.

G Detailed Validation Results

G.1 Layer Ablation

Based on our validation runs across different tasks and models, we study the layer ranges that, when ablated, result in the best invasive strategy for AttentionCD. As shown in Figure 10, for AttentionCD-JSD, we see that Low regions of the models result in the best contrastive signal, and for AttentionCD-DS, the High regions of the models result in the best contrastive signals when ablated.

G.2 Masking Threshold

We study how many attention heads should be masked from each decoder layer to result in reliable contrastive signals for AttentionCD with the help of our hyperparameter $k\%$ -heads. We range this value between 1% (where we mask only the

top-1% heads per layer of the model) and 100% (where we mask all the heads in the layer). From Figure 11, it is evident that for AttentionCD-JSD, masking top-1% heads per layer generally results in better contrastive signals, and for AttentionCD-DS, masking top-20% heads results in better signals.

We show in Figure 12 how masking attention heads in different layer ranges functions at different masking thresholds. For the question "Where did fortune cookies originate?", greedy decoding generates an incorrect response, falling under the common misconception of fortune cookies having originated in China. This is also the case with AttentionCD when we mask heads only from the Low layers of the model. The result does not get better even when we mask top-100% (all attention heads) from these layers. However, when we mask heads from High layers, we can see that AttentionCD generates the right response starting from top-25% heads masked.

G.3 CD Temperature

CD temperature is crucial in controlling the contribution of amateur logits from the ablated model. Prior work introduced static CD temperature as a hyperparameter (Chuang et al., 2024; Yu et al., 2025), and a few others resort to a dynamic CD temperature setting (Gema et al., 2025; Yuan et al., 2024). We study the effect of both settings and choose the best according to the dataset.

In Table 5, we compare and contrast setting a static versus a dynamic CD temperature. For the open-ended generation task with Natural Questions, we set a static CD temperature of 0.1, and for the multiple-choice QA task with TruthfulQA-MC, we use 1.0. Following Gema et al. (2025) and Yuan et al. (2024), we compare this setting against a dynamic CD temperature set per sample based on the entropy of the base model logits. It is evident

Model	Method	TruthfulQA MC			NQ	
		MC1 (↑)	MC2 (↑)	MC3 (↑)	EM (↑)	F1 (↑)
	Greedy	26.84	45.92	21.63	18.81	17.33
Llama-3.2-1B-Instruct	JSD + Dynamic CD	28.10	47.88	21.98	17.53	15.98
	JSD + Static CD	27.22	44.20	20.74	19.00	17.15
	DS + Dynamic CD	29.11	48.47	22.66	16.12	15.19
	DS + Static CD	28.99	44.71	21.40	18.84	17.41

Table 5: Comparison of setting a static CD temperature with a dynamic CD temperature.

from the results that the multiple-choice QA task benefits from a dynamic CD temperature, and the open-ended generation task benefits more from a static CD temperature. Motivated by this observation, we conducted our final evaluations, setting a dynamic CD temperature only for the multiple-choice QA task.

G.4 Repetition Penalty

Method	Repetition Penalty	TruthfulQA Gen		
		%Truth (↑)	%Info (↑)	%T*I (↑)
AttentionCD-JSD	1.2	56.08	95.57	51.65
	1.4	55.70	95.70	51.39
	1.6	55.57	95.57	51.14
	1.8	55.95	95.44	51.52
	2.0	55.70	95.70	51.52
AttentionCD-DS	1.2	53.54	94.56	48.23
	1.4	54.18	95.95	50.13
	1.6	55.06	96.58	51.65
	1.8	55.06	96.46	51.65
	2.0	55.44	96.46	52.03

Table 6: Comparison of AttentionCD variants at different repetition penalties. Model: Llama-3.21-1B-Instruct.

Repetition penalty is a decoding parameter that penalizes repetitive token generation by the model while ensuring coherence in the responses. Prior work (Chuang et al., 2024; Yu et al., 2025) set a repetition penalty of 1.2 to control the model generation after contrastive decoding. However, Chuang et al. (2024) also observes that techniques like Li et al. (2023) benefit from different values of repetition penalties. Thus, we study the influence of repetition penalties on AttentionCD variants.

In Table 6, we observe that AttentionCD-JSD benefits very little from setting a higher repetition penalty. However, AttentionCD-DS performs better at a higher repetition penalty of 2.0. While this finding helps us in further refining AttentionCD-DS

responses individually, to ensure a fair evaluation with other CD techniques, we ensured to set a common repetition penalty of 1.2 across CD techniques for our final evaluation.

