

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. Available under the MIT license.

CosmosQA [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 the model’s reading comprehension and commonsense reasoning capabilities. Similar to MMLU, we sampled 10,000 instances from the test set. Available under CC-BY-4.0 license.

HellaSwag [Zellers et al., 2019] evaluates the commonsense reasoning capabilities of the model by selecting the best sentence completion for a given sentence prompt, given a short text as a context. We also extracted 10,000 entities from this dataset. Available under the MIT license.

HaluDialogue 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. Available under the MIT license.

We chose datasets in order to cover the main formats of questions and common NLP tasks. Since our primary intention was to focus on the investigation and interpretability of attention heads’ roles in Question Answering, we limited ourselves to these four datasets. We did not try to cover as many benchmarks as possible.

A.2 Examples of questions from datasets

Listing 1: MMLU example

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 2: CosmosQA example

Context: My house is constantly getting messy and I can’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 3: HellaSwag example

Context: A young boy is wearing a bandana and mowing a large yard. he

<https://wilburone.github.io/cosmos/>

526 Question: Which of the following is the best ending to the given context
 527 ?
 528 Options:
 529 A. is unrelieved by the weeds and is barely smiling.
 530 B. walks away from the camera as he pushes the mower.
 531 C. moves and walks the mower but gets stuck because he is engaged in
 532 a game of ping pong with another boy.
 533 D. seems to be doing a whole lot of things and talks to the camera
 534 from behind a white fence.
 535 E. I don't know.
 536 F. None of the above.
 537

Listing 4: Halu Dialogue example

538 Context: [Human]: I like Pulp Fiction. What do you think about it? [
 539 Assistant]: I love it. It was written by Roger Avary [Human]: I
 540 heard he also wrote The Rules of Attraction. Do you know who is in
 541 that movie?
 542 Question: Which of the following responses is the most suitable one for
 543 the given dialogue?
 544 Options:
 545 A. Swoosie Kurtz is in it.
 546 B. Fred Savage is in it.
 547 C. Yes, it is a drama and crime fiction as well. Do you like crime
 548 fiction stories too?.
 549 D. No, it was not made into a film. However, it was adapted into a
 550 popular Broadway musical.
 551 E. I don't know.
 552 F. None of the above.
 553
 554

Listing 5: Simple Synthetic Dataset example

555 Question: Which of the following options corresponds to " optimal "?
 556 Options:
 557 A. ion.
 558 B. optimal.
 559 C. coins.
 560 D. jackie.
 561 E. I don't know.
 562 F. None of the above.
 563
 564

565 A.3 Prompt Templates and Examples

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

Listing 6: MMLU prompt template

571 Question:_{Text of the question}?\n
 572 Options:\n
 573 A._{Text of the option A}_.\n
 574 B._{Text of the option B}_.\n
 575 C._{Text of the option C}_.\n
 576 D._{Text of the option D}_.\n
 577 E._I don't know_.\n
 578 F._None of the above_.\n
 579 Answer:
 580
 581

Listing 7: CosmosQA/HellaSwag/Halu Dialogue prompt template

```

582 Context:_{The context of the question/situation or the dialog history}\n
583 Question:_{Text of the question}?\n
584 Options:\n
585 A._{Text of the option A}_.\n
586 B._{Text of the option B}_.\n
587 C._{Text of the option C}_.\n
588 D._{Text of the option D}_.\n
589 E._I don't know_.\n
590 F._None of the above_.\n
591 Answer:
592
593

```

594 The following is an example of a 1 shot prompt from MMLU. 2-3-4-5-shot prompts were built in the
 595 same way, and prompts for datasets with context were built the same way, except each question is
 596 preceded by its context. Note that in demonstrations, we add a single whitespace between “Answer:”
 597 and the correct choice letter; for example, “Answer: A”, but *never* “Answer:A”. This is done
 598 because sequences like “: A” and “:A” are differently split into tokens by the LLaMA tokenizer.
 599 The former produces the same tokens corresponding to the letter “A” as in the choice option line,
 600 while later yields a different version of “A”. From LLaMA’s point of view, these two versions of
 601 letters are separate entities and are NOT interchangeable. Removing those symbols of whitespace
 602 often leads to a noticeable drop in performance.

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

```

603 Question: A medication prescribed by a psychiatrist for major depressive
604 disorder would most likely influence the balance of which of the
605 following neurotransmitters?\n
606 Options:\n
607 A. serotonin .\n
608 B. dopamine .\n
609 C. acetylcholine .\n
610 D. thorazine .\n
611 E. I don't know .\n
612 F. None of the above .\n
613 Answer: A\n
614 Question: Meat should be kept frozen at what temperature in Fahrenheit?\n
615 n
616 Options:\n
617 A. 0 degrees or below .\n
618 B. between 10 and 20 degrees .\n
619 C. between 20 and 30 degrees .\n
620 D. 0 degrees or below .\n
621 E. I don't know .\n
622 F. None of the above .\n
623 Answer:
624
625

```

626 B Option-representative tokens analysis

627 For our method, it is crucial to identify particular option-representative tokens $\{t_i\}$ that concentrate
 628 the semantic information of each option. Given the causal nature of the attention mechanism in
 629 LLMs, the most logical choice is the last token following the option’s content—typically the “\n”
 630 token. We use this in most of our experiments, though other tokens are also worth analysing: the
 631 label itself, the period after the label, and the period after the option content (see Figure 2 left). We
 632 also tested the mean aggregated score across all tokens in the option’s content, but this approach
 633 yielded poor results. A detailed analysis of these variations for attention scores is shown in the right
 634 part of Figure 2. Our findings indicate that the period after the content and the end-of-line token
 635 are the most representative for our scores and the task. Interestingly, while the label token is almost
 636 useless in the zero-shot setup, which is consistent with [Lieberum et al., 2023], it performs well in
 637 the five-shot setup for certain heads. We hypothesise there exist multiple types of select-and-copy
 638 heads, each influencing the logits in distinct ways.

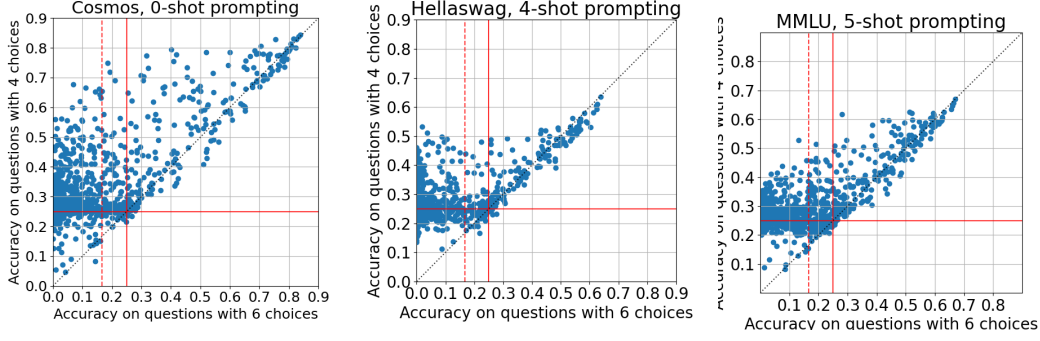


Figure 8: Correlation between heads QK-scoring accuracy on questions with 4 (‘A’-‘D’) and 6 (‘A’-‘F’) answer options. Solid red lines mark the accuracy level of 0.25, dashed red line – 0.167 (6 options random choice accuracy). Model LLaMA3.1-8B (base).

C Some more intuition on options ‘E’ and ‘F’

As mentioned in the main text, including 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 analyzing attention head roles, but that is somewhat beyond the scope of this article. Here, we would like to provide some intuition about 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 in Figure 8 show the correlation between the accuracy of heads using QK-scores on options without ‘E’-‘F’ (by y-axis) and their accuracy on questions with all six options (by x-axis). Here, only validation subsets were used. We present plots for some possible setups, but others follow similar patterns. 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 six choices and vice versa; see points around the diagonal $y = x$ in the upper-right quadrant.

We can also observe another significant 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. Therefore, 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 on 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.

The 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 the 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 that do not have MCQA in their functionality at all (they may perform other roles for LM).

Results on Figure 8 are given for LLaMA3.1-8B (base) model, but described patterns are typical for other models as well.

D Prompt design analysis

Large Language Models are sensitive to prompt design and this is a well-known issue. Appendix A.3 provides the details on the prompt templates we used in our main experiments. Those prompts do not contain instructions, so the primary intent of this section is to investigate effect of added instructions.

673 First, we considered 3 various formal style instructions to make the task clearer for the model. For
674 experiments on MMLU dataset we removed mentions of context from the instructions.

Listing 9: Explicit instructions

```
675  
676 1. Make your best effort and select the correct answer for the following  
677    question based on the context. You only need to output the option.  
678 2. Think logically and select the correct answer for the following  
679    question based on the context. You only need to output the option.  
680 3. Answer the question based on the context. Select the option you are  
681    most confident at.  
682
```

683 The questions prompted with instructions were built according to the modified template:

Listing 10: Modified template for prompt with instruction

```
684  
685 Instruction\n  
686 Context: \n [for MMLU this line is omitted]  
687 Question: {Text of the question}?\n  
688 ...  
689
```

690 We conducted experiments with ‘instructioned’ prompts in 0-shot setup for several models. The
691 setup was similar to the main experiments reported in this paper. The only difference was that
692 for evaluation we used a 1,000-sample subset of the original evaluation set (it was done to reduce
693 computational time; the class balance was preserved).

