LISTENING TO THE WISE FEW: SELECT-AND-COPY ATTENTION HEADS FOR MULTIPLE-CHOICE QA

Anonymous authors
Paper under double-blind review

ABSTRACT

A standard way to evaluate the abilities of LLM involves presenting a multiple-choice question and selecting the option with the highest logit as the model's predicted answer. However, such a format for evaluating LLMs has limitations, since even if the model knows the correct answer, it may struggle to select the corresponding letter simply due to difficulties in following this rigid format. To address this, we introduce new scores that better capture and reveal model's underlying knowledge: the Query-Key Score (QK-score), derived from the interaction between query and key representations in attention heads, and the Attention Score, based on attention weights. These scores are extracted from specific *select-and-copy* heads, which show consistent performance across popular Multi-Choice Question Answering (MCQA) datasets. Based on these scores, our method improves knowledge extraction, yielding up to 16% gain for LLaMA2-7B and up to 10% for larger models on popular MCQA benchmarks. At the same time, the accuracy on a simple synthetic dataset, where the model explicitly knows the right answer, increases by almost 60%, achieving nearly perfect accuracy, therefore demonstrating the method's efficiency in mitigating MCQA format limitations. To support our claims, we conduct experiments on models ranging from 7 billion to 70 billion parameters in both zero- and few-shot setups.

1 INTRODUCTION

Questions with multiple answer options are a common form of benchmarks evaluating question answering (Hendrycks et al., 2021), common sense (Zellers et al., 2019), reading comprehension (Huang et al., 2019), and other abilities of large language models. In multiple choice question answering tasks (MCQA), the model is provided with the question and multiple answer options, e.g. *"Question: How many natural satellites does the Earth have? Options: A. 0. B. 1. C. 2. D. 3. E. None of the above. F. I don't know."* Sometimes the context that might be helpful to give the answer is added before the question, such as a paragraph or a dialogue for reading comprehension or some common sense reasoning. The model is asked to output the letter denoting the correct answer option. This format is similar to certain real-life students' exams and shares some benefits with them: it is straightforward to evaluate, using automated tools.

On the other hand, for LLMs, especially smaller ones, understanding and adhering to a multiple-option format is not always trivial. The model's performance on a given multiple-option dataset depends not only on the ability to solve the task itself but also on its in-context learning or instruction-following capabilities. The model may produce correct answers with formatting issues, which hinders the automatic evaluation of the MCQA task. Consequently, some works delegate answer evaluation to another LLM instead of relying on exact string comparison (Wang et al., 2024). When assessing the logits of the model for options, LLMs can follow shallow patterns such as options distribution. Some LLMs are inclined to prefer the answer option "A", while others tend to choose "D" (Zheng et al., 2024a). All of these issues demonstrate pitfalls in the current MCQA evaluation process, especially for smaller LLMs.

However, the model's inability to follow the task format does not imply a lack of actual 'knowledge' regarding the correct answer. In this work, we show that while small LLMs generally perform poorly on MCQA benchmarks, their intermediate attention states can often provide better insights. Specifically, we introduce the method that uses the queries and keys within individual attention

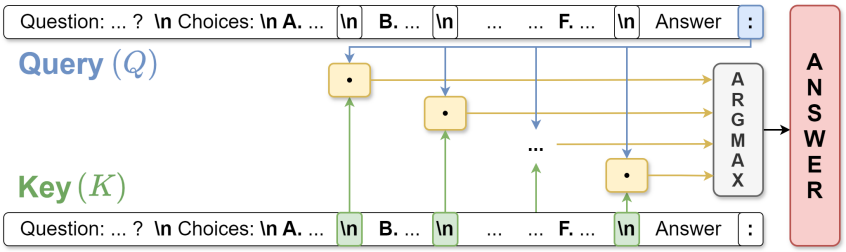


Figure 1: Our method calculates the Query-Key score between the end-of-line token of an answer option and the last token of the prompt for the designated head, from which we derive the answer.

heads to select an answer. We identify certain *select-and-copy* heads that can choose the option with semantically relevant information and transfer further its representation. Our findings suggest that LLMs process MCQA tasks more effectively in the middle layers but tend to revise this information in the later layers, leading to reduced performance. Our method belongs to the class of "white-box" techniques, meaning that we use the internal representations of LLMs to extract solutions for given tasks.

We identify the principal elementary algorithmic operation performed by pretrained Transformer models when answering multiple-choice questions. This task is complex, requiring the model to first compute a representation of semantic information contained in both the question and the options inside such specific heads. After completing this task, the model selects the most appropriate option using the query-key alignment mechanism, see section 3, and then copy and outputs the option. Based on this intuition, we propose to identify the heads in the model that perform this *select-and-copy* operation on the aggregated embeddings of the possible answers. Our results reveal the presence of such heads in all models we examined, ranging from 7 billion to 70 billion parameters. Remarkably, the best few performing heads are the same for different datasets. Moreover, the answers produced by these heads are significantly more accurate than the final output of the model, particularly in zero-shot scenarios.

Our contributions are as follows: (1) We demonstrate the presence of *select-and-copy* heads in LLMs of the wide range of 7-70B parameters, performing option selection operation for MCQA task; (2) We introduce QK-score, along with attention score, the option scoring methods based on key and query representations derived from such heads. This scoring leads to 9-16% improvement of the accuracy with task-specific heads; (3) We demonstrate that our method is more stable than the baselines to option permutations, renaming and also when supplementary options, like "I don't know", are added; (4) Our results provide further support for the hypothesis observed in other papers, e.g. Li et al. (2023b); Stolfo et al. (2023), that the representations of the semantic meaning of a phrase are encoded in certain heads in the query, key, and value vectors of the phrase's last tokens—namely, the end-of-sentence punctuation token or the end-of-line token; (5) We study the attention patterns of *select-and-copy* heads and their behavior under various conditions. This is a step towards better understanding how the LLMs work in general.

2 Related work

Question answering datasets are the standard way to measure capabilities of the Large Language Models to retain the knowledge, understand given texts, and perform reasoning. The results of such testing can be seen in a number of technical reports on new LLMs, such as LLaMA2 and LLaMA3, gpt-4o, or Claude 3 Opus (Touvron et al., 2023; Dubey et al., 2024; OpenAI, 2024; Anthropic, 2024). Also, MCQA tasks are often included in benchmarks for these models, as this task is easy for evaluation (Ye et al., 2024; Pal et al., 2022).

There are several setups for MCQA task. The first one, multiple choice prompting (MCP) is when we present the question and multiple answer options to model. It has many advantages over Cloze Prompting (CP), when a model is asked to complete partial inputs with a single probable word or phrase (Robinson & Wingate, 2023). While in CP normalized answer probabilities are used for evaluation, in MCP we can use the probabilities of the options' single tokens as a proxy. However,

recent works highlight some issues in evaluating models on MCQA tasks. Gupta et al. (2024) and Pezeshkpour & Hruschka (2024) show that permutation of the contents for options can significantly affect the accuracy using MCP. Similarly, Zheng et al. (2024a) describe selection bias for different LLMs and propose a debiasing method PriDe to boost the accuracy.

In our work, we investigate the inner mechanisms of LLMs, especially the role of attention heads in MCQA tasks. Functional roles of attention heads were analysed for transformer-based models from the very beginning of encoder-only models (Jo & Myaeng, 2020; Pande et al., 2021), and nowadays even more detailed approaches were developed for decoder-only models in the common track of mechanistic interpretability (Elhage et al., 2021; Olsson et al., 2022; Bricken et al., 2023). For example, *induction heads* identified by Elhage et al. (2021) play an important role in in-context learning (Olsson et al., 2022; Von Oswald et al., 2023), indirect object identification (Wang et al., 2023) and overthinking (Halawi et al., 2024). Additionally, there is a number of research connecting theoretically constructed networks with real pretrained language models, revealing elements such as constant heads (Lieberum et al., 2023), negative heads (Yu et al., 2024), and content gatherer heads (Merullo et al., 2024), among others. For more information on mechanistic interpretability and attention heads, we refer to Rai et al. (2024) and Zheng et al. (2024b).

In our work, we focus on *select-and-copy heads* that are used to select the right option for the MCQA task. The special case of such heads looking on option labels was mentioned in (Lieberum et al., 2023). However, we show that not only are other tokens more representative for MCQA, but they can be used to significantly increase accuracy compared to baseline.

Moreover, our experiments conclude that those heads that outperform the baseline on MCQA are located on the middle layers of LLM. It correlates with previous findings that many information is present in earlier layers, but is somehow lost or revised in later layers (Kadavath et al., 2022; Azaria & Mitchell, 2023; Liu et al., 2023; Zou et al., 2023; CH-Wang et al., 2024). These studies mainly focus on linear probes of hidden representations (Ettinger et al., 2016; Conneau et al., 2018; Burns et al., 2023), but we show that the disagreement between model output and inner structures can be captured in the level of query-key interactions and attention maps.

3 Attention as select-and-copy algorithm

In this section, we describe how attention mechanism can work as *select-and-copy* operation. Suppose we have a sequence of N token embeddings $\{\mathbf{x}_i\}_{i=1}^N$, which serve as an input to the corresponding attention head of transformer, each $\mathbf{x}_i \in \mathbb{R}^{d\times 1}$. In classical transformer architecture Vaswani et al. (2017), each attention head performs the transform of input embeddings:

$$\boldsymbol{o}_m = \sum_{n=1}^{N} a_{m,n}\boldsymbol{v}_n, \quad a_{m,n} = \frac{\exp\left(\frac{\boldsymbol{q}_m^\top \boldsymbol{k}_n}{\sqrt{d}}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\boldsymbol{q}_m^\top \boldsymbol{k}_j}{\sqrt{d}}\right)}, \tag{1}$$

where $\boldsymbol{q}_i = \boldsymbol{W}_q\boldsymbol{x}_i$, $\boldsymbol{k}_i = \boldsymbol{W}_k\boldsymbol{x}_i$, $\boldsymbol{v}_i = \boldsymbol{W}_v\boldsymbol{x}_i$ and $\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d_{model}\times d}$ are learned weight matrices. The resulting matrix $\boldsymbol{A} = \{a_{n,m}\}_{n,m=1}^N$ is stochastic, meaning that all its rows sum up to one. For decoder transformers, causal mask is applied to $\boldsymbol{A}$ before softmax: $a_{i,j} = 0, j > i$. Thus, from equation 1, for each token position k in decoder transformers we can write

$$\boldsymbol{o}_m = \sum_{n\le m} a_{m,n}\boldsymbol{v}_n \tag{2}$$

meaning that the m-th token of output embedding is the linear combination of values of the preceding tokens weighted by m-th row of the attention matrix $\boldsymbol{A}$. If all but one component here are close to zero, this transform can be considered as a conditional copy mechanism. Indeed, if $a_{m,j}$ is the only non-zero weight in the m-th row, then $a_{m,j} \approx 1$, and $\boldsymbol{o}_m \approx \boldsymbol{v}_j$ (by 2). Each token position from 0 to m can be considered as a cell storing the corresponding value vector; and attention weights $a_{m,0}, \dots, a_{m,m}$ are responsible for the *choice* which cell to copy to the m-th output.

Based on this, we came up with the idea of *select-and-copy* heads, which implement such copying mechanism. Namely, in this work, we are interested in finding heads in the model, which select the proper option and copy the information from it to the answer. In such heads, the attention of m-th row should be concentrated on a few selected tokens, where m is the output answer position.

In modern models, positional encoding information can be represented as the additional transform of queries and keys in Eq. 1. For example, in Rotary Position Embedding (RoPE) (Su et al., 2024) the rotation function $R_f(\cdot)$ is applied to them before taking dot product. Pre-softmax logit of the standard attention becomes $R_f(\boldsymbol{q}_m)^T R_f(\boldsymbol{k}_n) = R_g(\boldsymbol{q}_m, \boldsymbol{k}_n, m-n)$, introducing the dependency on the position shift $m-n$.

In this paper, we evaluate the efficiency of the answer options scoring derived from *select-and-copy* heads. We aim to choose heads which rely on options semantics rather than the position; to mitigate the effect of the relative position shift, we consider the QK-score, which does not use the positional shift when comparing the queries and keys (see details in the next section).

4 Approach

Consider some MCQA task with the corresponding dataset $\mathcal{D} = \mathcal{D}_{val} \cup \mathcal{D}_{test}$, where each instance represents the request to the model, consisting of prompt, question, and labelled answer options (Fig. 1). Given the request, the model should generate the label of the best option from the request.

To find the heads in the model that implement the described above option selection mechanism, we pick best-performing heads using $\mathcal{D}_{val}$ based on the accuracy there, and then evaluate their performance on much larger $\mathcal{D}_{test}$. If such heads are in fact fully responsible for option selection, the performance just on them should be at least comparable to the performance of the whole model. We prove this claim by experiments in Section 5.4. Another way to select such heads is proposed in Section 6; we demonstrate by attention maps analysis, that the best-performing heads indeed implement the option selection algorithm described above.

QK-score and Attention-Score. Given a data sample of MCQA task, we denote by q the question supported with context if applicable, by $o = \{o_1, o_2, ..., o_n\}$ the semantic content of the provided answer options, and the corresponding labels by $d = \{d_1, d_2, ..., d_n\}$ (e.g A/B/C/D); we believe that the labels are default-ordered. All these parts are concatenated to a string $q * d_1 * o_1 * \cdots * d_n * o_n *$, where $*$ stands for any kind of delimiters, usually punctuation marks or newline characters (Fig. 1). The model should estimate $P(d_i \mid q, d, o)$ – the probability of option d_i given the question q and contents of the answer o, concatenated with the answer options d.

Let $t_i, i \in \{1, 2, ..., n\}$ be the indices of tokens that incorporate knowledge about the corresponding answer options. We call them *option-representative tokens*. Choosing such tokens properly is important for the success of our algorithm. In most experiments, we use the end-of-line token after the i-th option content as t_i; other possible variants are presented in Fig. 2a. We study them in Sec. 6.

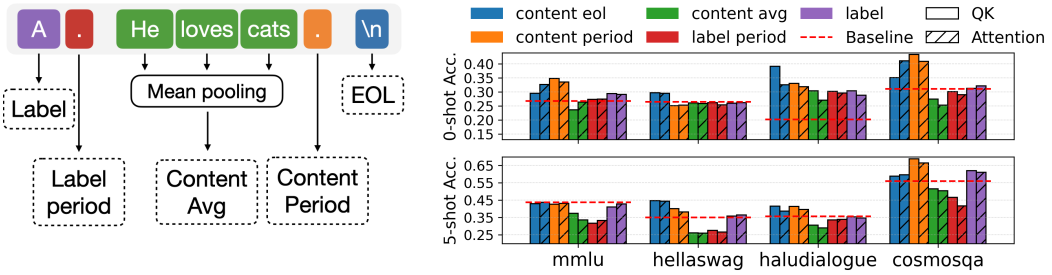
Let N be the length of the whole text sequence, and consider the head with index h from the layer l. Then, given (q, d, o), we can compute *QK-score* $S_{QK}^{(l,h)}(d_i)$ for option d_i (from query and key vectors), and *Attention-score* $S_{Att}^{(l,h)}(d_i)$ (from attention weights):

$$S_{QK}^{(l,h)}(d_i) = \boldsymbol{q}_N^{(l,h)\top} \boldsymbol{k}_{t_i}^{(l,h)}, \quad S_{Att}^{(l,h)}(d_i) = a_{N,t_i}^{(l,h)}, \quad i \in \{1, 2, ..., n\} \tag{3}$$

Our *QK-score* of i-th option is calculated as a dot product of the t_i-th key and the last query vector q_N (see Fig. 1). In *QK-score* we do not apply positional transformation, therefore it is not equal to the attention scores before softmax. The best token by *QK-score* does not necessarily correspond to the token with maximum attention, see Figure 8 for an example.