H Task-specific Heads vs. Retrieval Heads

Model	Method	NQ	
		EM (↑)	F1 (↑)
Llama-3.2-1B-Instruct	Greedy	18.81	17.33
	Global JSD Heads	18.89	17.33
	Global DS Heads	18.59	17.27
	Global Retrieval Heads	18.70	17.13
Llama-3.2-3B-Instruct	Greedy	30.83	28.48
	Global JSD Heads	31.83	29.20
	Global DS Heads	31.94	29.15
	Global Retrieval Heads	31.80	29.05
Llama-3.1-8B-Instruct	Greedy	36.98	34.90
	Global JSD Heads	37.20	34.78
	Global DS Heads	37.31	34.92
	Global Retrieval Heads	37.20	34.96

Table 7: Comparison of global task-specific heads against retrieval heads

While DeCoRe (Gema et al., 2025) masks only *retrieval heads* to build contrastive signals for contrastive decoding, AttentionCD masks task-specific attention heads to obtain the contrastive signal. In their original work, Gema et al. (2025) report evaluation results by (i) masking 10-100 *retrieval heads* and (ii) using different CD temperature settings — the one using dynamic CD temperature based on the entropy of base model logits performing the best, in general. Our work is different from theirs since we mask k%-heads per layer to study which layer ranges should be ablated for the maximal invasive strategy for Head-based CD.

To draw better insights into the different types of heads discussed, we evaluated both approaches under similar conditions. For this, (i) we set a

Method	TruthfulQA MC			TruthfulQA Gen			TriviaQA		NQ		Δ Average	Rank
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ			
	MC1	MC2	MC3	Truth%	Info%	T*I%	EM	F1	EM	F1		
	($\uparrow$)	($\uparrow$)	($\uparrow$)	($\uparrow$)	($\uparrow$)	($\uparrow$)	($\uparrow$)	($\uparrow$)	($\uparrow$)	($\uparrow$)		
Greedy	–	–	–	–	–	–	–	–	–	–	–	–
DoLa	2.01	-3.54	-6.12	-17.52	38.54	38.72	-0.15	-0.65	0.77	-2.21	5.0	4
ActD	-3.62	1.36	-3.52	-17.12	33.30	31.33	1.03	0.90	1.27	-0.01	4.5	6
END	1.41	1.06	3.19	-1.80	5.70	8.26	0.24	0.11	0.62	-1.62	1.7	7
PruneCD	7.76	2.39	2.86	-0.98	41.64	54.41	1.29	1.08	2.01	1.10	11.4	1
DeCoRe	8.87	3.56	0.70	-19.53	38.32	38.53	-4.87	-5.87	-5.70	-7.53	4.6	5
AttentionCD-JSD	4.78	1.92	-1.08	-16.75	43.58	47.12	1.18	1.32	2.05	0.58	8.5	3
AttentionCD-DS	7.20	2.67	0.94	-11.44	48.84	56.10	2.01	1.20	1.69	1.40	<u>11.1</u>	<u>2</u>

Table 8: Delta performance improvement of Layer-based and Head-based CD techniques over Greedy decoding. We **bold** the best scores and underline the second-best scores. Ranks are calculated based on the final average delta scores.

static CD temperature of 0.1 for Natural Questions, (ii) masked heads globally from the model (and not from each layer), and (iii) used a common threshold of top-10%. We report the results in Table 7. We observe that in this controlled setting, masking task-specific heads globally outperforms masking only *retrieval heads* globally.

I Delta Performance Improvement

To find the overall performance improvement of different CD techniques over greedy decoding, firstly, we find the per-model per-technique delta improvement over greedy decoding for each metric. We then average these delta improvement scores across models for a given technique and a given metric. These values are populated in Table 8 and denoted as “ Δ <metric-name>”. To find the final rank as an indicator of performance improvement over greedy decoding, we average the per-metric delta scores across the different tasks, given as “ Δ Average”. Ranks are then computed based on these average delta scores.

Based on the average delta scores, PruneCD performs the best overall, ranked # 1 with around +11.4% performance improvement over greedy decoding. AttentionCD-DS ranks # 2 with an average of +11.1% improvement over greedy decoding. AttentionCD-JSD follows shortly with an average improvement of over +8.5% over greedy decoding.