694 Normally, for each prompt individually, we selected the best head on the calibration data subset
695 and measured the performance of QK-scores from that head on the evaluation data. However, we
696 observed further stability of the QK-score method. Apparently, QK-score on heads chosen with the
697 default prompts (i.e. without instructions), can achieve near the same level of performance even
698 if prompt contains explicit instruction. And so, we took best heads found on calibration set with
699 no-instruction prompts (e.g. head (16, 19) for LLaMA3.1-8B on MMLU, and head (17, 24) on Hala
700 Dialogue, or head (14, 24) for LLaMA2-7B on MMLU) and evaluated their quality on calibration
701 sets when prompts are given with instructions. In the table this experiment is given in "QK-score +
702 fixed head" rows. We report results only for LLaMA2-7B-base/-chat, but for other models the overall
703 picture is similar.

704 The results are presented in Table 3. We see that in 0-shot setup on all datasets clear instructions in
705 the prompt improve the performance of both Baseline and QK-score methods compared to our default
706 prompts. However, this increase is not uniform and is much more pronounced for chat-/instruct-
707 tuned versions. Also, it can be noted that QK-score often has lower variance in both metrics. For the
708 family of LLaMA2 models, QK-score with ‘fixed head’ achieves the same level of performance as
709 QK-score with heads selected for each prompt individually. For the chat-tuned model there is some
710 difference between them (‘fixed head’ slightly degrades the performance).

Model	Method		MMLU	Cosmos QA	Hellaswag QA	Halu Dialogue
Qwen2.5-1.5B	Baseline	Acc	59.5 \pm 0.6	74.3 \pm 1.2	58.1 \pm 2.2	35.9 \pm 2.1
		PA	49.3 \pm 0.4	67.8 \pm 1.6	48.6 \pm 2.9	25.7 \pm 1.7
	QK-score	Acc	57.0 \pm 1.3	72.1 \pm 0.9	58.7 \pm 1.0	38.3 \pm 0.9
		PA	45.9 \pm 1.1	64.2 \pm 0.8	48.6 \pm 0.3	27.5 \pm 0.8
LLaMA2-7B	Baseline	Acc	29.2 \pm 1.5	37.3 \pm 1.2	29.4 \pm 0.2	26.5 \pm 1.1
		PA	10.6 \pm 0.5	17.3 \pm 0.9	9.1 \pm 0.1	7.2 \pm 1.2
	QK-score	Acc	32.2 \pm 2.6	51.9 \pm 0.3	37.4 \pm 0.5	35.6 \pm 0.3
		PA	13.5 \pm 1.9	33.8 \pm 0.7	18.4 \pm 1.6	12.8 \pm 0.9
	+ Fixed Head	Acc	33.0 \pm 1.5	51.9 \pm 0.3	36.4 \pm 1.5	34.4 \pm 1.5
		PA	15.9 \pm 1.0	33.8 \pm 0.7	17.9 \pm 2.0	14.5 \pm 1.5
LLaMA2-7B, chat	Baseline	Acc	44.1 \pm 1.1	62.1 \pm 0.1	42.9 \pm 0.4	25.8 \pm 1.6
		PA	28.6 \pm 0.8	47.0 \pm 0.9	25.0 \pm 1.5	12.6 \pm 0.8
	QK-score	Acc	44.9 \pm 0.4	64.0 \pm 1.2	49.5 \pm 0.6	44.0 \pm 0.5
		PA	30.1 \pm 1.2	51.0 \pm 1.4	34.6 \pm 1.1	25.2 \pm 0.8
	+ Fixed Head	Acc	44.5 \pm 0.8	62.8 \pm 0.4	44.4 \pm 0.8	35.8 \pm 0.6
		PA	28.9 \pm 0.1	46.3 \pm 0.9	30.1 \pm 1.5	16.6 \pm 0.8
LLaMA3-8B, Instruct	Baseline	Acc	61.1 \pm 0.5	84.5 \pm 0.5	67.9 \pm 0.8	60.7 \pm 0.5
		PA	52.4 \pm 0.6	79.3 \pm 0.7	59.3 \pm 0.9	51.1 \pm 1.0
	QK-score	Acc	61.5 \pm 0.1	88.3 \pm 0.1	70.8 \pm 1.9	65.3 \pm 1.4
		PA	52.9 \pm 0.0	84.4 \pm 0.2	61.8 \pm 2.5	53.5 \pm 4.4
LLaMA3.1-8B	Baseline	Acc	60.5 \pm 0.4	78.7 \pm 0.9	51.7 \pm 1.6	55.3 \pm 1.7
		PA	46.8 \pm 0.2	70.9 \pm 0.8	35.5 \pm 3.0	41.2 \pm 1.7
	QK-score	Acc	64.2 \pm 0.6	82.6 \pm 0.5	67.9 \pm 0.4	58.5 \pm 2.9
		PA	53.8 \pm 0.4	76.6 \pm 0.8	57.4 \pm 1.1	42.7 \pm 3.0
	+ Fixed Head	Acc	64.2 \pm 0.6	82.6 \pm 0.5	58.4 \pm 0.9	54.9 \pm 1.8
		PA	53.0 \pm 0.7	75.4 \pm 0.6	45.0 \pm 1.3	39.4 \pm 1.9
LLaMA3.1-8B, Instruct	Baseline	Acc	66.1 \pm 1.5	88.0 \pm 1.1	71.8 \pm 1.8	57.1 \pm 5.2
		PA	54.1 \pm 1.5	83.5 \pm 1.5	62.3 \pm 2.8	47.5 \pm 5.3
	QK-score	Acc	67.0 \pm 0.6	89.7 \pm 0.6	72.0 \pm 1.0	69.7 \pm 1.8
		PA	57.6 \pm 1.2	85.5 \pm 1.1	63.6 \pm 1.8	55.3 \pm 2.2
	+ Fixed Head	Acc	67.1 \pm 0.6	89.9 \pm 0.4	73.0 \pm 0.3	71.0 \pm 1.6
		PA	57.1 \pm 0.6	85.7 \pm 0.8	65.1 \pm 0.5	53.7 \pm 1.4

Table 3: Performance of Baseline and QK-score methods on prompts with explicit instructions. 0-shot setup.

Next, we explore more exotic examples. As it was noted in [Mozikov et al. \[2024\]](#), behaviour of LLM’s may change when models are prompted with different emotional states. Inspired by this work, we collected a set of 11 creative instructions: 8 instructions asking the model to emulate an emotion and 3 instructions asking to answer the question while roleplaying a fictional character. The setup of experiments was the same as the one used for explicit instructions. The results are presented in Table 4. Here we can see that, in general, our creative instructions result in the decrease of performance of both methods, and once again, this effect is more noticeable for chat-/instruct-tuned models.

As for prompts with clear instructions, evaluation performance of QK-score on best head fixed on the ‘no-instruction’ prompt is close to performance of QK-score on best head selected specifically for the prompt. This observation suggests that heads chosen for QK-scoring in our method are stable across a wide range of possible prompts (thus justifying the omission of clear instructions in our main experiments is justified).

Model	Method		MMLU	Cosmos QA	Hellaswag QA	Halu Dialogue
Qwen2.5-1.5B	Baseline	Acc	56.9 ± 2.9	71.4 ± 2.7	56.7 ± 3.7	38.4 ± 2.5
		PA	45.9 ± 3.2	64.2 ± 3.2	47.3 ± 3.8	27.2 ± 2.5
	QK-score	Acc	55.3 ± 2.0	70.6 ± 1.3	56.1 ± 1.8	38.2 ± 2.5
		PA	43.3 ± 1.9	61.3 ± 1.4	45.8 ± 2.9	23.6 ± 5.7
LLaMA2-7B	Baseline	Acc	29.2 ± 1.8	29.6 ± 5.0	27.3 ± 1.3	23.9 ± 1.2
		PA	10.0 ± 1.5	11.5 ± 4.2	8.1 ± 0.9	5.7 ± 0.8
	QK-score	Acc	28.2 ± 1.8	47.0 ± 6.0	38.3 ± 1.6	34.4 ± 1.4
		PA	10.5 ± 2.0	27.6 ± 7.2	20.2 ± 2.0	13.7 ± 2.1
	+ Fixed Head	Acc	28.9 ± 4.6	47.0 ± 5.9	33.5 ± 4.6	31.7 ± 2.5
		PA	13.3 ± 2.9	28.3 ± 5.9	15.5 ± 3.5	13.6 ± 2.2
LLaMA2-7B, chat	Baseline	Acc	38.2 ± 6.4	55.4 ± 4.6	36.7 ± 4.5	22.8 ± 4.9
		PA	23.3 ± 5.4	39.2 ± 5.8	19.5 ± 3.6	10.7 ± 2.5
	QK-score	Acc	39.0 ± 3.3	55.4 ± 4.6	36.7 ± 4.5	40.0 ± 3.0
		PA	22.6 ± 4.8	44.2 ± 5.9	25.7 ± 3.7	20.3 ± 4.0
	+ Fixed Head	Acc	34.7 ± 7.9	57.8 ± 3.9	37.7 ± 1.5	34.8 ± 3.4
		PA	21.2 ± 5.3	37.5 ± 6.9	22.0 ± 2.1	16.4 ± 1.9
LLaMA3-8B, Instruct	Baseline	Acc	56.2 ± 5.0	80.0 ± 4.1	55.8 ± 14.3	55.2 ± 9.2
		PA	47.6 ± 4.7	74.4 ± 4.6	46.5 ± 14.2	46.0 ± 9.3
	QK-score	Acc	60.8 ± 1.8	86.3 ± 1.2	64.8 ± 5.9	63.4 ± 5.7
		PA	52.8 ± 2.8	81.5 ± 2.0	55.7 ± 6.8	52.6 ± 7.8
LLaMA3.1-8B	Baseline	Acc	55.5 ± 1.7	75.6 ± 1.5	33.2 ± 9.5	49.6 ± 4.1
		PA	43.5 ± 2.3	67.2 ± 1.9	18.2 ± 6.9	35.3 ± 4.2
	QK-score	Acc	62.9 ± 1.1	80.6 ± 1.0	62.2 ± 2.2	57.4 ± 2.4
		PA	52.1 ± 1.7	73.4 ± 1.3	49.9 ± 2.8	42.1 ± 3.5
	+ Fixed Head	Acc	63.2 ± 1.0	80.6 ± 1.1	47.8 ± 5.6	50.2 ± 2.8
		PA	52.0 ± 1.2	72.1 ± 1.3	34.9 ± 5.3	35.2 ± 3.3
LLaMA3.1-8B, Instruct	Baseline	Acc	60.9 ± 6.3	82.7 ± 4.1	54.9 ± 18.3	42.2 ± 14.2
		PA	49.8 ± 5.9	77.6 ± 4.4	44.4 ± 17.7	32.9 ± 13.3
	QK-score	Acc	65.5 ± 2.0	88.2 ± 1.1	64.1 ± 8.1	64.9 ± 5.6
		PA	55.9 ± 2.2	83.3 ± 1.1	53.8 ± 8.5	48.6 ± 6.2
	+ Fixed Head	Acc	65.5 ± 2.0	88.0 ± 1.2	63.6 ± 10.4	65.6 ± 5.7
		PA	55.6 ± 1.8	82.8 ± 1.7	54.1 ± 10.0	44.8 ± 7.5