For each method, the prediction is straightforward: we take the option, for which the score gives maximum. By applying softmax function to scores we could also estimate the head and score specific probabilities for options.

Choosing the predicting heads. We do not aggregate heads predictions. Instead, we use the scores from the single best head, which is selected by the accuracy on the validation set $\mathcal{D}_{val}$. In Section 5.4, we report the results obtained from the best heads chosen separately for each dataset and each number of shots (i.e. number of examples provided in the prompt). In Section 6, we show that for each model, there exist universal heads working well on the most tasks and number of shots. Furthermore, we demonstrate that such universal heads can be found without access to labelled validation data.



(a) (b)

Figure 2: (a) Scheme for option-representative token types. (b) Performance of *QK-score* and *Attention-score* for different option-representative tokens on Llama2-7B base.

5 EXPERIMENTS

5.1 DATASETS

We experiment on four challenging real-world MCQA datasets from LLM benchmarks: **MMLU** (Hendrycks et al., 2021), **CosmosQA** (Huang et al., 2019), **HellaSwag** (Zellers et al., 2019) and **HaluDialogue**, which is a "dialogue" part of HaluEval (Li et al., 2023a). All of them consist of questions with four possible answer options and some additionally have a context to be used to give the answer. More details about each dataset can be found in Appendix A.1. Additionally, we introduce **Simple Synthetic Dataset** (SSD) created as a synthetic task in MCQA setting that will allow to estimate the ability of the model deal with the bare task format. Tasks from SSD do not require any factual knowledge from the model. The main version of this dataset contains questions of the form "Which of the following options corresponds to "`<word>`" ?" and contains 2.500 examples. Options include a word from the question and 3 random words, all mixed in a random order and marked by letters 'A'-'D'. Other variations of this dataset have another number of options, sampled and named by the same principle. These other versions are described in more details in Appendix H.

Finally, following Ye et al. (2024) in all five datasets we specially modified questions by adding two extra options "`E. None of the above.`" and "`F. I don't know.`" that are intended to aggregate the uncertainty of LLM. Despite adding these two options, there are *NO* questions for which 'E' or 'F' are correct answers. Examples from all datasets are listed in the Appendix A.2, as well as prompt formatting we used.

Following the previous approach by Zheng et al. (2024a), with fixed N-shot setup, we select $\mathcal{D}_{val}$ as 5% of $\mathcal{D}$ for each dataset that is dedicated to assessing each head's performance. Based on this evaluation, the best head is chosen and applied to other questions in the dataset.

5.2 BASELINES

The standard approach for MCQA is to use output probabilities from LLM for all options d_i to choose the predicted option $\hat{d}$:

$$\hat{d} = \arg\max_{d_i} P(d_i \mid q, d, o), \tag{4}$$

where q is the question, $o = \{o_1, o_2, ..., o_n\}$ are option contents, and $d = \{d_1, d_2, ..., d_n\}$ are the options labels (e.g A/B/C/D). In our experiments, we refer to this method as `Baseline`.

In recent work (Zheng et al., 2024a),it was proposed to mitigate option selection bias by averaging the results over options permutation. The idea is to use the set of all cyclic permutations $\mathcal{I} = \{(i, i+1, ..., n, 1, ..., i-1)\}_{i=1}^{n}$ to calculate the debiased probability:

$$\tilde{P}(d_i \mid q, d, o) = \frac{1}{|\mathcal{I}|} \sum_{I \in \mathcal{I}} \log P(\pi_I(d_i) \mid q, d, \pi_I(o)) \tag{5}$$

Since computing probabilities for all permutations for each question is expensive, authors propose to estimate the prior distribution for option IDs on test set which is 5% of all samples, and use it to

debias new samples. In our experiments we refer to this method as `PriDe`. The test set is the same as we use for the best heads selection.

5.3 EXPERIMENTAL SETUP

Our main experiments were carried out according to the following pipeline: first, we took a frozen pre-tranied Transformer LLM (its weights were not modified in any of the experiments); then, we passed through it questions from the validation subset and for each head of the model and each question obtained best in terms of *QK-score* answer. After that, we chose a single head on which the highest accuracy was achieved (if several heads appeared to have equal accuracy scores we chose one from the lower level of the model; although, in our experiments this happened extremely rare). Then we obtained the predictions via the baseline method and via *QK-score* on the chosen head. Finally, we perform random shuffle of options in all questions and repeat the abovementioned procedure: it is done to correctly compute the Permutation Accuracy metric. Note that it may be two different heads that achieve best *QK-scores* on validation set before and after option permutation.

We report two quality metrics on the test subset: accuracy of predicted answers (from the first run) and *Permutation Accuracy* (PA) metric. The latter was introduced in Gupta et al. (2024) and is, in a sense, accuracy stable for choice permutation. PA metric is computed as the percentage of questions for which model choose correct choices before and after the random permutation of options. $\text{PA} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}_i \mathrm{I}_i^p$, where N is the dataset size, I_i is the indicator value equals to 1 iff model's answer on question i is correct, while I_i^p equals to 1 iff model gives correct answer on question i after its options (their texts not letters) were permuted. Answer options "`E. None of the above.`" and "`F. I don't know.`" are special and therefore are exempt from shuffling.

The prompt templates we use in our experiments are provided in Appendix A.3. In few-shot regimes before asking the question we provide a model with demonstrations in the same format except that the true answers (single capital letter for the correct option) are given after each example separated by single whitespace. The examples are separated from each other and from the actual question by single line breaks. The demonstrations are the same for every question in the given dataset. The set of examples for $(k+1)$-shot prompts contains the set of examples for k-shot prompts and one new example. The demonstrations were chosen from the first fifteen entries of the validation set, and their choice was mostly arbitrary, but we tried to filter out questions that we considered suboptimal from the perspective of an English-speaking human expert.

5.4 RESULTS

Figure 3 demonstrates the results of our method for LLaMA2-7B model. We observe an impressive improvement by 7-16% on all the datasets in zero-shot regime. Although *QK-scores* is not completely robust to option permutations, it is more stable than the baseline: the relative performance drop by PA metric is less than the baseline on all the datasets. In the few-shot regime, our approach is on par or outperforms other methods, with the most visible improvement on Halu Dialogue dataset by 5-9% depending on the number of shots.

PriDe results are added on Figure 3 for the comparison. PriDe in the most cases performs better than the baseline, but sometimes fails in zero-shot regime. Our analysis reveals that this method is not robust for additional uncertain options "E" and "F". We additionally provide experiments without such options in Appendix B, where PriDe performs better in few-shot regimes, but still loses in 0-shot setup. But overall, in all cases and for any options set, QK score outperforms PriDe.

We also applied our method to larger models of LLaMA family: LLaMA2 (-13B, -70B) and LLaMA3 (-8B, -70B) as well as to their chat/instruction-tuned versions. Table 1 presents the results of our method for large models in zero-shot regime; full version including few-shot regimes is provided in Appendix G. Overall, the results are in line with those obtained for LLaMA2. For all the smaller models of 8B and 13B size in zero-shot settings, our approach outperforms the baseline on all the datasets, both on accuracy and permutation accuracy, with the improvement up to huge 27%, achieved on HellaSwag dataset with LLaMA3-8B model. With larger models, MMLU is the most difficult benchmark for our method, likely because questions from it are aimed at measuring general knowledge, while our method by design focuses more on the semantic relations between the question and the possible answers.

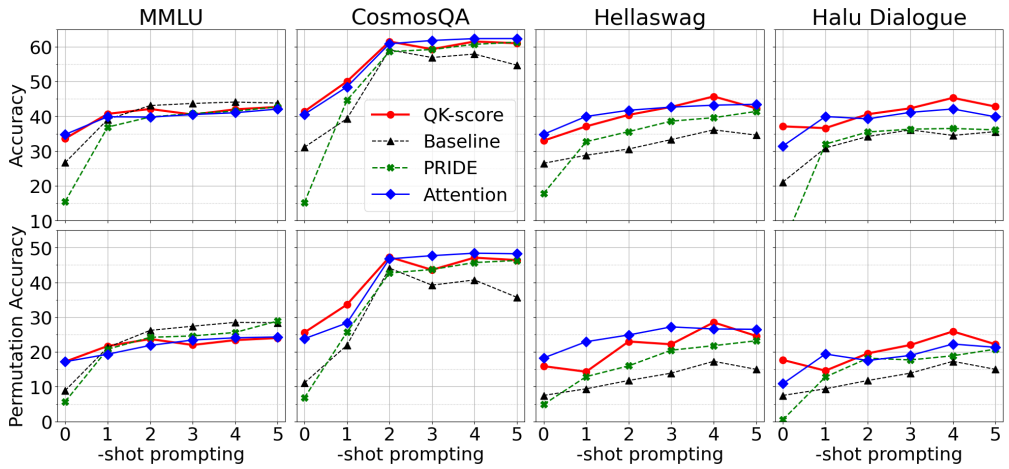


Figure 3: Comparison of different methods for LLaMA2-7B (base) on various Q&A datasets. Reported metrics are Accuracy (Acc) and Permutation Accuracy (PA).

Method		LLaMA... -30B	-65B	2-13B	2-70B	3-8B	3-70B	LLaMA... (chat, instruct) 2-13B	2-70B	3-8B	3-70B
MMLU											
Baseline	Acc	**50.4**	**48.3**	34.6	**59.7**	60.3	**75.3**	47.4	57.7	60.5	**78.2**
	PA	**37.9**	**35.7**	22.4	**48.5**	50.4	**68.8**	34.6	45.9	47.7	**70.1**
QK-score	Acc	45.2	46.2	**42.2**	56.7	**61.0**	74.5	**49.7**	**58.9**	**63.0**	77.9
	PA	30.7	32.1	**25.9**	39.2	**51.5**	66.0	**38.3**	**47.1**	**49.3**	67.9
Cosmos QA											
Baseline	Acc	59.9	**65.7**	29.6	65.5	54.9	82.0	48.1	68.5	85.4	91.6
	PA	**47.5**	**53.1**	19.4	**56.3**	39.3	75.7	36.8	58.3	71.0	82.5
QK-score	Acc	**60.1**	63.5	**58.2**	**69.5**	**70.6**	**87.6**	**67.7**	**84.8**	**88.6**	**94.1**
	PA	44.4	50.8	**44.3**	56.2	**60.9**	**81.7**	**51.6**	**75.9**	**75.1**	**88.1**
Hellaswag QA											
Baseline	Acc	35.2	33.4	36.8	71.6	33.5	**82.5**	41.6	61.4	67.4	**86.8**
	PA	16.5	13.7	17.1	62.9	15.8	**76.1**	25.8	49.0	27.8	71.2
QK-score	Acc	**43.9**	**53.8**	**52.9**	**74.9**	**60.9**	82.1	**50.8**	**73.0**	**72.5**	86.3
	PA	**21.5**	**35.0**	**38.8**	**63.3**	**50.8**	75.2	**37.3**	**64.9**	**36.3**	**72.8**
Halu Dialogue											
Baseline	Acc	36.3	**46.7**	41.0	39.4	46.6	44.3	49.4	39.4	62.1	68.8
	PA	21.1	**29.8**	22.2	25.4	29.1	33.5	32.6	26.6	42.6	63.8
QK-score	Acc	**44.8**	42.4	**47.2**	**58.4**	**52.3**	**67.8**	**56.2**	**58.1**	**64.7**	**76.7**
	PA	**27.6**	22.5	**30.2**	**42.6**	**36.7**	**57.9**	**42.5**	**42.8**	**46.6**	**65.6**

Table 1: Comparison of different base models in zero-shot setup on various Q&A datasets. Reported metrics are Accuracy (Acc) and Permutation Accuracy (PA). Best results are highlighted in **bold**.

Regarding performance of models on synthetic dataset SSD, in Figure 5b we can see that in baseline zero-shot setting LLaMA2-7B struggles to point onto the correct option, meanwhile, our method allows to extract the needed information from the model and gain much better quality. The figure shows accuracy for *QK-score* from five best heads (we denote them by their *(Layer, Head)* indices). Three of these heads can also be seen in the Table 10; the other two ((8, 8) and (12, 15)) are unique for this particular dataset.

6 Analysis

Choosing option-representative tokens. To compare our scores, we need to select option-representative tokens $\{t_i\}$, where the semantic information about each option semantics is con-

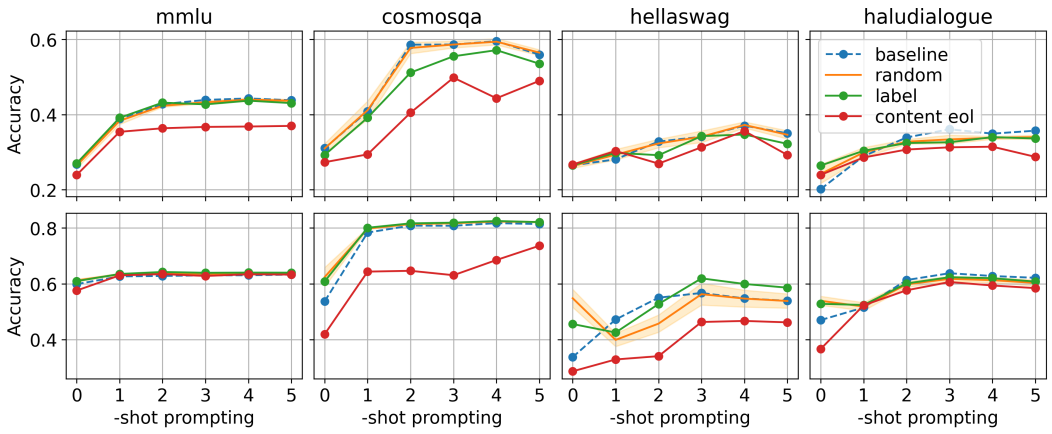


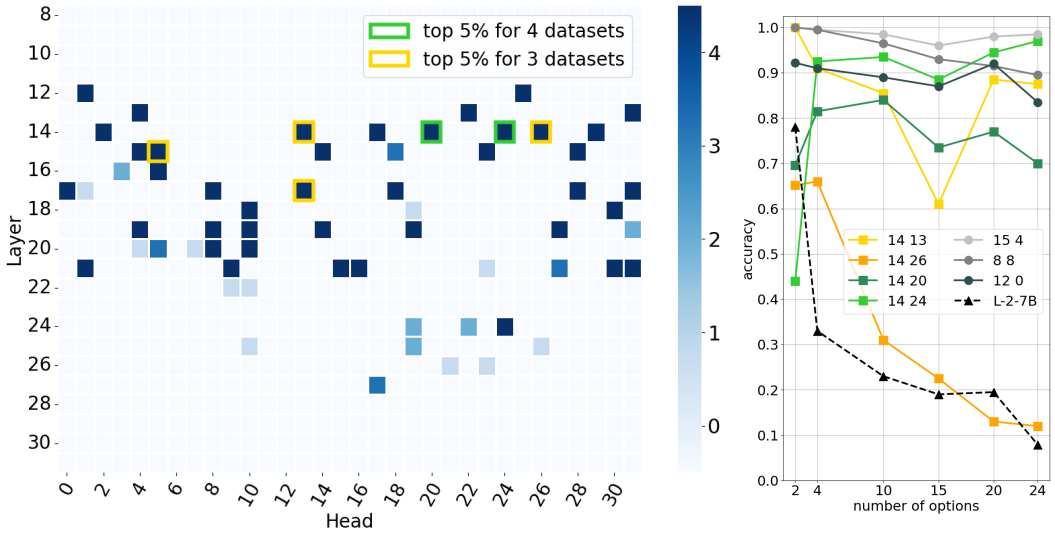
Figure 4: Zero-ablation of heads for LLaMA2-7B (upper) and LLaMA3-8B (lower)

centrated. Due to the causal nature of the attention in LLMs, the logical choice is the last token after the content of the option, which is the end-of-line token. We use it in most of our experiments, although there are other tokens worth analysing: label itself, period after label and period after option content (see Fig. 2a). We also experimented with the mean aggregated score through all tokens in the content of the option, but it gave poor results. The detailed analysis of such variations for attention scores is presented in Fig. 2b. We observe that the period after content and the end-of-line tokens are the most representative of our scores. There is an interesting finding concerning label token: despite it being almost useless in 0-shot setup as was shown in (Lieberum et al., 2023) also, we can see the better performance for 5-shot setup, in different heads. We hypothesise that there exist several types of "select-and-copy" heads, which influence the logits differently.