Table 4: Performance on prompts asking to emulate an emotion (or roleplay a character) and then answer the question. 0-shot setup

E Numerical results for comparison of QK-score with other methods

Table 5 provides numerical results for our main experiments with QK-scores from heads of the LLaMA3.1-8B (base) model that are presented in Figure 3 in the main text. As can be seen from the table, for MMLU and CosmosQA PRIDE shows the best performance, while QK-score yields results slightly worse (though within reasonable tolerance). It can be explained by the fact that PRIDE was initially aimed at improving the model performance by increasing stability of its predictions, and QK-score is more an analysis tool than a method for evaluation improvement.

In the same way, Table 6 provides numerical results for the LLaMA2-7B (base) model for comparison.

F Best Heads

We utilized the minimum accuracy percentiles to determine stable heads that can be seen in Figures 9 and 10. The head is considered “stable” if (1) *QK-score* gives both high average accuracy across tasks and (2) accuracy on each task is better than at least 90% of heads. From these figures, we can see that

Method		...-shot prompting					
		0	1	2	3	4	5
MMLU							
Baseline	Acc	59.0	63.3	63.3	62.6	62.7	63.7
	PA	48.4	52.6	52.9	52.1	52.0	53.3
PRIDE	Acc	62.4	64.0	63.3	63.0	63.4	64.1
	PA	52.9	53.8	54.2	53.7	53.5	54.1
Attention score	Acc	62.7	62.1	62.1	62.5	62.6	62.9
	PA	51.0	51.1	50.9	52.0	50.4	52.3
QK-score	Acc	63.4	62.1	61.8	62.5	62.7	62.0
	PA	52.8	52.0	51.0	53.6	52.8	53.4
Synthetic	Acc	61.7	60.8	60.0	62.9	63.7	63.5
	PA	53.6	52.8	51.9	52.9	53.5	53.4
Cosmos QA							
Baseline	Acc	80.6	86.2	86.4	86.1	86.5	86.1
	PA	73.0	80.7	80.4	80.4	80.7	80.5
PRIDE	Acc	84.7	86.5	86.8	86.6	86.9	87.0
	PA	79.2	81.9	82.2	82.2	82.6	82.5
Attention score	Acc	84.2	85.5	85.5	85.2	85.2	85.0
	PA	76.6	80.1	80.0	77.3	78.0	77.9
QK-score	Acc	84.0	86.5	85.6	85.9	85.6	85.4
	PA	78.3	80.6	80.0	80.1	78.8	77.8
Synthetic	Acc	83.1	84.5	84.2	83.5	84.1	84.0
	PA	77.6	80.6	78.9	78.3	79.2	79.4
Hellaswag QA							
Baseline	Acc	38.8	60.7	56.5	65.1	63.2	63.6
	PA	22.2	46.2	41.2	51.6	49.8	50.8
PRIDE	Acc	63.0	62.5	61.1	67.7	64.8	63.7
	PA	50.9	49.4	47.6	56.7	53.0	52.1
Attention score	Acc	59.5	64.1	64.7	66.8	66.1	64.9
	PA	47.9	54.2	54.9	54.6	54.5	54.8
QK-score	Acc	63.4	64.1	63.4	64.2	65.7	64.9
	PA	53.6	54.2	53.8	55.6	56.5	55.8
Synthetic	Acc	49.1	51.0	60.0	62.5	63.7	63.5
	PA	37.7	38.1	52.2	55.0	55.0	54.5
Halu Dialogue							
Baseline	Acc	51.2	50.9	60.6	61.7	61.2	61.3
	PA	37.0	39.2	49.6	50.8	47.7	49.1
PRIDE	Acc	58.3	51.5	61.1	62.5	62.5	62.3
	PA	47.4	40.1	50.9	52.8	51.6	52.3
Attention score	Acc	56.6	56.3	65.2	67.6	68.6	67.3
	PA	44.8	43.2	53.9	56.4	57.9	52.5
QK-score	Acc	57.0	54.5	65.2	67.7	67.6	67.0
	PA	42.6	41.6	49.9	57.0	53.7	53.1
Synthetic	Acc	51.3	51.0	60.1	62.9	63.6	62.9
	PA	37.5	39.4	50.2	52.7	53.7	53.0

Table 5: Comparison of different methods for LLaMA3.1-8B (base) on various Q&A datasets. Reported metrics are Accuracy (Acc) and Permutation Accuracy (PA). The best results are highlighted in **bold**.

LLaMA-3-8B has more “stable” heads than LLaMA-2-7B, and LLaMA-3.1-8B, in turn, surpasses LLaMA-3-8B in the number of “stable” heads.

For LLaMA-2-7B (Figure 10(a)), the heads from the 14th layer show the highest accuracy on almost all percentiles again. We also listed the top 1% pairs for all datasets based on accuracy in Table 7. There is a noticeable overlap between heads for various setups, and, once again, all of them are in middle layers of the model.

		...-shot prompting					
Method		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 6: Comparison of different methods for LLaMA2-7B (base) on various Q&A datasets. Reported metrics are Accuracy (Acc) and Permutation Accuracy (PA). The best results are highlighted in **bold**.

Dataset	Best (Layer, Head)
MMLU	(16, 19), (17, 24), (16, 26), (17, 26)
HaluDialogue	(14, 5), (14, 21), (14, 2), (17, 24)
HellaSwag	(14, 5) (17, 24), (17, 29), (16, 19)
CosmosQA	(17, 24), (16, 19), (16, 26), (17, 29)

Table 7: Common top 1% heads across all n -shots based on accuracy for different datasets for LLaMA-3.1-8B

If we compare the performance of the “stable” heads with results obtained with preceding calibration in Figure 11, (14, 24) and (14, 20) are frequently chosen from the validation set. However, even when they do not, their performance is comparable to that of their validation-chosen counterparts, except for HaluDialogue. Besides, we tested the heads (14, 24), (14, 20), (14, 26), and (14, 13) for stability against increasing the number of options in SSD dataset (see Figure 14) and against changing the symbols that denote options, following Alzahrani et al. [2024] (see Appendix H). We also added other heads performing well on the SSD dataset to these plots for comparison.

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

Table 8: Common top 1% heads across all n -shots based on accuracy for different datasets for LLaMA-2-7B

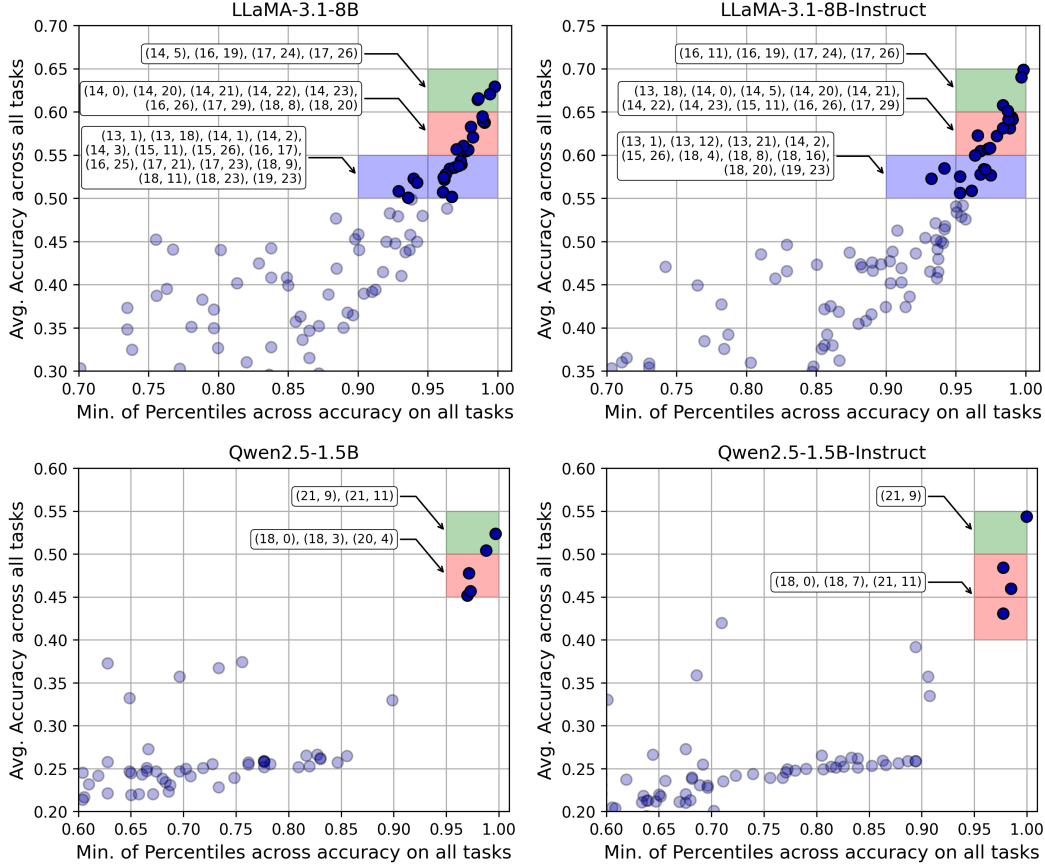


Figure 9: Stable heads for QK -score for 0-shot setup across all tasks for LLaMA-3.1-8B and Qwen2.5-1.5B (Base and Instruct). “ k -th Minimum of Percentiles” means that the head is better than k share of all heads for all tasks.

For LLaMA-3.1-8B-base, most heads from the right-top corner of the left part of Figure 9 were already seen in the Figure 4 as heads in a green frames, i.e. top-5% across all four real datasets (namely, heads (16, 26), (18, 20), (14, 1), (14, 22), (17, 24), (16, 19), (14, 0), (18, 8), (14, 21), (17, 26), (19, 23), (17, 29), (18, 9), (14, 2), (14, 5), (14, 23), (14, 20)). Several of these heads also turned out to be very stable on the synthetic dataset, even when the number of options was increased (see Figure 14 for examples of such heads) and when the language was changed (see Table 10).

We also provided the results for LLaMA-3.1-8B-Instruct at the right part of the Figure 9 to compare the “stable head distribution” with the base version. Interestingly, there is a big intersection between best stable heads of base and Instruct versions of the model.

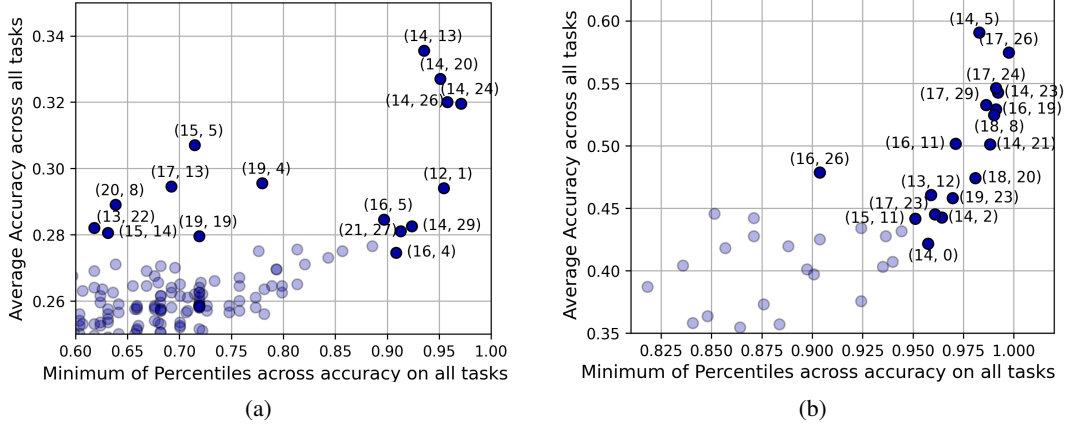


Figure 10: Stable heads for QK-score in (a) LLaMA2-7B and (b) LLaMA3-8B for 0-shot setup across all tasks. “ k -th Minimum of Percentiles” means that the head is better than k share of all heads for all tasks.

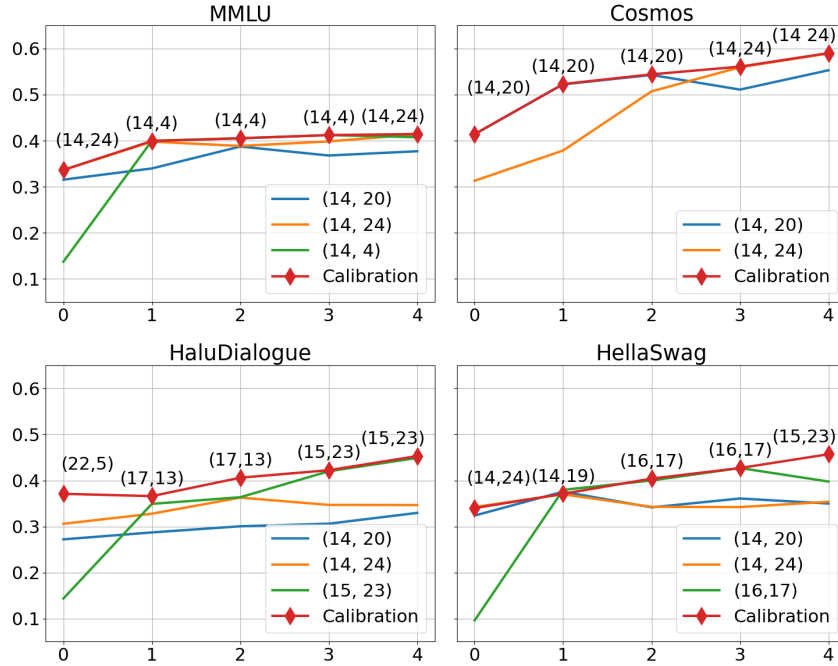


Figure 11: Accuracy of the best performing heads and several of the most robust heads – (14, 24), (14, 20) in LLaMA-2-7B

757 G Heads ablation

758 In this section, we present the results of zero-ablation for the select-and-copy mechanism applied
 759 to two additional models: LLaMA-3.1-8B-Instruct (Fig. 12) and DeepSeek-R1-Distill-Qwen-7B
 760 (Fig. 13). Attention heads were selected based on above-random performance according to the
 761 QK-score. For comparison, we also include results from random ablation, where an equal number of
 762 attention heads were ablated at random.

763 Random ablations were conducted across five independent runs, and we report the mean and standard
 764 deviation of model accuracy.

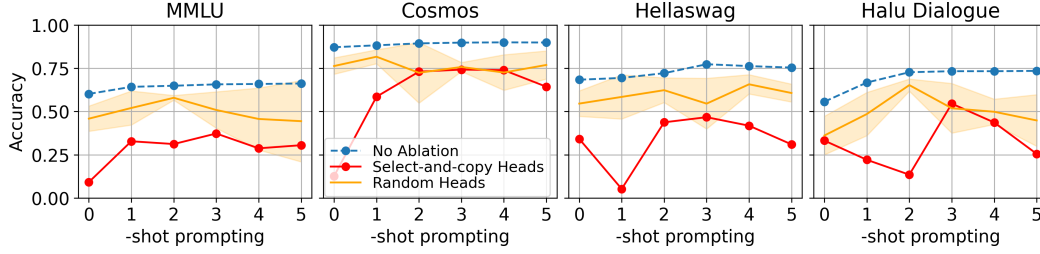


Figure 12: Zero-ablation of *select-and-copy* and random heads for LLaMA3.1-8B-Instruct

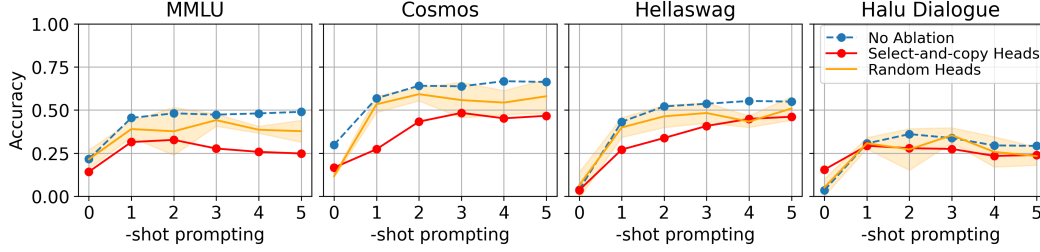


Figure 13: Zero-ablation of *select-and-copy* and random heads for DeepSeek-R1-Distill-Qwen-7B

765 Although performance degradation due to zero-ablation remains substantial in finetuned and aligned
 766 models, in some cases it is comparable to that observed in random ablation. We hypothesize that
 767 this may suggest the involvement of other attention heads (associated with in-context learning or
 768 alignment) in the decision-making process.

769 H Behaviour of the best heads under the change of options symbols and 770 options amount

771 Aside from the standard version of the Simple Synthetic Dataset (SSD), which includes four essential
 772 options and two additional options, “E” and “F” (described in Section 5.1), we also considered
 773 alternative versions of the SSD with varying numbers of possible options. For instance, the version
 774 corresponding to the number “10” on the x-axis of Figure 5 contains ten essential options (A, B,
 775 C, D, E, F, G, H, I, J) and two special options: “K. I don’t know” and “L. None of the above” (see
 776 Example 11). In these experiments, we used 200 examples from each version of the dataset to
 777 compute the attention scores.