Select-and-copy heads ablation. To investigate *select-and-copy* heads and their relation to the performance of the model, we use zero-ablation of heads (Olsson et al., 2022) to analyze the causal relationship between *select-and-copy* heads and model output. In this method, we replace the output of a selected set of heads with a zero vector, effectively removing their contribution to the residual stream. We experiment with a set of the 10 best heads based on the *Attention-score* with EOL and label option-representative tokens and report the results in Fig. 4. Additionally, we perform random ablations by selecting 10 random heads and aggregating the results from 5 runs. To ensure a fair comparison, we select random heads only from the middle layers (12 to 20), where the top heads are also located. We observe a significant drop in accuracy, sometimes below random performance, for the most of dataset, when ablating heads by EOL token, indicating a causal relationship with the model's output. Some additional experiments where logit lens (Nostalgebraist, 2020) is utilized to further evaluate the layer-wise dynamics of the answer and *select-and-copy* heads in the model can be found in Appendix F.

Best heads. As choosing the best head on validation set requires a sufficient amount of training data, we would like to determine whether universal heads exist, that perform on par with the calibrated for particular topic heads. Moreover, finding such heads would help mitigate the effects of a poorly chosen validation set, when discrepancies exist between the questions in the validation and test sets.

To illustrate how best-performing heads change in different setups, we select the best heads on the mixes across datasets and across shots, and select the 5% ot the best heads for each mix based on mean accuracy of each head. See Appendix D for more details. The result is shown on Figure 5a. This heatmap highlights the most stable heads, which appears among the best in several mixed tasks: when "shots" are mixed (framed cells), or when datasets are mixed (coloured cells). Most notably, the majority of robust heads in this sense lay within 12 and 21 layers. The most universal heads w.r.t. to the dataset change are (14,24) and (14,20). They appeared in the top 5% pairs in mixed-shot setup for all four datasets. They also demonstrate high performance on the synthetic data when the number of options is increased up to 24, as shown on Figure 5b, while the performance of the baseline method drops below random. These results provide an additional evidence that the selected heads indeed able to perform the option selection task based on



(a) (b)

Figure 5: (a) Heatmap for (layer, head) indices for the best performing heads in LLaMA2-7B. The top 5% heads were selected for each N-shot setup, with all 4 datasets combined. The intensity of color indicates the maximal N where this pair appears. The framed cells indicate best-performing pairs that are uniform for 3 or 4 datasets. The first 8 layers are omitted because no interesting heads are found there. (b) Synthetic Dataset QK-score accuracy for various numbers of options (number of options is plotted on x axis, varies from 0 to 24) in zero-shot for LLaMA2-7B. Different colors of the lines correspond to different heads. "Square" markers correspond to the heads, performing well across real datasets (they are "framed" on Figure 5a), and "round" markers correspond to the heads that work well on the synthetic dataset specifically. The "triangle"-marked dotted line reflects the baseline model's performance.

option content. For more detailed analysis for 0-shot performance see analysis using percentiles in Appendix E.

Attention patterns analysis. Figure 6 reflects the typical attention pattern together with QK scores for our most stable head (14, 24); attention patterns of the other best heads across our tasks - (14, 20), (14, 26), and (14, 13) (right top corner of the Figure 7) are reflected in Appendix, Figure 8. We can see that the attention weights are concentrated on option-representative tokens, namely $\backslash n$ symbols after options, with the highest weight on the correct option, and exactly that is expected from *select-and-copy* heads. Interestingly, that QK score gives the clearer picture.

Finding best heads without validation labels. Based on this observation, we can propose an algorithm to find such stable heads without a labeled validation set. Namely, such heads should have heavy attention weights on option-representative tokens and high variability in the options they attend to. Thus, we can score each head using a product of two values: 1) sum of average attention weights to all $\backslash n$ symbols after options on this head; 2) a frequency of "choosing" any option aside of the most popular one (see formal definitions at Appendix I). If some head doesn't attend on the options, then the first value is close to zero, meanwhile when the head is "looking" on options, but "chooses" the same option most of the time, the second value is close to zero. Multiplying these two values yields low scores for heads that consistently "look" at the same option or ignore options entirely. Conversely, heads that "look" at diverse options receive high scores. Sorting all the heads of LLaMA2-7B model by this score, we see that the named four heads have very high score. Namely, they get into top-20 heads if scored across real datasets we use, and they get into top-10 heads if scored on our synthetic dataset, see Figure 34.

Selection bias. Following previous studies on selection bias Pezeshkpour & Hruschka (2024); Zheng et al. (2024a), we investigate our methods towards the tendency to choose specific option rather then choosing a correct answer. We observe that among best heads we also have uneven distribution in predictions, which are corrected as well when increasing the number of shots. How-

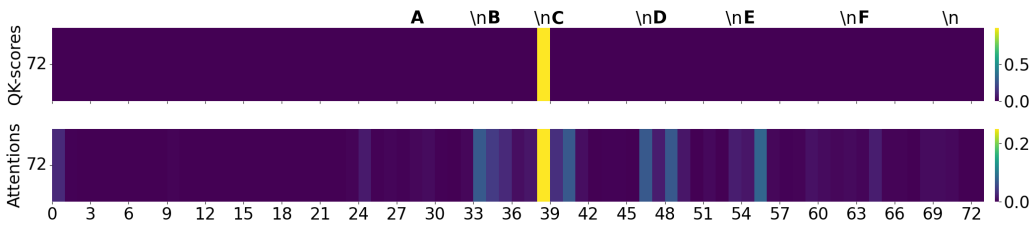


Figure 6: QK-scores after softmax (upper part of the diagram) and attentions (lower part of the diagram) for the last token on the 0-shot MMLU example on (14, 24) head. The task is "`Question: What singer appeared in the 1992 baseball film 'A League of Their Own'\nOptions:\nA. Brandy.\nB. Madonna.\nC. Garth Brooks.\nD. Whitney Houston.\nE. I don't know.\nF. None of the above.\nAnswer:`". Full version is on Figure 8.

ever, there is an interesting pattern that two best heads distributions are complementing each other, i.e $S_{QK}^{(14,20)}$ is biased to options "A" and "D" an $S_{QK}^{(14,24)}$ - to options "B" and "C". More detailed information can be found in Fig. 35 in Appendix J.

7 Conclusion

In this work, we introduced two novel scoring mechanisms: *QK-score* and *Attention-score*, derived from internal mechanism of LLM that can help to improve the performance on multiple-choice question answering tasks. Our experiments demonstrated significant improvements (up to 16%) across popular benchmarks, and even more striking results (up to 60%) on a synthetic dataset designed to test the model's understanding of task format.

We identified a subset of attention heads, which we termed *select-and-copy* heads that play a critical role in these performance gains. These heads are relatively stable across different datasets and exist universally across model scales, and we explored their causal effect on task performance. Our findings suggest that these specialized heads have the potential to deepen our understanding of LLMs' capabilities not only for MCQA but for other reasoning tasks as well.

This work opens up new avenues for further research into the internal dynamics of LLMs, including a deeper exploration of attention mechanisms and their role in complex task-solving that requires selection and copying information from the text.

8 Limitations

Our method cannot be applied to models without an access to attention matrices. Also, our method is not applicable on scarce-resource tasks, even though one can utilize the heads we marked as robust enough. Besides, MCQA task itself was criticized for oversimplification (Balepur et al., 2024).

References

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13787–13805, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.744. URL https://aclanthology.org/2024.acl-long.744.

Anthropic. Introducing the next generation of claude, 2024. URL https://www.anthropic.com/news/claude-3-family.

Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10308–10330, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.555.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

Sky CH-Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. Do androids know they're only dreaming of electric sheep? In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 4401–4420, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.260.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2126–2136, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1198. URL https://aclanthology.org/P18-1198.

Abhimanyu Dubey et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 134–139, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-2524. URL https://aclanthology.org/W16-2524.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. Changing answer order can decrease mmlu accuracy, 2024.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Tigr1kMDZy.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*. OpenReview.net, 2021. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2021.html#HendrycksBBZMSS21.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://www.aclweb.org/anthology/D19-1243.

Jae-young Jo and Sung-Hyon Myaeng. Roles and utilization of attention heads in transformer-based neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3404–3417, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.311. URL https://aclanthology.org/2020.acl-main.311.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6449–6464, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.397. URL https://aclanthology.org/2023.emnlp-main.397.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=aLLuYpn83y.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. *arXiv preprint arXiv:2307.09458*, 2023.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 4791–4797, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fpoAYV6Wsk.

Nostalgebraist. Interpreting gpt: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

OpenAI. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022.

Madhura Pande, Aakriti Budhraja, Preksha Nema, Pratyush Kumar, and Mitesh M. Khapra. The heads hypothesis: A unifying statistical approach towards understanding multi-headed attention in bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13613–13621, May 2021. doi: 10.1609/aaai.v35i15.17605. URL https://ojs.aaai.org/index.php/AAAI/article/view/17605.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=yKbprarjc5B.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics ACL 2024*, pp. 7407–7416, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-acl.441.

Fanghua Ye, Yang MingMing, Jianhui Pang, Longyue Wang, Derek F Wong, Yilmaz Emine, Shuming Shi, and Zhaopeng Tu. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*, 2024.

Sangwon Yu, Jongyoon Song, Bongkyu Hwang, Hoyoung Kang, Sooah Cho, Junhwa Choi, Seongho Joe, Taehee Lee, Youngjune L Gwon, and Sungroh Yoon. Correcting negative bias in large language models through negative attention score alignment. *arXiv preprint arXiv:2408.00137*, 2024.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=shr9PXz7T0.

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models: A survey. *arXiv preprint arXiv:2409.03752*, 2024b.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023. URL https://arxiv.org/abs/2310.01405.

A Datasets

A.1 Datasets Details

Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021) contains 4-way questions on the variety of topics related to STEM, the humanities, the social sciences, and other fields of knowledge. We sample 10,000 instances from the test set to utilize them in our experiments.

CosmosQA[1] (Huang et al., 2019) together with question and answer options additionally contains text paragraph that is supposed to be used by a model to give the final answer. The purpose is to evaluate reading comprehension and commonsense reasoning capabilities of the model. Similar to MMLU, we sampled 10,000 instances from the test set.

HellaSwag (Zellers et al., 2019) evaluates the commonsense reasoning capabilities of the model through selecting the best sentence completion for a given sentence prompt, given a short text as a context. We, once again, extracted 10,000 entities from this dataset.

Halu Dialogue is a "dialogue" part of HaluEval (Li et al., 2023a) dataset with about 10,000 examples. Here a model is asked to choose an appropriate continuation of a dialogue from four possible options.

A.2 Examples of questions from datasets

[1] https://wilburone.github.io/cosmos/

```
Question: Where is the Louvre museum?
Options:
    A. Paris.
    B. Lyon.
    C. Geneva.
    D. Vichy.
    E. I don't know.
    F. None of the above.
```

Listing 1: MMLU example

```
Context: My house is constantly getting messy and I ca n't keep up . I am
     starting at a new school with no one I know and it is 4 times bigger
     than UAF . I am now going to have to balance school , homework ,
    kids , bill paying , appointment making and cleaning when I can
    barely keep up without the school and homework ( keep in mind this is
     a full time GRADUATE program at a fairly prestigious school ) . We
    are in financial crisis .
Question: What is causing the narrator 's recent stress ?
Options:
    A. They are moving to a new house .
    B. I would have tried to guess their password and alternatively gone
    to a coffee shop for wifi.
    C. They are moving to a new university .
    D. They are moving to a new house for the kids .
    E. I don't know.
    F. None of the above.
```

Listing 2: CosmosQA example

```
Context: A young boy is wearing a bandana and mowing a large yard. he
Question: Which of the following is the best ending to the given context?
Options:
    A. is unrelieved by the weeds and is barely smiling.
    B. walks away from the camera as he pushes the mower.
    C. moves and walks the mower but gets stuck because he is engaged in
    a game of ping pong with another boy.
    D. seems to be doing a whole lot of things and talks to the camera
    from behind a white fence.
    E. I don't know.
    F. None of the above.
```

Listing 3: HellaSwag example

```
Context: [Human]: I like Pulp Fiction. What do you think about it? [
    Assistant]: I love it. It was written by  Roger Avary [Human]: I
    heard he also wrote The Rules of Attraction. Do you know who is in
    that movie?
Question: Which of the following responses is the most suitable one for
    the given dialogue?
Options:
    A. Swoosie Kurtz is in it.
    B.  Fred Savage is in it.
    C. Yes, it is a drama and crime fiction as well. Do you like crime
    fiction stories too?.
    D. No, it was not made into a film. However, it was adapted into a
    popular Broadway musical.
    E. I don't know.
    F. None of the above.
```

Listing 4: Halu Dialogue example

```
Question: Which of the following options corresponds to " optimal "?
Options:
    A. ion.
    B. optimal.
    C. coins.
    D. jackie.
    E. I don't know.
    F. None of the above.
```

Listing 5: Simple Synthetic Dataset example

A.3 Prompt Templates and Examples

Variable parts are highlighted in **bold**; whitespace placing is marked by underscores; position of line-breaks is explicitly shown by symbols '\n' (note that the last line always ends without whitespace or line break). In our datasets we ensured that each question ends with question mark, and each choice ends with point (single whitespace before it does not affect the logic of tokenization by LLaMA tokenizer).

```
Question:␣{Text of the question}?\n
Options:\n
A.␣{Text of the option A}␣.\n
B.␣{Text of the option B}␣.\n
C.␣{Text of the option C}␣.\n
D.␣{Text of the option D}␣.\n
E.␣I␣don't␣know␣.\n
F.␣None␣of␣the␣above␣.\n
Answer:
```

Listing 6: MMLU prompt template

```
Context:␣{The context of the question/situation or the dialog history}\n
Question:␣{Text of the question}?\n
Options:\n
A.␣{Text of the option A}␣.\n
B.␣{Text of the option B}␣.\n
C.␣{Text of the option C}␣.\n
D.␣{Text of the option D}␣.\n
E.␣I␣don't␣know␣.\n
F.␣None␣of␣the␣above␣.\n
Answer:
```

Listing 7: CosmosQA/HellaSwag/Halu Dialogue prompt template

Following are an example of 1-shot prompts from MMLU. 2-3-4-5-shot prompts were built in the same way and prompts for dataset with context are built the same way, except each question is preceded by its context. Note that in demonstrations we add a single whitespace between "`Answer:`" and the correct choice letter; for example, "`Answer:  A`", but *NEVER* "`Answer:A`". This is done because sequences like "`:  A`" and "`:A`" are differently split into tokens by LLaMA tokenizer, and the former produces the same tokens corresponding to letter "`A`" as in the choice option line, while later yields a different version of "`A`". From LLaMA's point of view, these two versions of letters are separate entities and are NOT interchangeable. Removing those symbols of whitespace in many cases leads to noticeable drop in performance.

```
Question:␣A␣medication␣prescribed␣by␣a␣psychiatrist␣for␣major␣depressive␣
    disorder␣would␣most␣likely␣influence␣the␣balance␣of␣which␣of␣the␣
    following␣neurotransmitters?\n
Options:\n
A.␣serotonin␣.\n
B.␣dopamine␣.\n
C.␣acetylcholine␣.\n
```

```
D.␣thorazine␣.\n
E.␣I␣don't␣know␣.\n
F.␣None␣of␣the␣above␣.\n
Answer:␣A\n
Question:␣
    Meat should be kept frozen at what temperature in degrees Fahrenheit?
    \n
Options:\n
A.␣0 degrees or below␣.\n
B.␣between 10 and 20 degrees␣.\n
C.␣between 20 and 30 degrees␣.\n
D.␣0 degrees or below␣.\n
E.␣I␣don't␣know␣.\n
F.␣None␣of␣the␣above␣.\n
Answer:
```

Listing 8: An example of 1-shot prompt for a question from MMLU dataset

B SOME MORE INTUITION ON OPTIONS 'E' AND 'F'

As we mentioned in the main text, inclusion of fictional, though always incorrect, choices "`E. None of the above`" and "`F. I don't know`" in every question was aimed at creating the "uncertainty sinks". However, they are also beneficial for the analysis of attention head roles, but that is somewhat beyond the scope of this article. Here we would like to provide some intuition to it.