Listing 11: Modification of SSD with ten options - example

```

778 Which of the following options corresponds to " mediterranean "?
779 Options:
780   A: acceptance
781   B: specialties
782   C: charitable
783   D: typically
784   E: access
785   F: jose
786   G: findlaw
787   H: colonial
788   I: mediterranean
789   J: data
790   K: I don't know.
791   L: None of the above.
  
```

794 Figure 14 is an extended version of Figure 5 showing more heads for LLaMA2-7B (left), LLaMA3-
 795 8B (center) and LLaMA3.1-8B (right). The heads for this figure are taken from the upper right
 796 sections of Figures 10 and 9 as they are the most stable across real datasets.

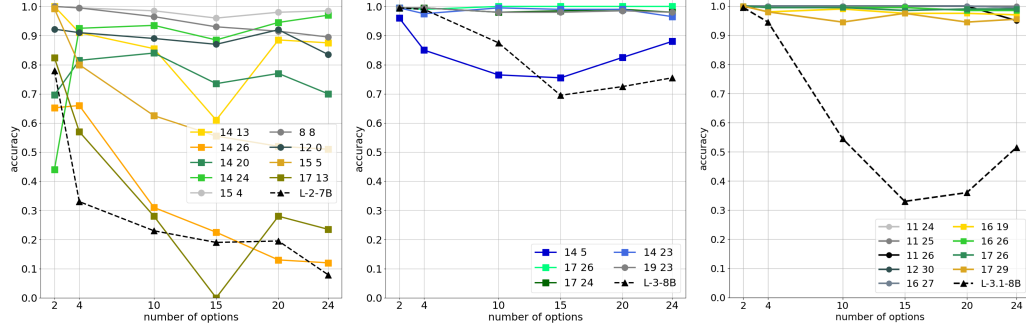


Figure 14: The results for various numbers of options in the Simple Synthetic Dataset (SSD) in a zero-shot setting are shown for LLaMA2-7B (left), LLaMA3-8B (center) and LLaMA3.1-8B (right). Different line colors represent the QK dot products from different heads. “Square” markers indicate heads that perform well across real datasets, while “round” markers represent heads that perform well on the synthetic dataset. Interestingly, more heads from newer versions of the LLaMA model show stability under increasing the number of options.

In Figures 15 and 16, we return to the standard 4-option SSD dataset but use different symbols for the option labels. The upper plot includes the renamed special options “E” for “I don’t know” and “F” for “None of the above”, while the lower plot omits them for the LLaMA2-7B model.

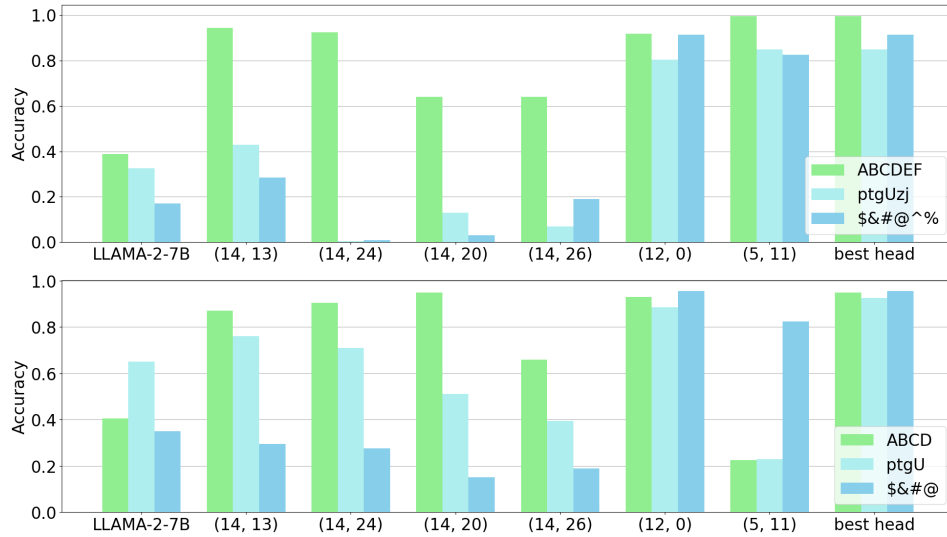


Figure 15: Performance of the QK-score from the best heads of the LLaMA2-7B for different option symbols, with “uncertainty” options (i.e. “I don’t know” and “None of the above”) presented (upper figure) and not presented (lower figure)

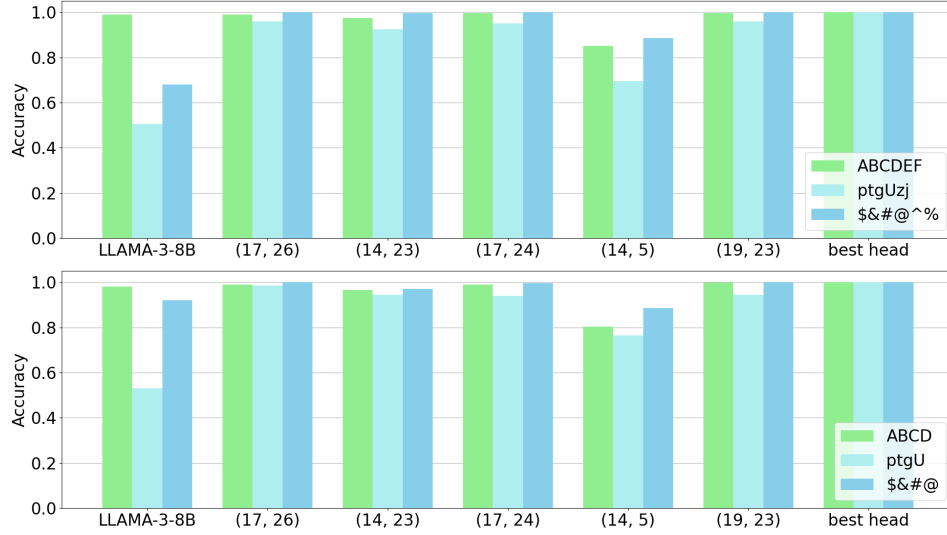


Figure 16: Upper row: performance of the QK-score from the best heads of the LLaMA3-8B for different option symbols, with “uncertainty” options (i.e. “I don’t know” and “None of the above”) presented (upper figure) and not presented (lower figure)

808 The accuracy of the best four heads of LLaMA-2-7B in Figure 10a declines in these new setups, but
809 the head (12, 0) remains stable across all setups. Another interesting head is (5, 11): its accuracy
810 is high for all setups with “uncertainty” and for “\$&#” setup, but drops abruptly for “ABCD” and
811 “ptgU”. Studying such “anomalies” is a subject for future research.

804 We plot the results from the same setup for the LLaMA3-8B model in Figure 16. Interestingly, the
805 best heads of the newer LLaMA3-8B model (see Figure 10b) exhibit significantly greater stability
806 across the evaluated setups compared to those of the older LLaMA2 model.

807 I Synthetic dataset in different languages

808 We regenerated our synthetic dataset using three languages in addition to English. Figure 17 shows that
809 the general distribution of QK-scores across heads of the LLaMA2-7B model on these multilingual
810 datasets remains largely unchanged; for example, layers 8–15 still contain the most performant heads.
811 However, differences in the performance of individual heads are also observed. Additionally, we
812 show accuracies for top-10 performant heads in Table 9. We coloured green the heads that perform
813 best across four real datasets (see Figure 10). Additionally, we highlighted in bold the heads that
814 appear in the top-10 for all four languages.

815 As shown, 7 out of the top-10 best heads are shared across synthetic datasets in different languages,
816 including two “green” heads that are also the best across our real datasets. This significant overlap
817 suggests a substantial degree of universality among the identified heads. Interestingly, the QK-scores
818 for the best heads are somewhat lower for English compared to the other languages we analysed.
819 However, we cannot draw definitive conclusions from this observation without further investigation.
820 A more thorough study of how QK-scores and the best-performing heads vary with the dataset’s
821 language remains a topic for future research and is beyond the scope of this paper.