We performed experiments on a modified version of our datasets, where questions include only 4 "meaningful" choices, i.e. options 'A'-'D' only. Scatterplots on Figure 9 show the correlation between accuracy of heads using QK-scores on options without 'E'-'F' (by y-axis) and their accuracy on questions with all 6 options (by x-axis). Here, only validation subsets were used. We present plots for few of the possible setups, but other follow similar pattern. From these charts we can see that if a head reaches good accuracy answering 4-choice questions, it usually will reach nearly the same accuracy on questions with 6 choices and vice versa, see points around the diagonal $y = x$ in the upper-right quadrant.

We can also observe another major trend: horizontal stripe near y-level 0.25. It can be explained in the following manner: in the data used, ground-truth answers are perfectly balanced – that is, for every choice 'A'-'D' 25% of the questions have it as the correct answer. And if a head reaches 4-choice accuracy of $\approx 25\%$, it falls into one of the three categories:

1. This head chooses only one option in all questions. Usually it is the last one of the list.
2. This head "guesses" answers, choosing options nearly randomly and "independent" from their meanings.
3. This head "understands" questions, but is genuinely bad at answering them.

Addition of choices 'E' and 'F' drops the performance of the first type heads down to nearly 0%, second type – to around 16.7%; QK-scoring accuracy of the third type heads, however, usually remains the same.

Thus, we can conclude that choices 'E' and 'F' cause little effect on performance of good heads, but, at the same time, their inclusion creates separation between heads that are bad at Multiple Choice Question Answering and heads which do not have MCQA in their functionality at all (they may perform other roles for LM).

C NUMERICAL RESULTS FOR COMPARISON OF QK-SCORE WITH OTHER METHODS

Table 2 provides numerical results for our main experiments with QK-scores from heads of LLaMA2-7B model that are presented on Figure 3 in the main text.

Method		...-shot prompting 0	1	2	3	4	5
MMLU							
Baseline	Acc	26.7	39.1	**43.1**	**43.7**	**44.1**	**43.8**
	PA	8.9	21.3	**26.2**	**27.4**	**28.5**	**28.4**
PRIDE	Acc	15.5	36.9	39.8	40.8	41.5	42.7
	PA	5.7	20.8	24.2	24.6	25.6	28.9
Attention score	Acc	**34.8**	39.9	39.8	40.5	41.0	42.1
	PA	**17.2**	19.4	21.9	23.4	24.1	24.3
QK-score	Acc	33.6	**40.7**	42.1	40.5	42.0	42.7
	PA	**17.2**	**21.7**	23.7	22.0	23.4	24.0
Cosmos QA							
Baseline	Acc	31.1	39.3	59.1	56.9	57.9	54.7
	PA	11.1	21.9	44.1	39.2	40.7	35.7
PRIDE	Acc	15.2	44.6	58.6	59.2	60.7	61.3
	PA	6.8	25.7	42.7	43.7	45.7	46.3
Attention score	Acc	40.6	48.5	60.9	**61.8**	**62.3**	**62.3**
	PA	23.8	28.3	46.8	**47.7**	**48.4**	**48.2**
QK-score	Acc	**41.4**	**50.0**	**61.5**	59.3	61.5	61.0
	PA	**25.6**	**33.6**	**47.3**	43.6	47.1	46.4
Hellaswag QA							
Baseline	Acc	26.5	28.8	30.6	33.3	36.1	34.6
	PA	7.5	9.4	11.8	13.9	17.3	15.0
PRIDE	Acc	17.8	32.7	35.6	38.6	39.6	41.4
	PA	4.9	12.9	16.0	20.5	21.8	23.2
Attention score	Acc	**34.8**	**40.0**	**41.7**	42.6	43.2	**43.5**
	PA	**18.3**	**22.9**	**24.9**	**27.2**	26.6	**26.5**
QK-score	Acc	33.0	37.1	40.4	**42.7**	**45.7**	42.3
	PA	15.9	14.3	23.0	22.2	**28.5**	24.6
Halu Dialogue							
Baseline	Acc	21.1	30.9	34.2	36.1	34.5	35.6
	PA	5.4	10.2	14.3	18.9	16.8	20.7
PRIDE	Acc	3.0	32.0	35.5	36.3	36.5	36.1
	PA	0.5	12.8	18.2	17.7	18.9	20.8
Attention score	Acc	31.4	**39.9**	39.3	41.1	42.1	39.9
	PA	10.9	**19.4**	17.5	19.0	22.3	21.3
QK-score	Acc	**37.1**	36.6	**40.6**	**42.3**	**45.3**	**42.8**
	PA	**17.7**	14.6	**19.6**	**22.0**	**25.9**	**22.2**

Table 2: Comparison of different methods for LLaMA2-7B (base) on various Q&A datasets. Reported metrics are Accuracy (Acc) and PErmutation Accuracy (PA). Best results are highlighted in **bold**.

D BEST HEADS

Setup	Best (Layer, Head)
0-shot	**(14, 24)**
1-shot	(15, 5), **(15, 23)**, (14, 20)
2-shot	**(14, 24)**, (15, 5), (15, 4) (18, 10), **(15, 23)**, (16, 17)
3-shot	**(14, 24)**, (15, 5), (15, 4) (18, 10), **(15, 23)**, (14, 26), (17, 18)
4-shot	**(14, 24)**, (15, 5), (14, 4) (15, 4), (18, 10), **(15, 23)** (14, 20), (14, 26), (17, 18), (16, 17)

(a) Mixed top heads based on datasets

Dataset	Best (Layer, Head)
MMLU	**(14, 24)**, (15, 4), (17, 0), **(14, 20)**, (20, 10), (18, 30)
HaluDialogue	(14, 29), **(14, 24)**, (14, 26)
HellaSwag	(15, 5), (15, 4), (18, 10), **(14, 20)**, (14, 13), (13, 22)
CosmosQA	**(14, 24)**, (15, 5), (15, 4), (18, 10), (17, 0), (15, 23), **(14, 20)**, (14, 26), (14, 13), (18, 30)

(b) Mixed top heads based on shots.

Figure 10: Top 1% heads based on accuracy for different ways of mixing

```
mmlu_top_heads = {
    0: [(14, 24), ...],
    1: [(14, 20), ...], ... # for all 5 shots  top 1% heads for MMLU
}

hellaswag_top_heads = {
    0: [(15, 10), ...],
    1: [(14, 20), ...], ... # for all 5 shots top 1% heads for HellaSwag
}

top_heads = [[] for i in range(5)]  # 0 shot to 4 shot

for index in range(5):
    top_heads[index] = mmlu_top_heads[index] + hellaswag_top_heads[index]

best_heads_across_shots = set(top_heads_for_each_shot[0])

for index in range(1, 5):
    top_heads_for_each_shot &= set(top_heads_for_each_shot[index])
```

Listing 9: Example of calculation of best heads for all 5 shots for two datasets

E STABILITY OF BEST HEADS

Then, we utilized the minimum of accuracy percentiles as a way to determine stable heads that can be seen on Figure 11a. Again, the heads from 14th layer are showing the highest accuracy on almost all percentiles. We also listed the top 1% pairs for all setups based on accuracy in Table 10. There is a noticeable overlap between heads for various setups, and, once again, all of them are middle layers of the model.

If we compare the performance of the "stable" heads with results obtained with preceding calibration on Figure 7, (14, 24) and (14, 20) are frequently chosen from validation set, but even when they do not, their performance is comparable to their validation-chosen counterparts, except for HaluDialogue. Besides, we tested the heads (14, 24), (14, 20), (14, 26), and (14, 13) for a stability against increasing the amount of options in SSD dataset (see Figure 5b) and against changing the symbols that denote an options, following Alzahrani et al. (2024) (see Appendix H). We also added other heads that are performing well on SSD dataset to these plots for comparison.

F LOGIT LENS EXPERIMENT

We follow Halawi et al. (2024) and track the accuracy in the intermediate layers using logit lens (Nostalgebraist, 2020). Denoting $\boldsymbol{h}^{(l)} \in \mathbb{R}^d$ as a hidden state corresponding to last token in

layer l, we extract intermediate probabilities for options d_i using:

$$P_l(d_i \mid q, d, o) = \text{Logits}^{(l)}_{t_i}, \qquad \text{Logits}^{(l)} = \text{Softmax}(\boldsymbol{W}_U \cdot \text{LayerNorm}(\boldsymbol{h}^{(l)})) \tag{6}$$

Fig. 12 demonstrates the results for LLaMA2-7B base model, which shows some interesting patterns. In most cases, we see the improvements after the 12 layer for all setups excluding 0-shot. As we compare it with the maximal accuracy over *Attention-score* for two different types of option-representative tokens, we see the similar trend. However, the peak accuracy is seen in the middle layers, after which it degrades. Interestingly, that the logit lens performance demonstrate a sudden performance drop around 20. This indicates that some alternative "thoughts" about the answer emerges at this point, being overlapped further by the correct answer.

G Comprehensive results for experiments on larger models

Here we provide complete results of our experiments with QK-scores on four main datasets (MMLU, CosmosQA, HellaSwag and Halu Dialogue) for larger models. As before, reported metrics are Accuracy and Permutation Accuracy.

- Figure 13 contains results for LLaMA2-13B, and Figure 19 for its chat-tuned version
- Figure 14 contains results for LLaMA2-70B, and Figure 20 for its chat-tuned version
- Figure 15 contains results for LLaMA3-8B, and Figure 21 for its instruct-tuned version
- Figure 16 contains results for LLaMA3-70B, and Figure 22 for its instruct-tuned version
- Figure 17 contains results for LLaMA-30B
- Figure 18 contains results for LLaMA-65B.

Note that in our experiments the accuracy scores for these baseline models are somewhat lower than ones you can see in the original technical reports (Touvron et al., 2023; Dubey et al., 2024). The main reason for this is that we added additional "E" and "F" options that were not used in those reports; some differences in prompts and particular examples for few-shot learning also could play a role. Also note that in many experiments we focus on zero-shot scenario without chain-of-thoughts prompting, that has received less attention in the original technical reports.

H Behaviour of the best heads under the change of options symbols and options amount

Aside from the standard version of Simple Synthetic Dataset (with four essential options and two additional options "E" and "F"), described in Section 5.1, we considered alternative versions of SSD containing various numbers of possible options. For example, the variation of the dataset that corresponds to the number "10" on the x-axis of the Figure 5b contains ten essential options - A, B, C, D, E, F, G, H, I, J, - and two special options - "K. I don't know" and "L. None of the above" (see the Example 10). Also note that in these experiments we used 200 examples from each version of the dataset to get the attention scores.

Figure 31 is an extended version of the Figure 5b, that includes more heads for LLaMA2-7B (left) and the similar experiment for several heads of LLaMA3-8B (right), four of which are taken from the upper right part of the Figure 11b as the most stable across real datasets.

```
Which of the following options corresponds to " mediterranean "?
Options:
    A: acceptance
    B: specialties
    C: charitable
    D: typically
    E: access
    F: jose
    G: findlaw
    H: colonial
```

Method	0-shot Orig.	A	B	C	D	5-shot Orig.	A	B	C	D
Baseline	26.6	73.4 (+46.8)	7.9 (-18.7)	0.6 (-26.0)	28.3 (+1.7)	43.9	41.7 (-2.2)	53.4 (+9.5)	38.2 (-5.7)	42.5 (-1.4)
$S_{QK}^{(14,20)}$	31.4	43.9 (+12.5)	9.0 (-22.4)	10.8 (-20.6)	60.2 (+28.8)	37.6	78.2 (+40.6)	40.5 (+2.9)	25.2 (-12.4)	12.4 (-25.2)
$S_{QK}^{(14,24)}$	33.6	30.7 (-2.9)	58.5 (+24.9)	33.2 (-0.4)	14.4 (-19.2)	43.0	14.3 (-28.7)	49.7 (+6.7)	53.0 (+10.0)	51.9 (+8.9)
$S_{QK}^{(15,23)}$	26.2	30.9 (+4.7)	2.1 (-24.1)	5.0 (-21.2)	63.6 (+37.4)	36.3	71.2 (+34.9)	36.3 (+0.0)	14.7 (-21.6)	27.0 (-9.3)

Table 3: Selection bias for different methods on MMLU 0-shot and 5-shot using LLaMA2-7B. The table compares original accuracy (for task to predict A/B/C/D/E/F) and recall only on subset with single ground truth option (i.e. only questions with answer A).

```
I: mediterranean
J: data
K: I don't know.
L: None of the above.
```

Listing 10: Modification of SSD with ten options - example

On Figure 32 we return to standard 4-options SSD dataset but change the symbols for option labels. We include (renamed) special options "E" and "F" for the upper plot and omit them for the lower plot for LLaMA2-7B model, and Figure 33 shows the same but for LLaMA3-8B model.

I Head scoring without validation set

Let $\hat{\mathcal{D}}$ be some unlabelled MCQA dataset. Then, for each head we may calculate a score

$$HeadScore = \left(\frac{1}{|\hat{\mathcal{D}}|} \sum_{\hat{\mathcal{D}}} \sum_{i=1}^{n} a_{Nt_i} \right) \left(\frac{1}{|\hat{\mathcal{D}}|} \mathbb{I}\{\arg\max_i (a_{Nt_i}) \neq \hat{i}\} \right),$$

where $\hat{i}$ denotes the most frequent option for the given head; head indices (l, h) are omitted. The left component here denotes the average amount of attention, concentrated on the option-representative tokens $t_i, i = 1, \ldots, n$. The right component reflects the frequency of the situation, when the largest attention among the options falls on the option other than $\hat{i}$, i.e. any not the most frequent option.

The results of ranking heads according to this scores are presented on Figure 34.

J Selection Bias

We investigate our methods towards the tendency to choose specific option rather then choosing a correct answer. Fig. 35 presents a selection bias for baseline and 3 heads for *QK-score* in 0-shot and 5-shot regimes.