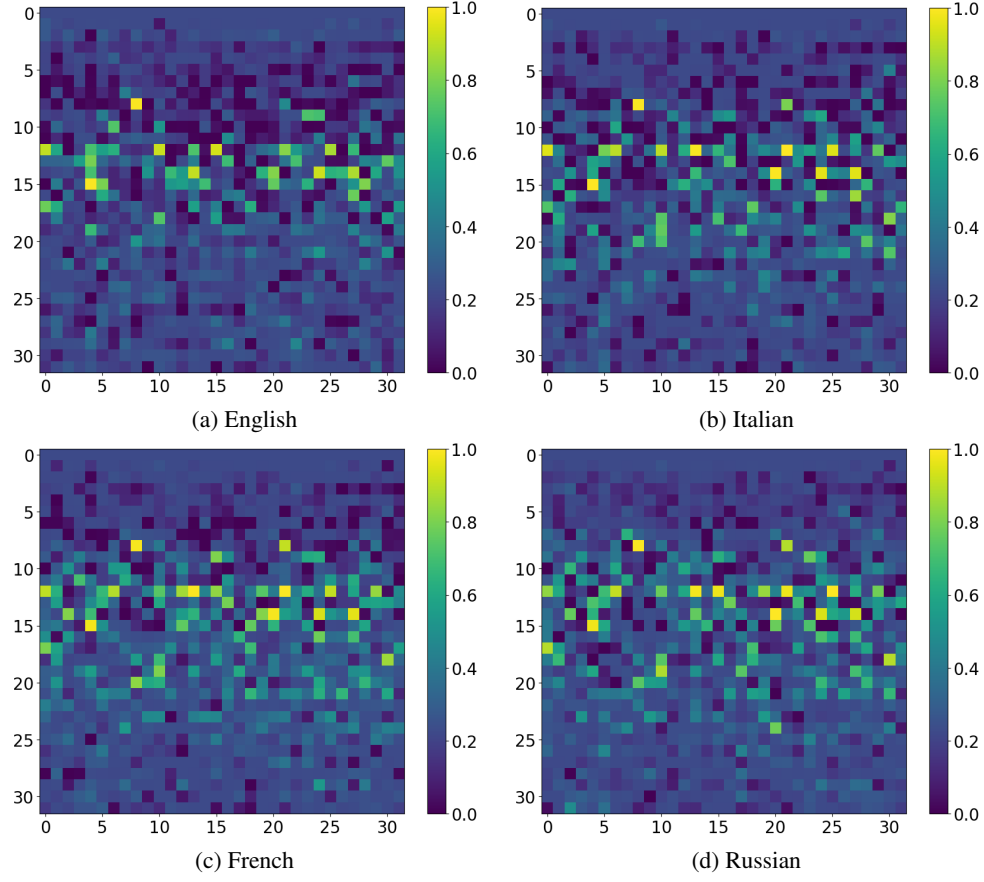


Figure 17: Performance of QK-score across different heads of LLaMA2-7B on a synthetic dataset generated in multiple languages

Language	Best (Layer, Head)	Max Acc	Min Acc
English	(8, 8), (15, 4), (12, 15), (14, 24), (12, 10) (14, 13), (14, 27), (12, 25), (12, 21), (14, 20)	0.995	0.815
Italian	(12, 21), (15, 4), (8, 8), (14, 20), (12, 13) (14, 24), (14, 27), (12, 0), (12, 25), (12, 10)	1.000	0.900
French	(12, 21), (12, 13), (8, 8), (14, 20), (15, 4) (14, 27), (14, 24), (8, 21), (12, 25), (12, 0)	1.000	0.905
Russian	(12, 25), (14, 20), (8, 8), (12, 13), (12, 15) (12, 21), (15, 4), (14, 24), (14, 27), (12, 6)	1.000	0.910

Table 9: Top-10 best heads per language for LLaMA 2-7B, sorted by decreased accuracy

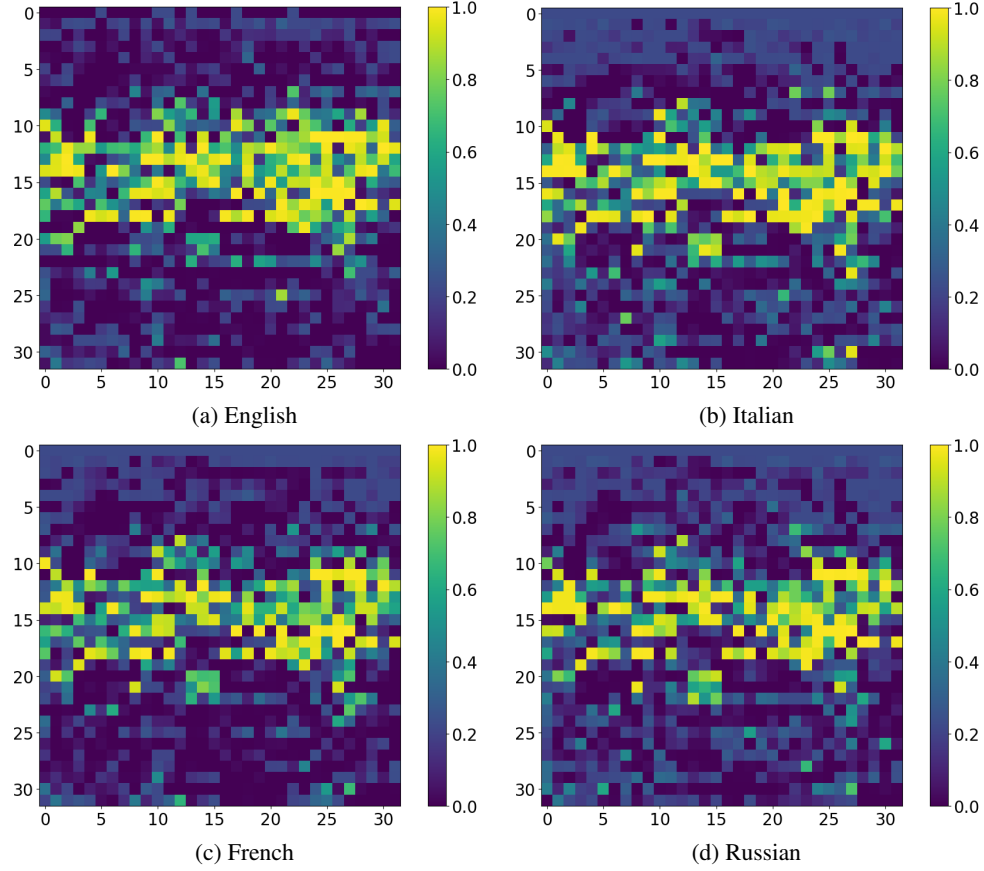


Figure 18: Performance of QK-score across heads of LLaMA3.1-8B on a multi-language SSD

Language	Best (Layer, Head)	Max Acc	Min Acc
English	(11, 24), (11, 25), (11, 26), (12, 30), (14, 6) (15, 26), (16, 24), (16, 27), (17, 28), (11, 10) (11, 27), (11, 28), (12, 14), (13, 2), (13, 9) (13, 10), (13, 16), (13, 23), (16, 17), (16, 25) (16, 26), (17, 26), (18, 9), (13, 13), (14, 2) (15, 10), (17, 21), (18, 8), (18, 18), (12, 20)	1.000	0.985
Italian	(11, 24), (11, 25), (11, 27), (12, 14), (12, 30) (13, 9), (13, 10), (13, 13), (13, 15), (13, 16) (14, 18), (15, 26), (16, 17), (16, 19), (16, 27) (17, 21), (18, 9), (18, 18), (11, 4), (11, 10) (11, 28), (13, 1), (14, 2), (14, 13), (14, 30) (16, 10), (16, 11), (16, 21), (16, 26), (18, 24)	1.000	0.995
French	(11, 4), (16, 27), (18, 9), (11, 24), (11, 25) (11, 27), (11, 28), (12, 14), (13, 9), (16, 17) (16, 19), (16, 24), (16, 26), (17, 26), (11, 10) (12, 30), (13, 16), (10, 0), (13, 13), (14, 6) (16, 21), (17, 29), (18, 11), (13, 1), (16, 8) (16, 25), (17, 21), (18, 8), (12, 28), (13, 15)	1.000	0.975
Russian	(11, 25), (11, 27), (14, 6), (16, 27), (17, 28) (11, 24), (11, 26), (12, 30), (13, 13), (13, 16) (16, 19), (16, 24), (16, 25), (16, 26), (17, 29) (18, 9), (13, 15), (14, 14), (14, 21), (16, 8) (16, 17), (17, 26), (18, 8), (18, 18), (13, 9) (13, 10), (14, 0), (14, 3), (15, 20), (15, 24)	1.000	0.985

Table 10: Top-30 best heads per language for LLaMA 3.1-8B, sorted by decreased accuracy

Figure 18 shows that similar patterns are observed for LLaMA3.1-8B, with the key difference being that the average quality of the heads in the middle layers is significantly higher than in LLaMA2-7B. Table 10 lists the top 30 heads for LLaMA3.1-8B, and again, we observe several heads that are universal across languages (marked in bold). As with LLaMA2-7B, we highlight in green the heads that perform best across four real-world datasets. We report the top 30 heads instead of the top 10 because LLaMA3.1-8B contains a large number of “good” heads with near-perfect performance – substantially more than LLaMA2-7B – making a top-10 selection unrepresentative for this model.

J Best heads on synthetic dataset for Qwen 2.5-1.5B

In Figure 19, we present the accuracy of the QK-score for each head of Qwen 2.5-1.5B-base and -instruct on our synthetic datasets in four languages. These diagrams confirm that the layers with the best heads in Qwen 2.5-1.5B are closer to the final layer than in LLaMA-family models. Specifically, the best-performing heads in Qwen 2.5 are concentrated in layers 16–22, out of a total of 28 layers. Besides, earlier layers contain many heads with performance close to zero. Despite these differences, the overall pattern resembles the corresponding heatmaps for LLaMA-2-7B and LLaMA-3.1-8B shown in Figures 17 and 18, particularly in very early and late layers as both do not contain strongly pronounced select-and-copy heads. We also see that some especially “stable” heads, such as (18, 3), (20, 4) and (21, 11), which perform well across real datasets in Qwen 2.5-1.5B-base, remain among the best heads for the synthetic dataset in all languages for both Qwen 2.5-1.5B-base and Qwen 2.5-1.5B-Instruct models, as shown in Figure 19. This again illustrates the persistence of some select-and-copy heads across many datasets and languages, as discussed in our paper.