Table 3 shows the selection bias in terms of recall. We can see that most methods (especially in 0-shot setup) concentrates on single or several options.

K Synthetic dataset in different languages

We regenerated our synthetic dataset using three languages in addition to English. Figure 36 shows that the general distribution of QK-scores across heads of LLAMA-2-7B model on these datasets remains largely unchanged; for instance, layers 8–15 still contain the most performant heads everywhere. However, differences in the performance of individual heads are also observed.

Below are the top-10 best-performing heads for each language, sorted by decreasing accuracy:

- EN: **(8, 8)**, **(15, 4)**, (12, 15), **(14, 24)**, (12, 10), (14, 13), **(14, 27)**, **(12, 25)**, **(12, 21)**, **(14, 20)**, accuracy decreasing from 0.995 to 0.815;
- IT: **(12, 21)**, **(15, 4)**, **(8, 8)**, **(14, 20)**, (12, 13), **(14, 24)**, **(14, 27)**, (12, 0), **(12, 25)**, (12, 10), accuracy decreasing from 1.0 to 0.9;
- FR: **(12, 21)**, (12, 13), **(8, 8)**, **(14, 20)**, **(15, 4)**, **(14, 27)**, **(14, 24)**, (8, 21), **(12, 25)**, (12, 0), accuracy decreasing from 1.0 to 0.905;
- RU: **(12, 25)**, **(14, 20)**, **(8, 8)**, (12, 13), (12, 15), **(12, 21)**, **(15, 4)**, **(14, 24)**, **(14, 27)**, (12, 6), accuracy decreasing from 1.0 to 0.91.

We colored green those heads that perform best across four real datasets (see Figure 5a). Additionally, we highlighted in bold those heads that are common across the top-10 in all four languages.

As shown, 7 out of top-10 best heads are shared across synthetic datasets on different languages (including two "green" heads that are also the best across our real datasets). It is significant overlap, which gives us hope for a substantial degree of universality among the identified heads. It is also interesting that the QK-scores for the best heads are somewhat lower for English compared to the other languages we analyzed. However, we cannot draw final conclusions from this observation without further investigation. A more thorough study into how exactly QK-scores and the best-performing heads vary with the dataset's language remains a topic for future research and is beyond the scope of this paper.

L RESULTS FOR OTHER MODEL FAMILIES

Here we provide results of our experiments with QK-scores on four main datasets (MMLU, CosmosQA, HellaSwag and Halu Dialogue) for the models from other families. As before, reported metrics are Accuracy and Permutation Accuracy.

- Figure 23 contains results for Dolly V2-3B
- Figure 24 contains results for Gemma-2B
- Figure 25 contains results for Phi-3.5-Instruct
- Figure 26 contains results for Qwen-2.5-1.5B-Instruct tuned
- Figure 27 contains results for Qwen-7B
- Figure 28 contains results for Qwen-7B-Instruct
- Figure 29 contains results for Qwen-14B
- Figure 30 contains results for Qwen-14B-Instruct

M BEST HEADS ON SYNTHETIC DATASET FOR QWEN 2.5-1.5B

In Figure 37a we provide an accuracy of the QK-score for each head of Qwen 2.5-1.5B-base on our Synthetic dataset. Interestingly, the layers with the best heads are closer to the final layer compared to those in the LLaMA-family models (7B and larger). Namely, the best heads are concentrated around layers 16-22, while the model itself has a total of 28 layers. Figure 37b shows similar tendency for Qwen-2.5-1.5B-Instruct: the best heads are concentrated around layers 13-22 in this case. However, overall, the situation resembles the corresponding heatmap for LLAMA-2-7B, shown in Figure 36a.

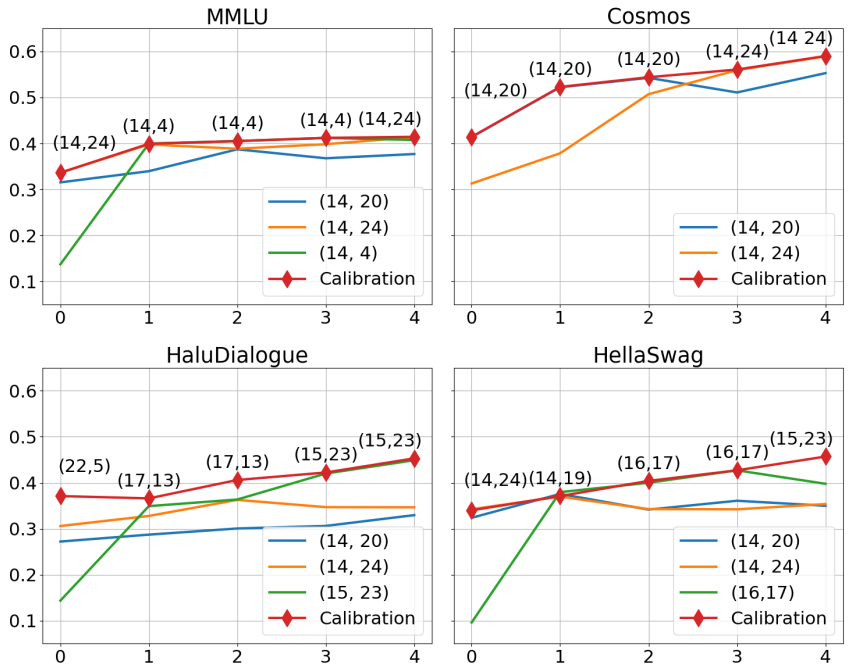


Figure 7: Accuracy of the best performing heads and of the most robust heads - (14, 24), (14, 20)

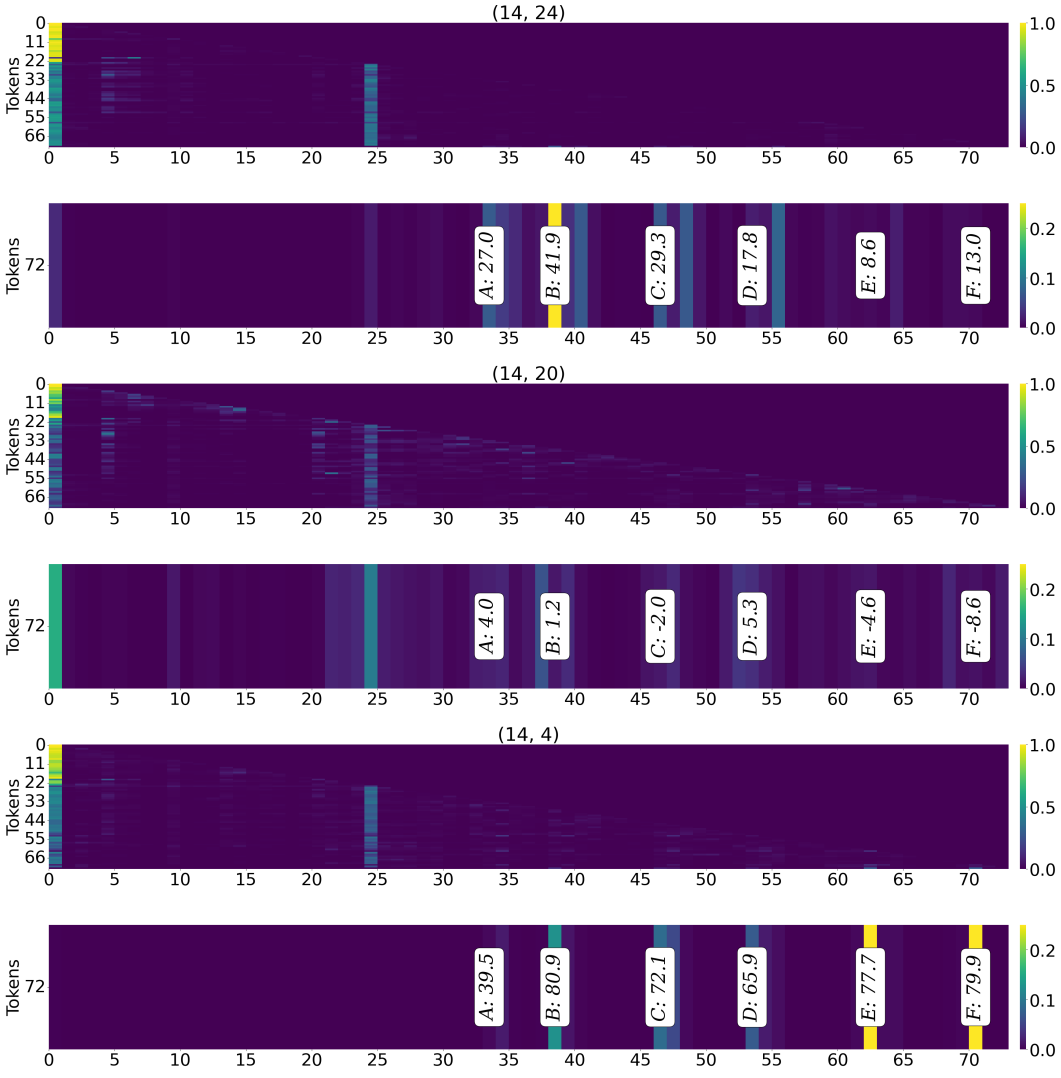


Figure 8: Attention maps of (14, 24), (14, 20) and (14,4) pairs (Head, Layer) for 0-shot setting for MMLU example: `Question: What singer appeared in the 1992 baseball film 'A League of Their Own'? \nOptions: \nA. Brandy.\nB. Madonna.\nC. Garth Brooks.\nD. Whitney Houston.\nE. I don't know.\nF. None of the above.\nAnswer:`. Second plot for each pair corresponds to the same, but scaled to the end-of-text-sequence attention map. Values in annotated cells are corresponding QK-score values. End of each option is denoted with \n symbols. 33th token is the end of A option, 38th token is the end of B option, 46th token - the end of C option, 53th token - the end of D option, 62th token - the end of E option, 70th token - the end of F option. The answer from QK-score of (14, 24) and (14, 4) is B, of (14, 20) is D. The correct answer for this example is B.

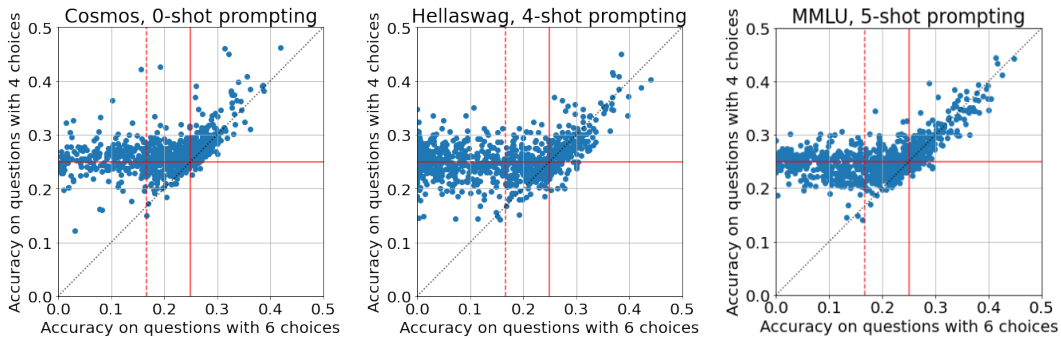
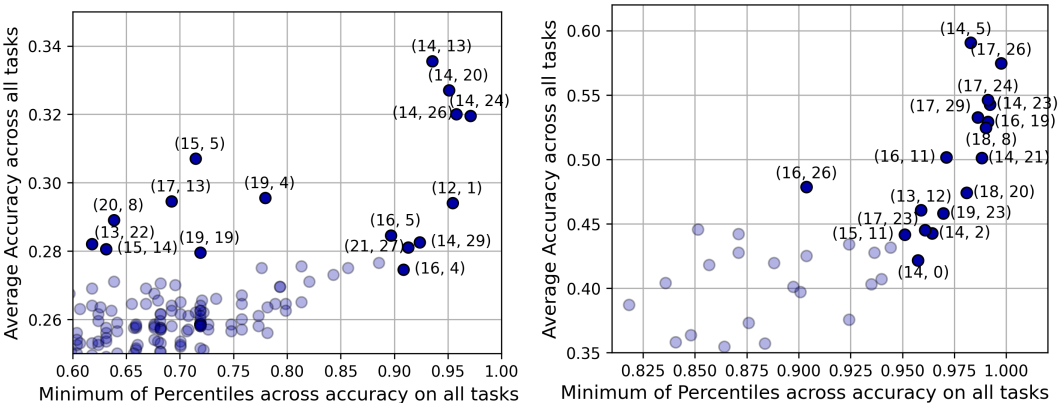


Figure 9: Correlation between heads QK-scoring accuracy on questions with 4 (‘A’-‘D’) and 6 (‘A’-‘F’) answer options. Solid red lines mark the accuracy level of 0.25, dashed red line – 0.167 (6 options random choice accuracy).



(a) (b)

Figure 11: Stable heads for *QK-score* in (a) LLaMA2-7B and (b) LLaMA3-8B for 0-shot setup across all tasks. “k-th Minimum of Percentiles” means that the head is better than k share of all heads for all tasks.

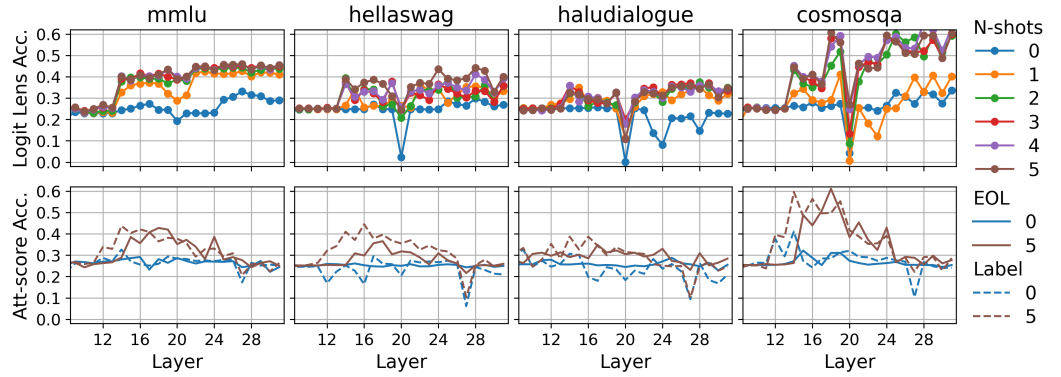


Figure 12: Logit Lens results on LLaMA2-7B base model for 0-shot and few-shot setups (upper) and a comparison to maximal accuracy per layer via *Attention-score* (lower)

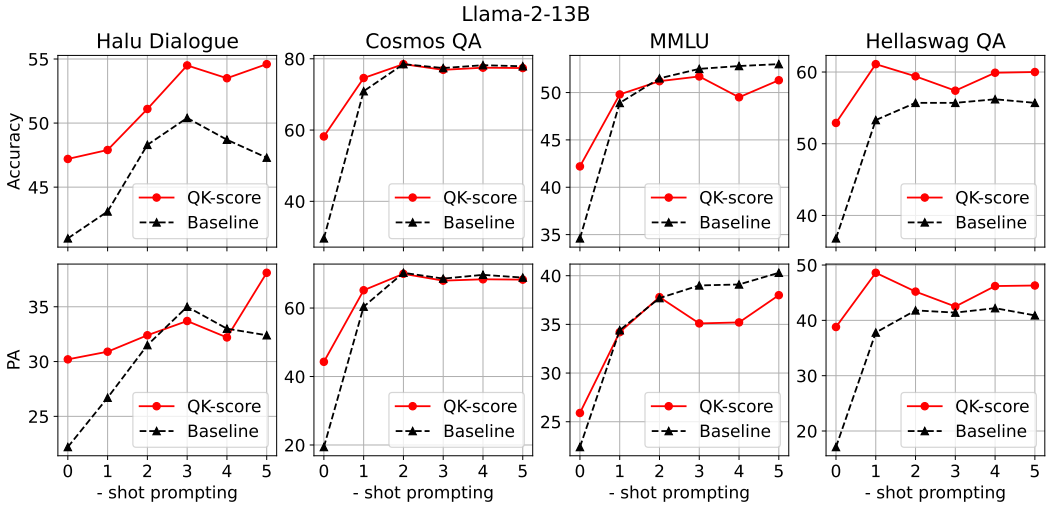


Figure 13: Comparison of different methods for LLaMA2-13B (base) on various Q&A datasets.

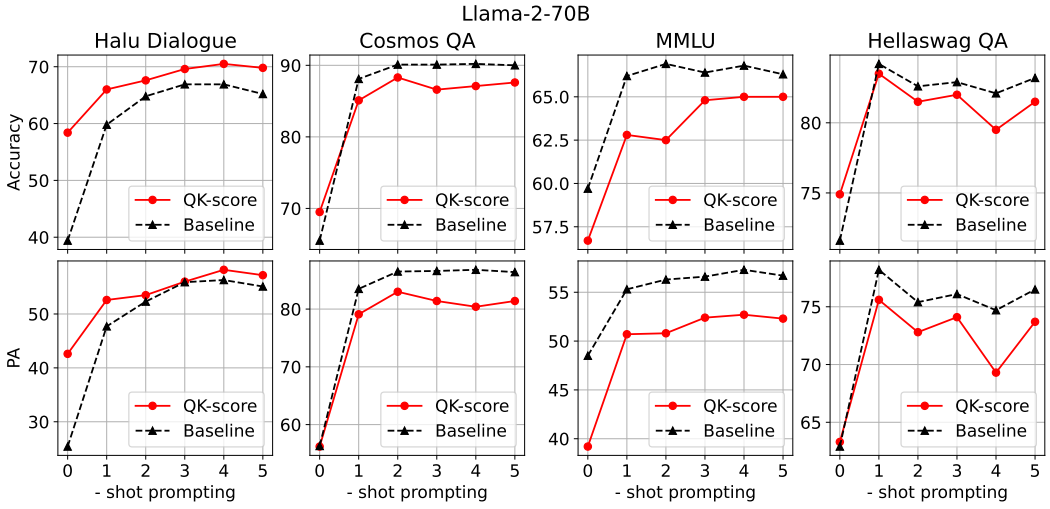


Figure 14: Comparison of different methods for LLaMA2-70B (base) on various Q&A datasets.

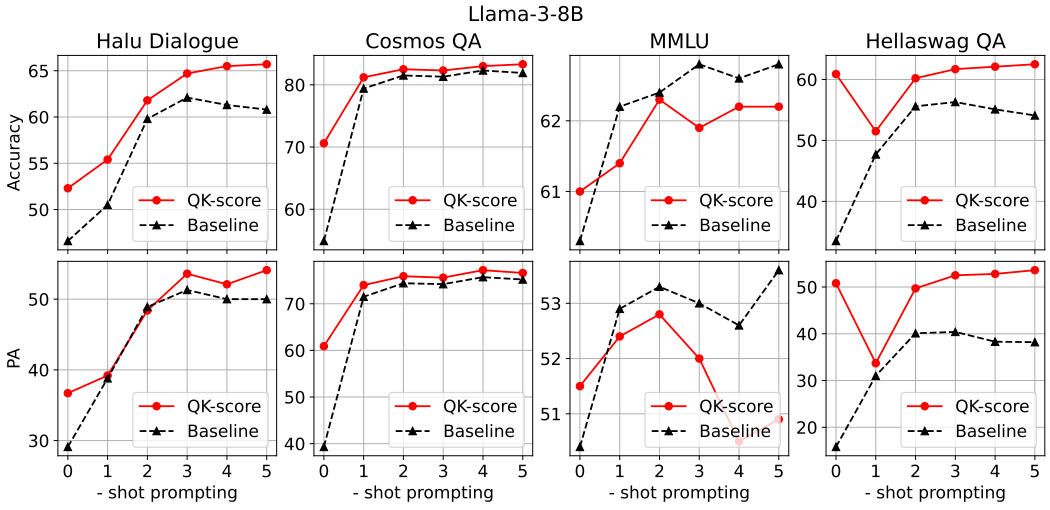


Figure 15: Comparison of different methods for LLaMA3-8B (base) on various Q&A datasets.

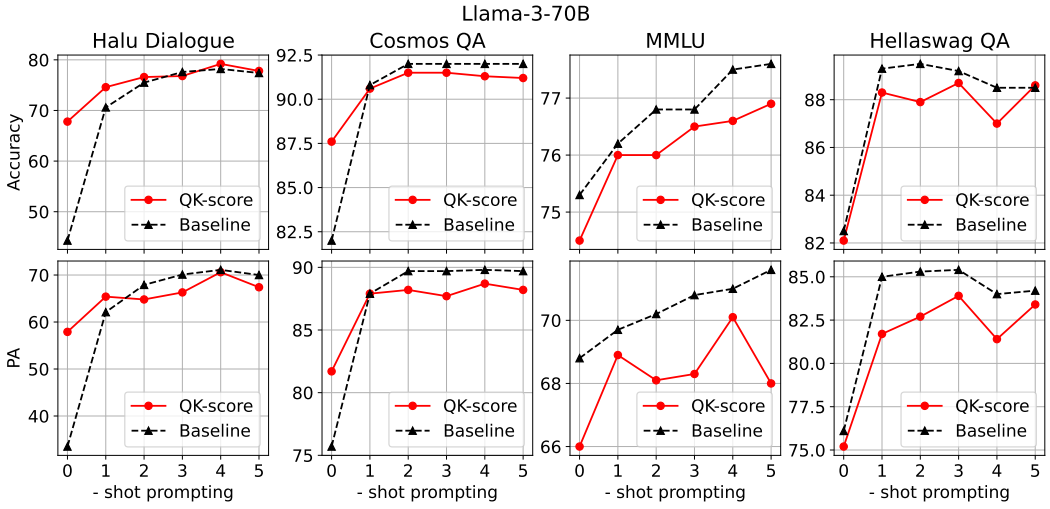


Figure 16: Comparison of different methods for LLaMA3-70B (base) on various Q&A datasets.

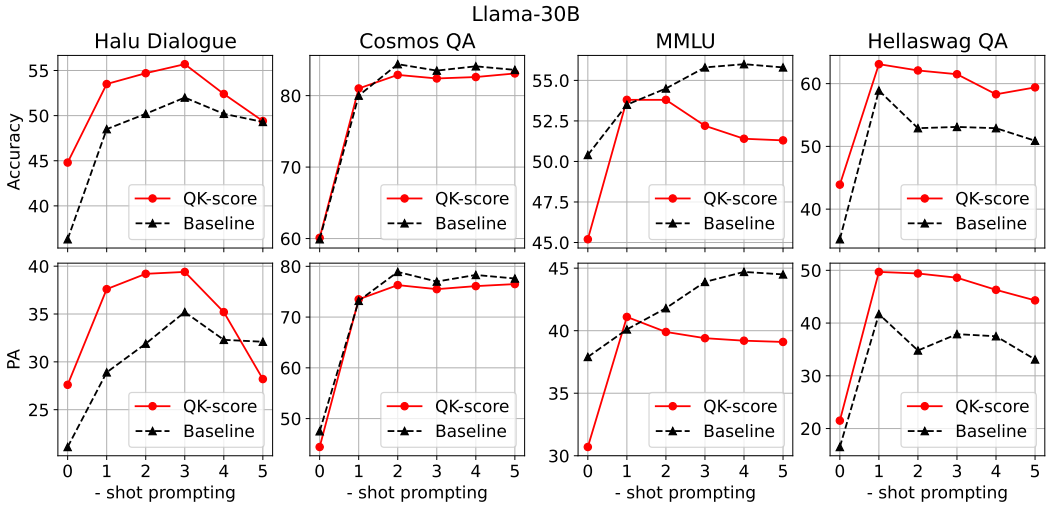


Figure 17: Comparison of different methods for LLaMA-30B (base) on various Q&A datasets.

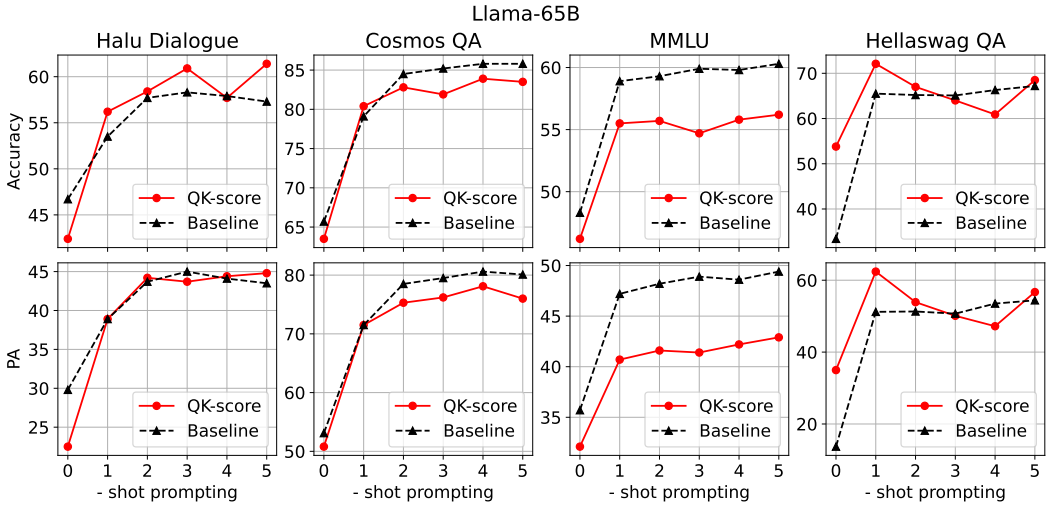


Figure 18: Comparison of different methods for LLaMA-65B (base) on various Q&A datasets.

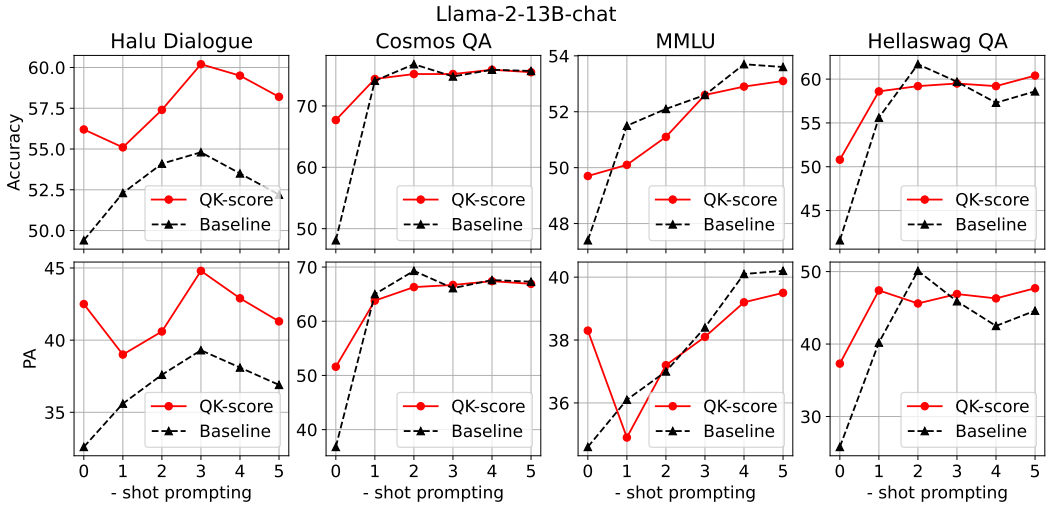


Figure 19: Comparison of different methods for LLaMA2-13B-chat on various Q&A datasets.

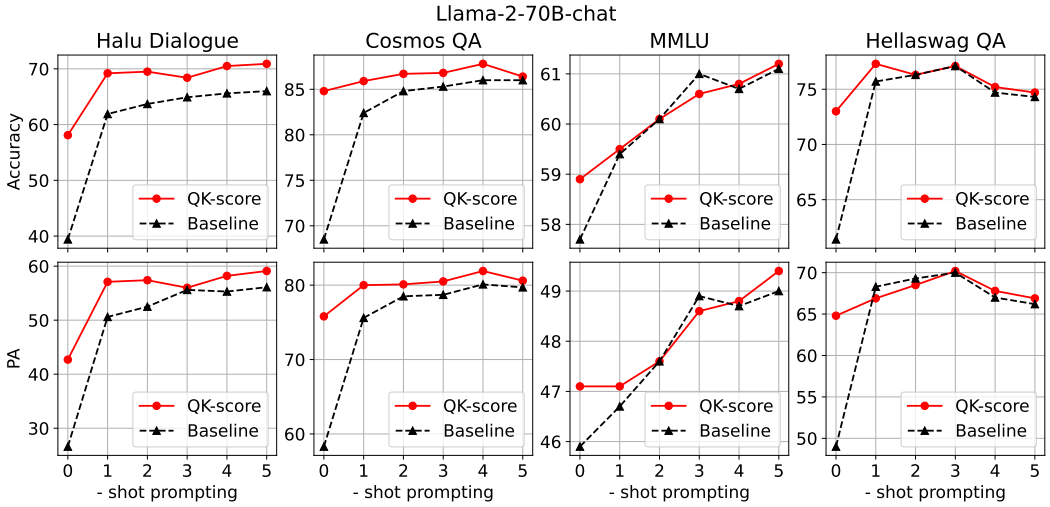


Figure 20: Comparison of different methods for LLaMA2-70B-chat on various Q&A datasets.

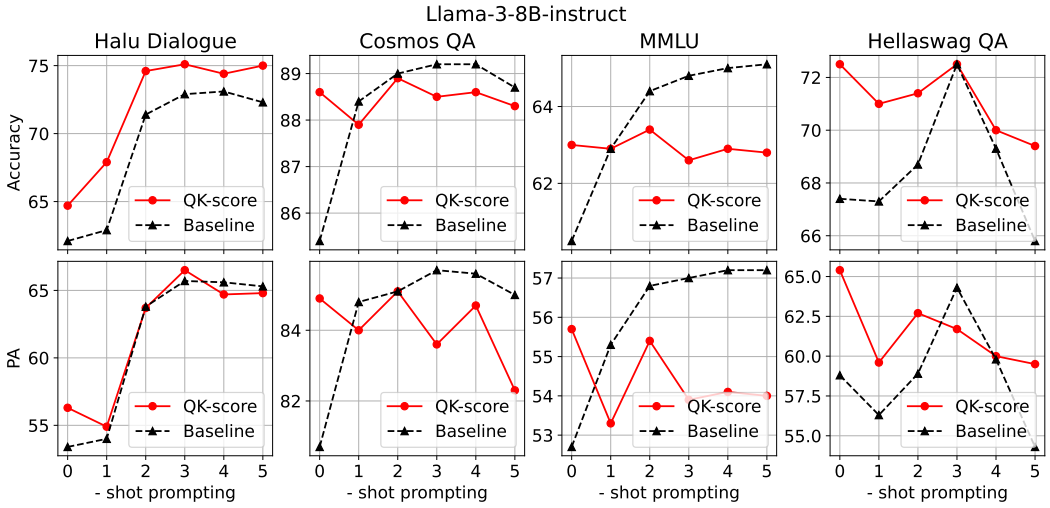


Figure 21: Comparison of different methods for LLaMA3-8B-instruct on various Q&A datasets.