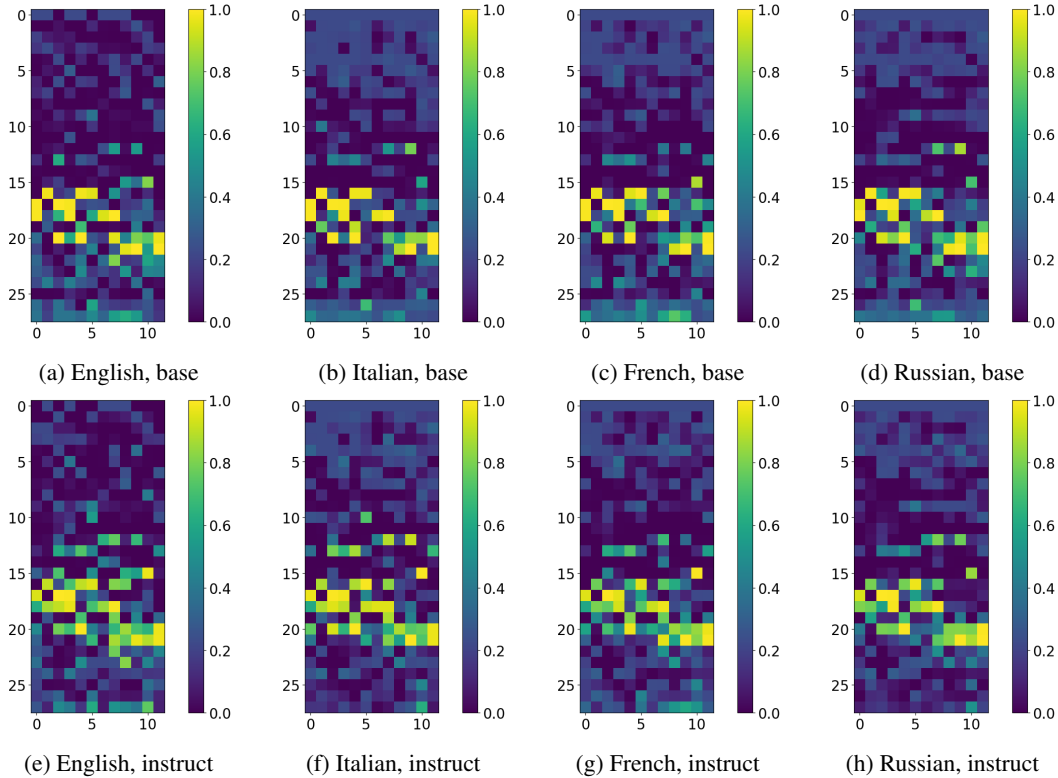


Figure 19: Performance of QK-score across different heads of Qwen 2.5-1.5B-base (upper row) and Qwen 2.5-1.5B-Instruct (lower row) on a multi-language SSD

842 K Head scoring without validation set

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

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

843 where \hat{i} denotes the most frequent option for the given head; head indices (l, h) are omitted. The
 844 left component represents the average amount of attention concentrated on the option-representative
 845 tokens $t_i, i = 1, \dots, n$. The right component reflects the frequency of the situation, when the largest
 846 attention among the options falls on the option other than \hat{i} , i.e., any option other than the most
 847 frequent one.

848 The results of ranking heads according to these scores for LLaMA models are presented in Figure 20
 849 We applied similar technique for Qwen-2.5-1.5B as well; the results are presented in the Figure 21.
 850 Here, we calculated the attention score without applying RoPE.

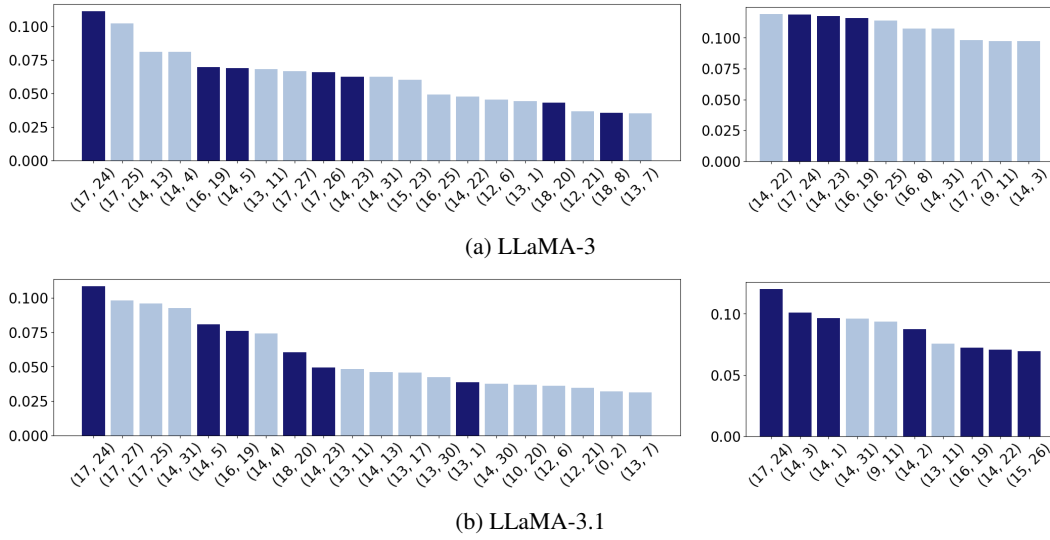


Figure 20: Left: average top heads scores of LLaMA models across real datasets (first twenty). Right: Top head scores on the Simple Synthetic Dataset (first ten). Dark blue marks the best heads, stable across all real datasets (i.e. heads from top-right corner of the Figures 9 and 10). Note that for calculating this score, we did not use the dataset labels.

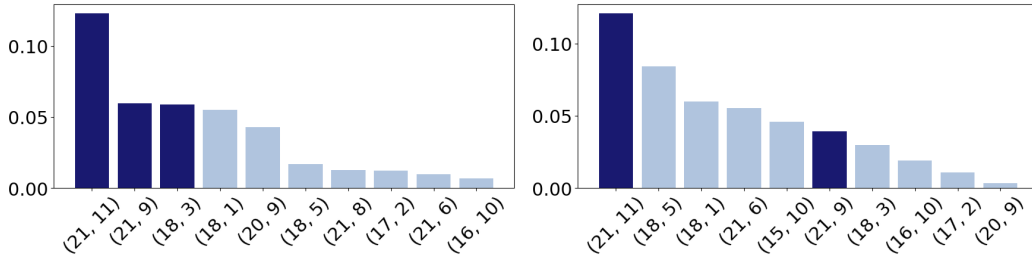


Figure 21: Left: average top heads scores for Qwen-2.5-1.5B-base across real datasets (first ten). Right: the same for Qwen-2.5-1.5B-instruct. Dark blue marks those heads that are the best across all four real datasets (see Figure 9).

851 For all three models (LLaMA-3-8B, LLaMA-3.1-8B and Qwen-2.5-1.5B), shown in the Figure 20,
 852 we see that the head with the highest unsupervised score, calculated on unlabelled examples from real

853 datasets, also exhibits consistently high accuracy across all these datasets, appearing in the top-right
854 corner of Figures 9 or 10. Moreover, many other top-performing heads of these models also rank
855 among those with the highest unsupervised scores on real and/or synthetic datasets.

856 L Selection Bias

857 We investigate our methods in relation to the tendency to favor specific answer choices over the
858 correct ones. Specifically, we compute how frequently the model itself or the QK -score across the
859 top three attention heads selects each option. Figure 22 illustrates the selection bias (expressed as a
860 percentage) for the MMLU task under both 0-shot and 5-shot settings.

861 Our observations indicate that the models exhibit a distinct selection bias, and the QK -score sometimes
862 reflects a similar distribution. However, the top heads often display differing distributions over the
863 answer choices. Notably, in the five-shot setting, these distributions tend to align more closely with
864 the distribution of correct answers.

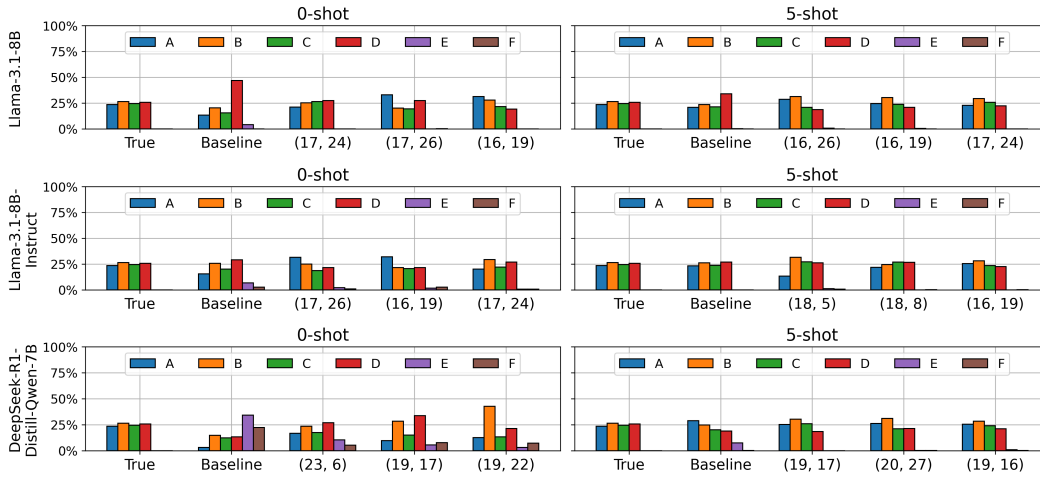


Figure 22: Distribution of predictions across options for different methods on MMLU 0-shot and 5-shot setup. (l, h) depicts the distribution for $S_{QK}^{(l, h)}$

865 M Comprehensive results for experiments on larger models from llama 866 family

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

- 870 • Figure 23 contains results for LLaMA2-13B, and Figure 30 for its chat-tuned version
- 871 • Figure 24 contains results for LLaMA2-70B, and Figure 31 for its chat-tuned version
- 872 • Figure 28 contains results for LLaMA-30B, and Figure 29 contains results for LLaMA-65B.
- 873 • Figure 25 contains results for LLaMA3-8B, and Figure 33 for its instruct-tuned version
- 874 • Figure 26 contains results for LLaMA3-70B, and Figure 34 for its instruct-tuned version
- 875 • Figure ?? contains results for LLaMA3.1-8B Instruct tuned, while results for its untuned
876 version are given in the main text.
- 877 • Figure 27 contains results for LLaMA3.1-70B, and Figure 35 for its instruct-tuned version
- 878 • Figure 36 contains results for LLaMA3.3-70B-instruct

879 Our baseline accuracies are slightly lower than those in the original model reports [Touvron et al.
880 2023, Dubey et al. 2024] for two methodological reasons: (i) we evaluate with six answer options

(A–F) rather than the four used previously; and (ii) we benchmark the strict no-chain-of-thought setting, a regime that received limited coverage in the original papers.