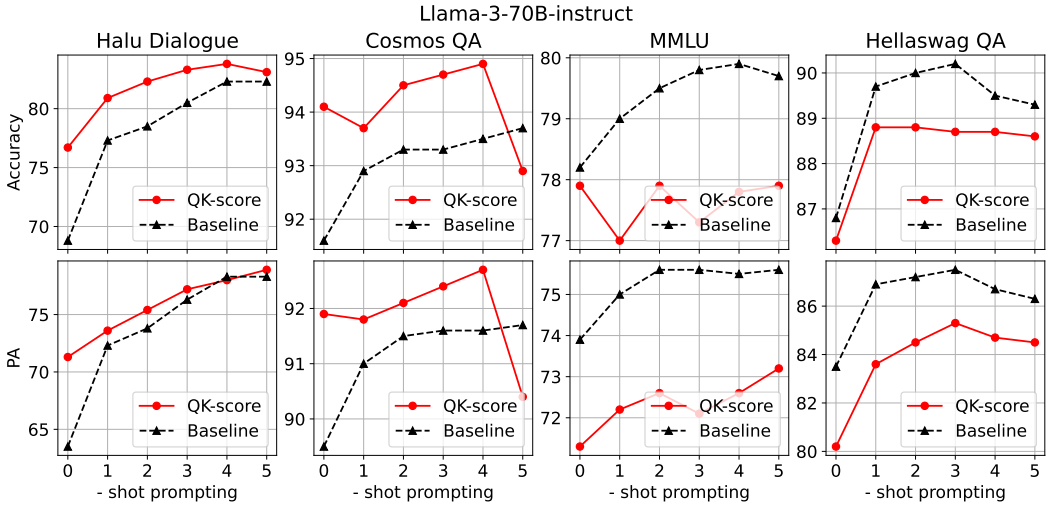


Figure 22: Comparison of different methods for LLaMA3-70B-instruct on various Q&A datasets.

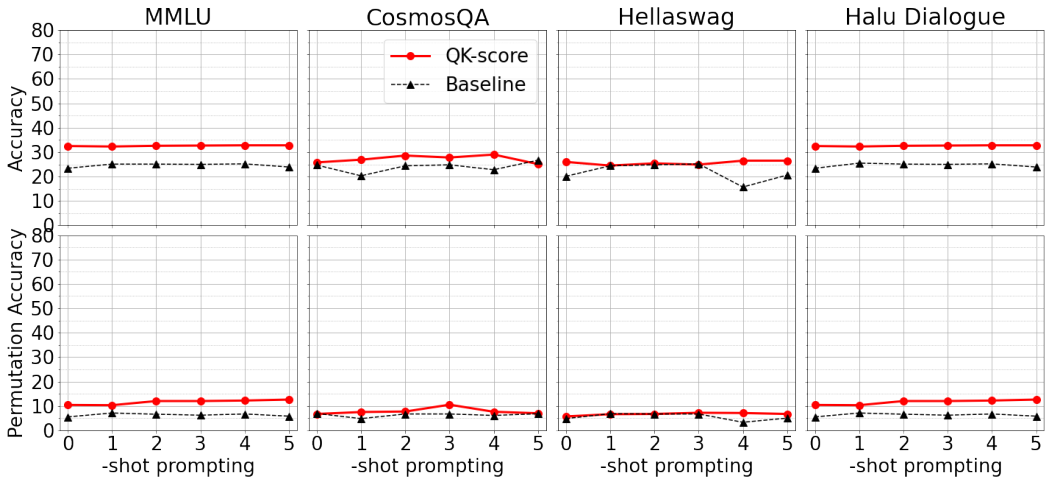


Figure 23: Comparison of different methods for Dolly V2-3B on various Q&A datasets.

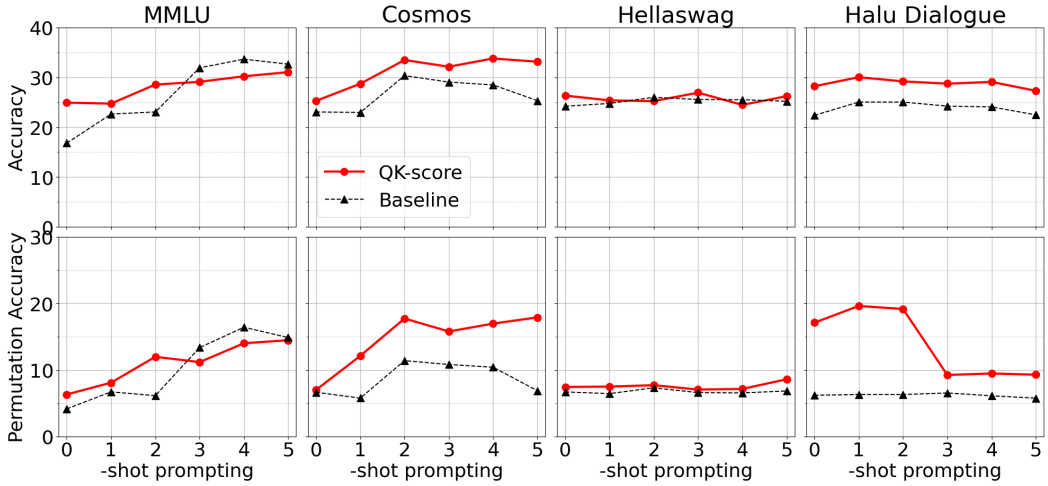


Figure 24: Comparison of different methods for Gemma-2B on various Q&A datasets.

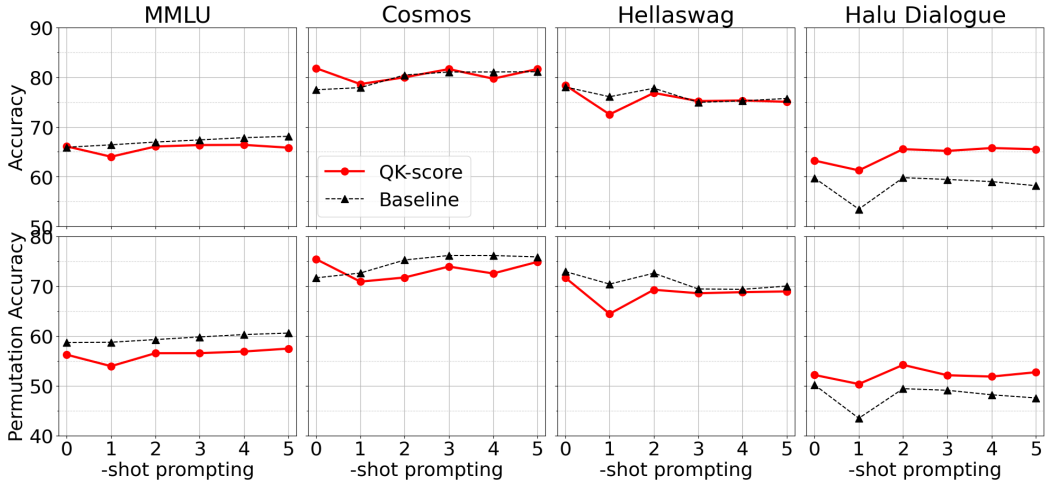


Figure 25: Comparison of different methods for Phi-3.5-mini (instruct tuned) on various Q&A datasets.

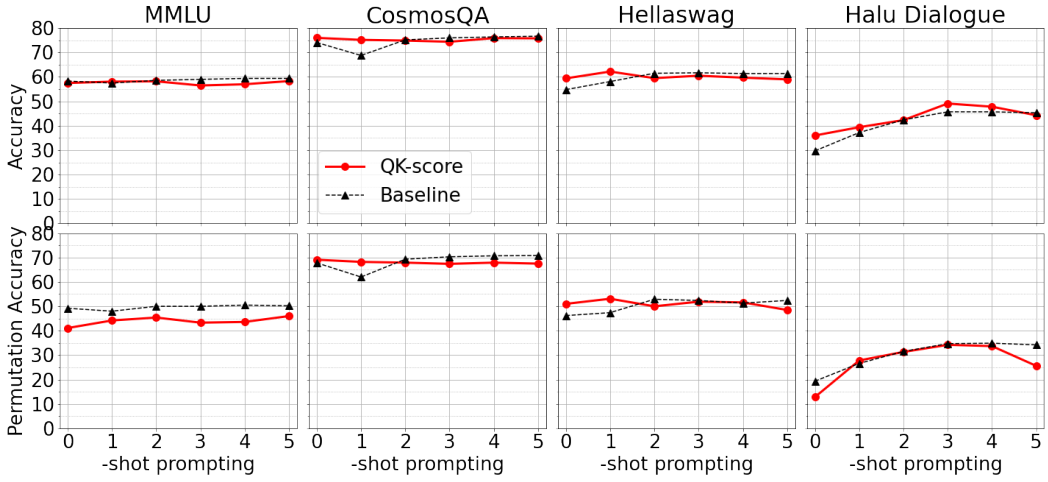


Figure 26: Comparison of different methods for Qwen-1.5B-Instruct on various Q&A datasets.

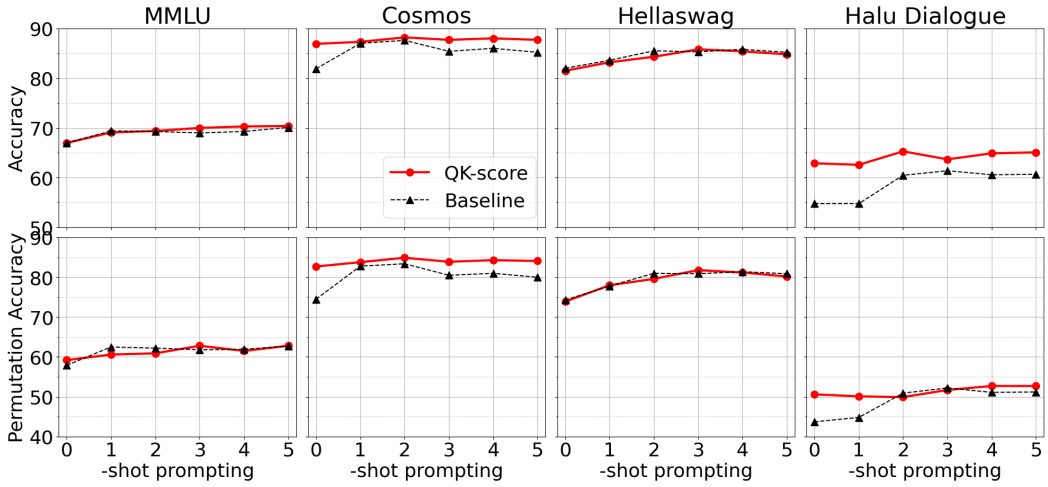


Figure 27: Comparison of different methods for Qwen-7B on various Q&A datasets.

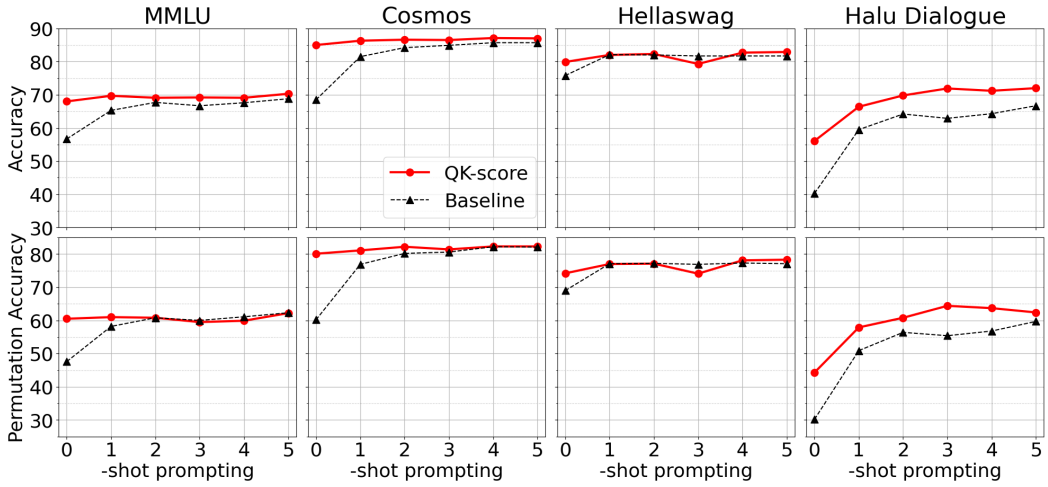


Figure 28: Comparison of different methods for Qwen-7B-Instruct on various Q&A datasets.

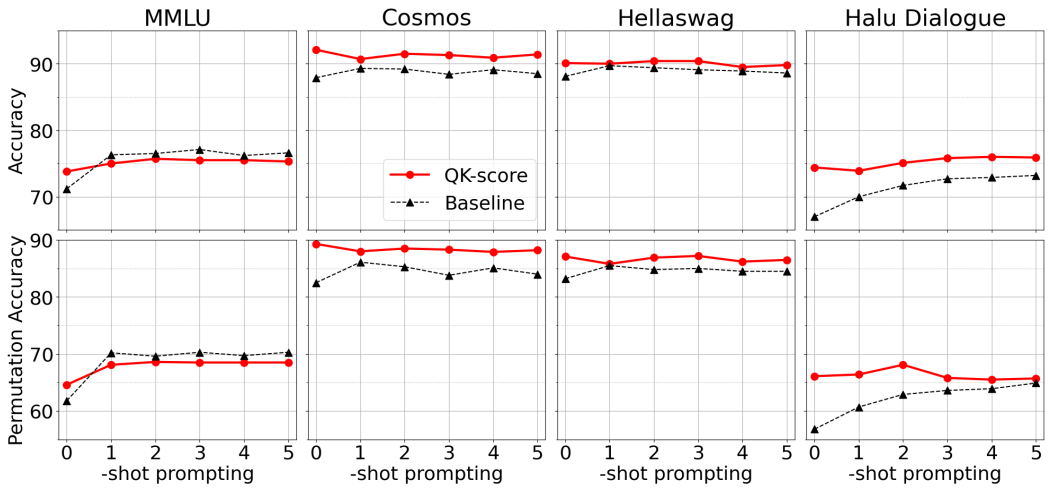


Figure 29: Comparison of different methods for Qwen-14B on various Q&A datasets.

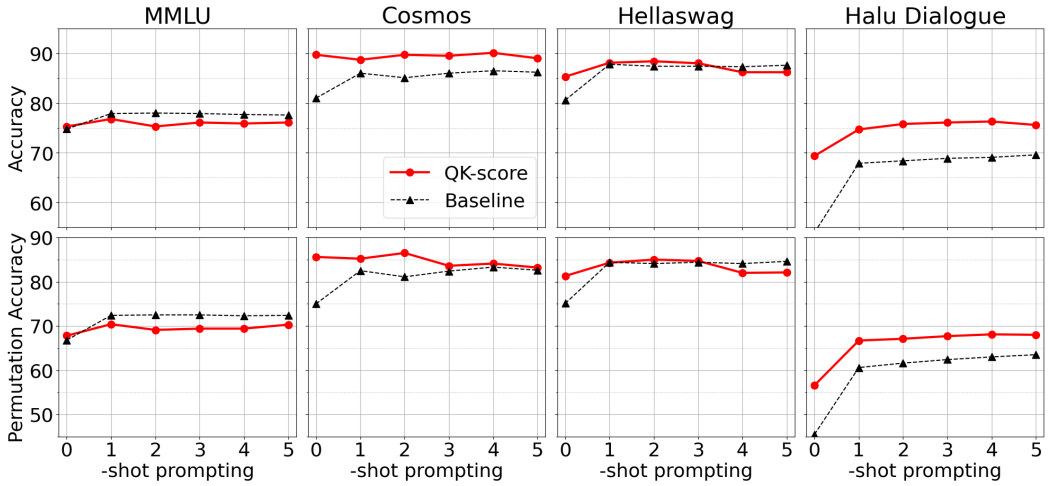


Figure 30: Comparison of different methods for Qwen-14B-Instruct on various Q&A datasets.

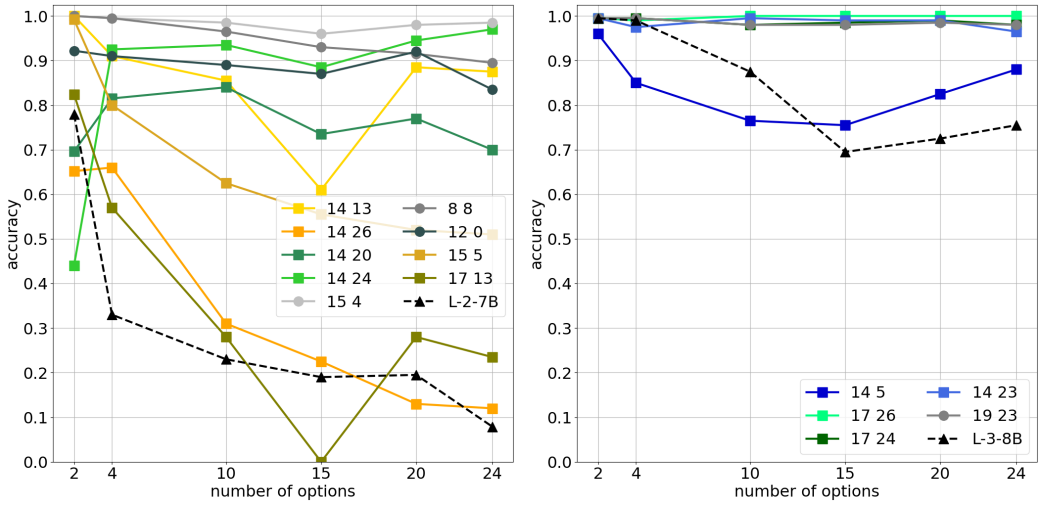


Figure 31: The results for a various numbers of options in the Simple Synthetic Dataset in zero-shot for LLaMA2-7B (left) and LLAMA3-8B (right). Different colors of the lines correspond to the results of QK dot products from different heads. "Square" markers correspond to the heads, working well across real datasets, and "round" markers correspond to the heads that work well on the synthetic dataset.

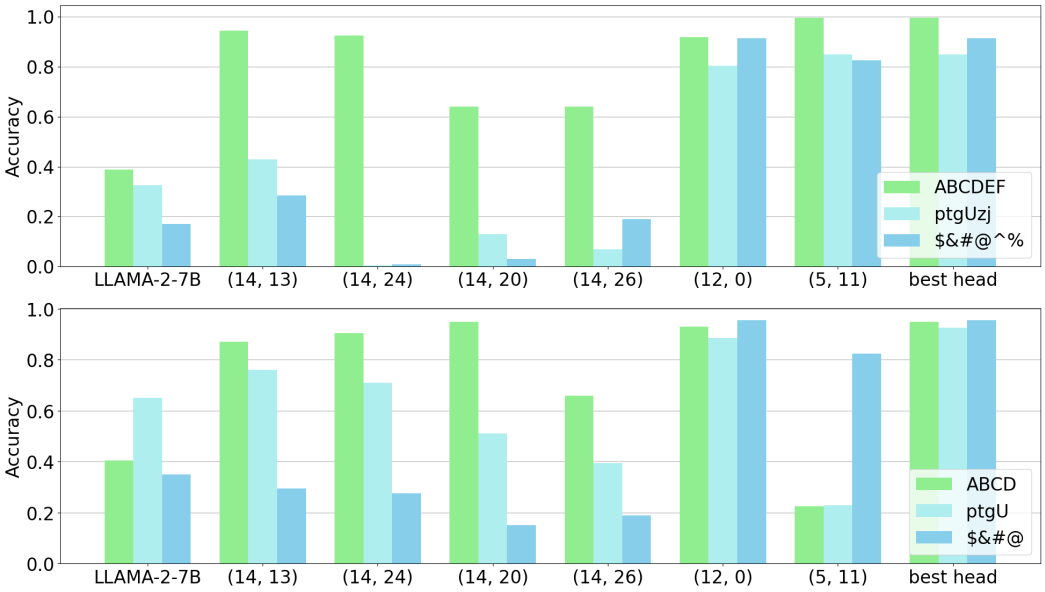


Figure 32: Performance of the QK-score from the best heads of LLaMA2-7B model for different options symbols when "uncertainty" options (i.e. "I don't know" and "None of the above") are presented (upper figure) and not presented (lower figure). The accuracy of the best four heads from the Figure 11a declines in these new setups, but the head (12, 0) keeps being stable across all of setups. Another interesting head is the head (5, 11): it's accuracy is high for all setups with "uncertainty" and for "$&# setup, but drops abruptly for "ABCD" and "ptgU". Studying such "anomalies" is a matter for future research.

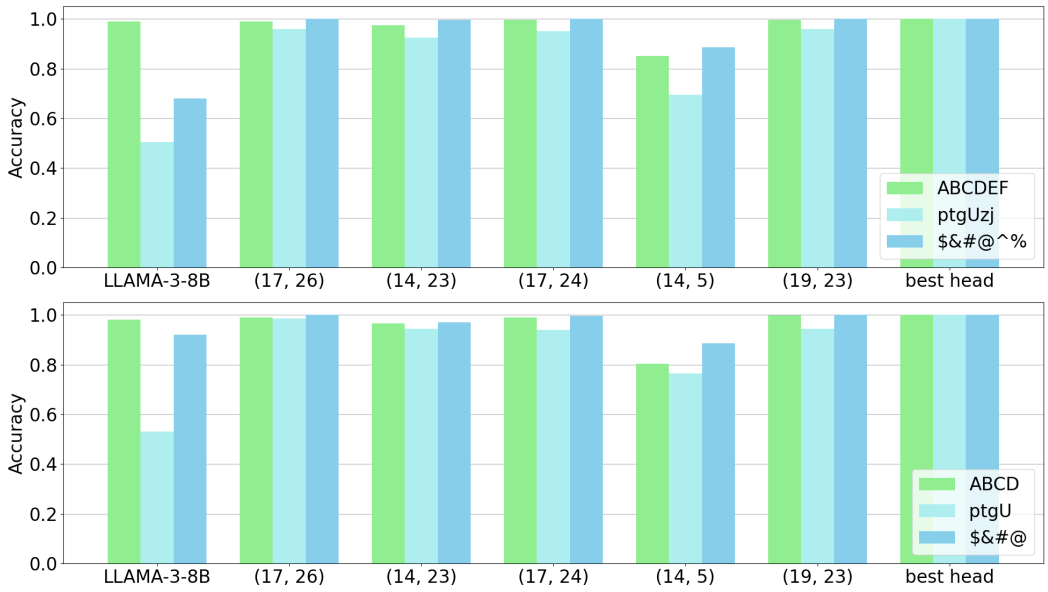


Figure 33: Performance of the QK-score from the best heads of LLaMA3-8B model for different options symbols with and without "uncertainty" options. Interestingly, the best heads of the LLAMA3-8B model (see Figure 11b) are significantly more stable across considered setup.

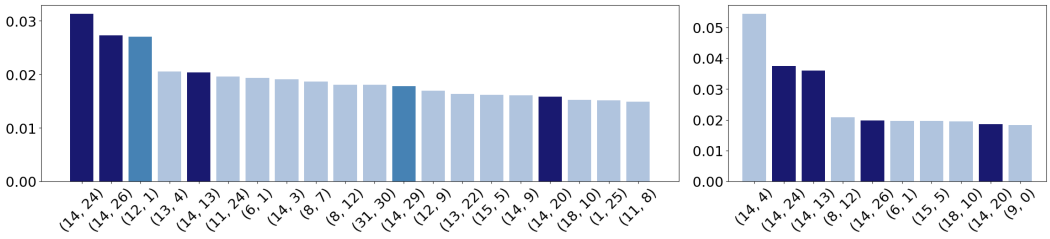


Figure 34: Left: average top heads scores across real datasets (first twenty). Dark blue marks four heads from the right top corner of Figure 11a. Medium blue marks other heads with minimum of accuracy percentiles on all tasks more then 0.9. As we can see, the first two heads that get the best scores across real datasets, belong to the group of the best heads from the right top corner of Figure 11a. Right: top heads scores on Simple Synthetic Dataset (first ten). Here, the top-scored head (14, 4) doesn't appear at the right top corner of Figure 11a, but it appears at the Figure 7 as one of the best heads for MMLU dataset. Note that for calculating this score we didn't use the dataset labels.

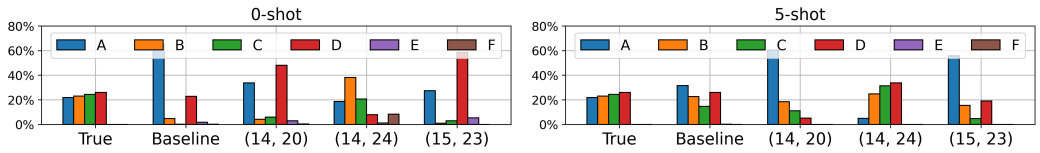
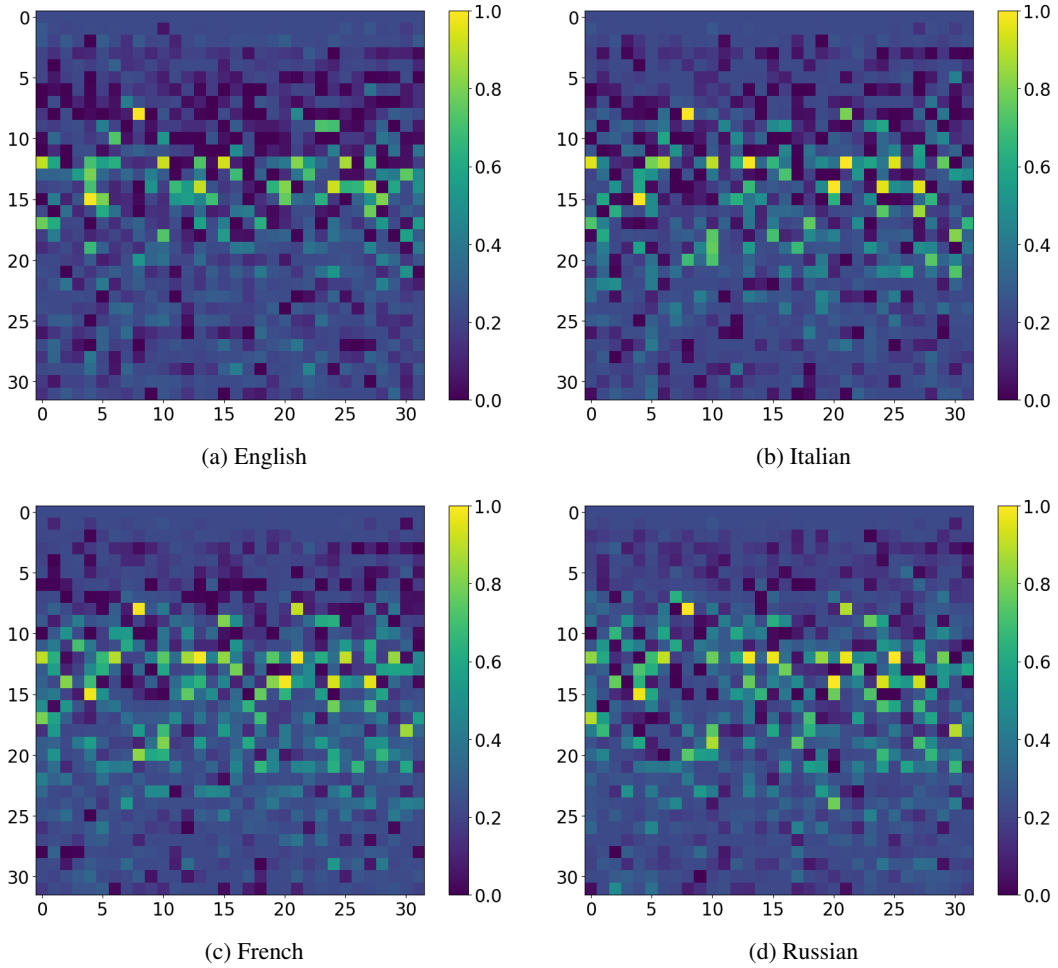


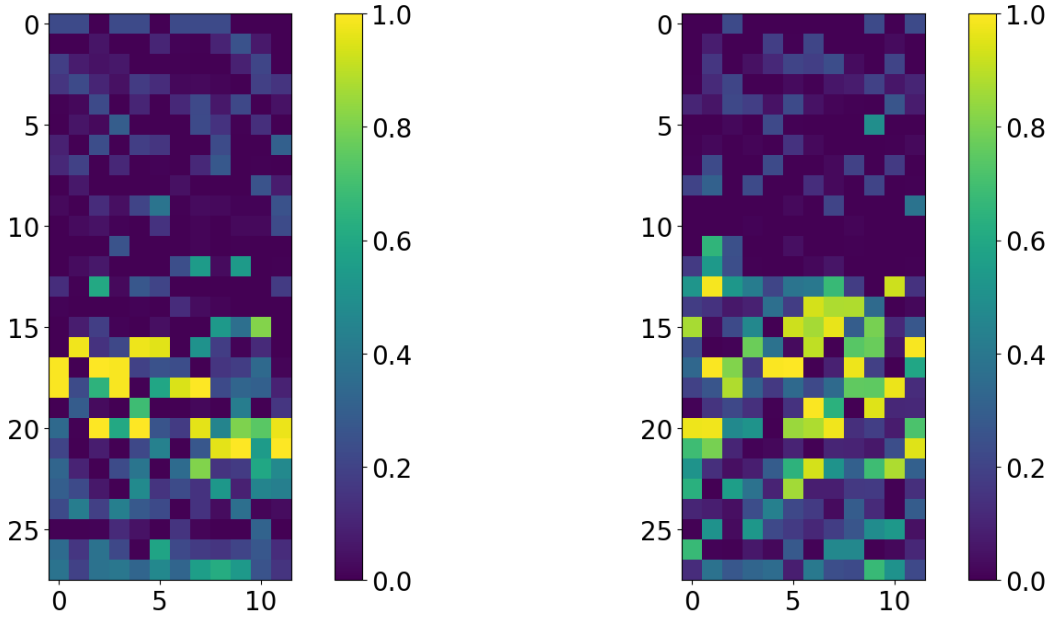
Figure 35: Distribution of predictions across options for different methods on MMLU 0-shot (upper) and 5-shot (lower) setup. (l, h) depicts the distribution for $S_{QK}^{(l,h)}$



(a) English (b) Italian

(c) French (d) Russian

Figure 36: Performance of QK-score across different heads of LLAMA-2-7B on a synthetic dataset generated in multiple languages



(a) base version

(b) instruct version

Figure 37: Performance of QK-score across different heads of Qwen 2.5-1.5B