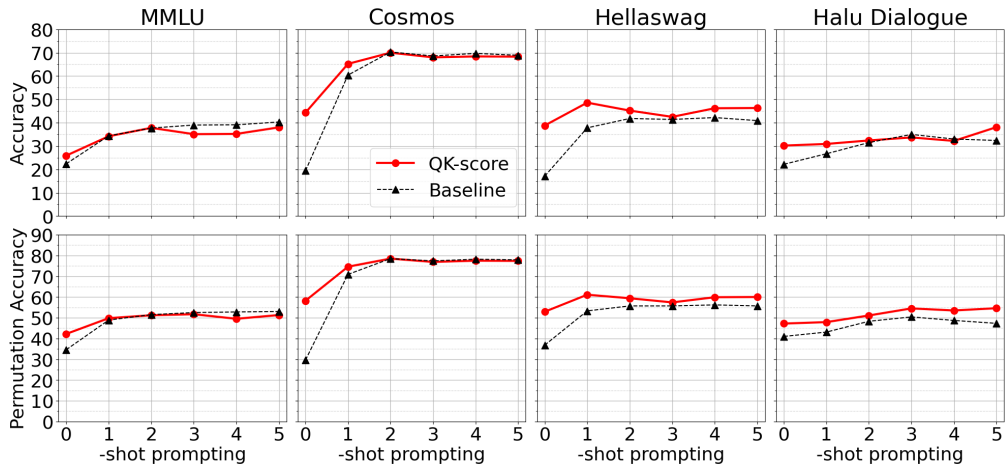


Figure 23: Comparison of different methods for LLaMA2-13B (base) on various Q&A datasets.

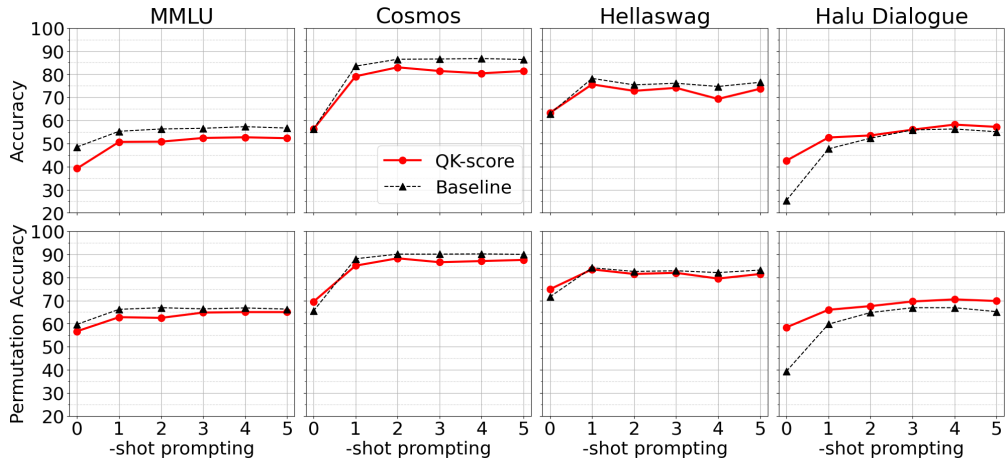


Figure 24: Comparison of different methods for LLaMA2-70B (base) on various Q&A datasets.

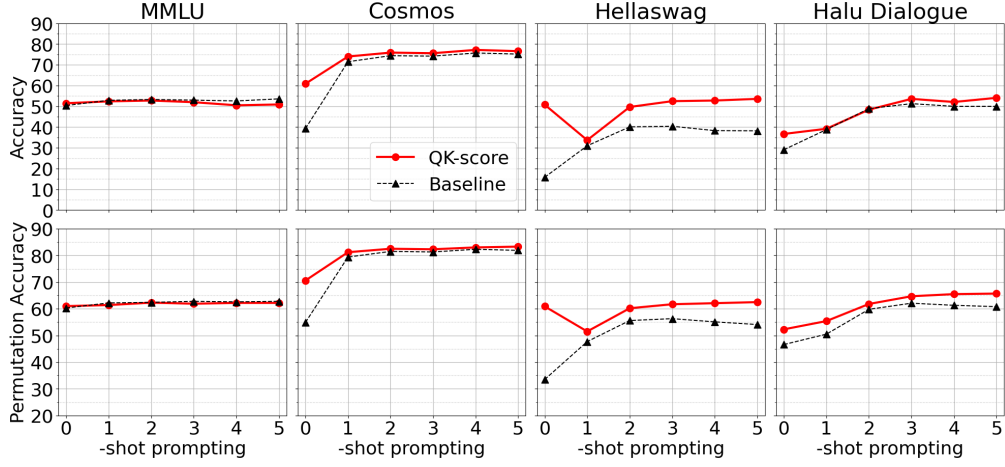


Figure 25: Comparison of different methods for LLaMA3-8B (base) on various Q&A datasets.

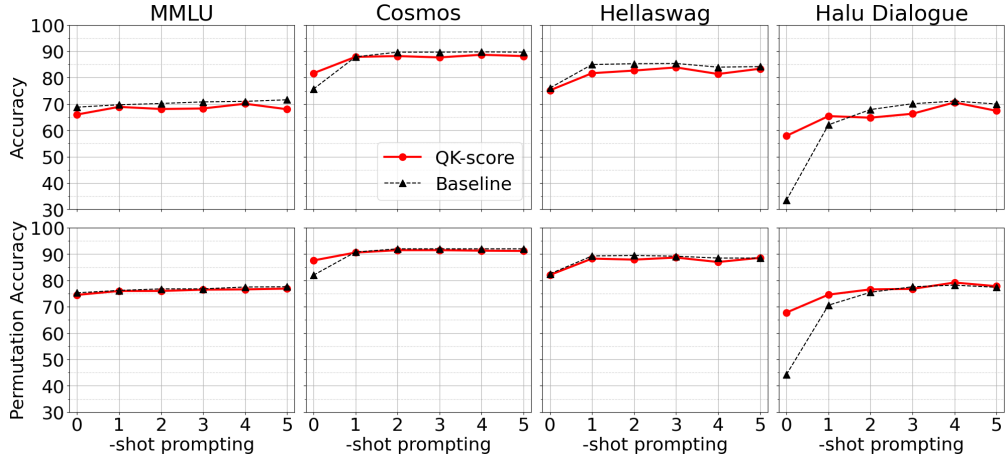


Figure 26: Comparison of different methods for LLaMA3-70B (base) on various Q&A datasets.

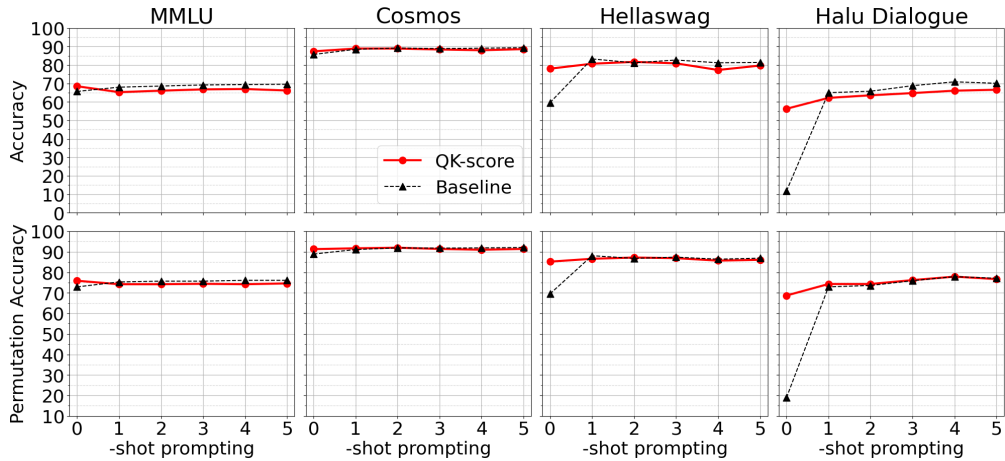


Figure 27: Comparison of different methods for LLaMA3.1-70B (base) on various Q&A datasets.

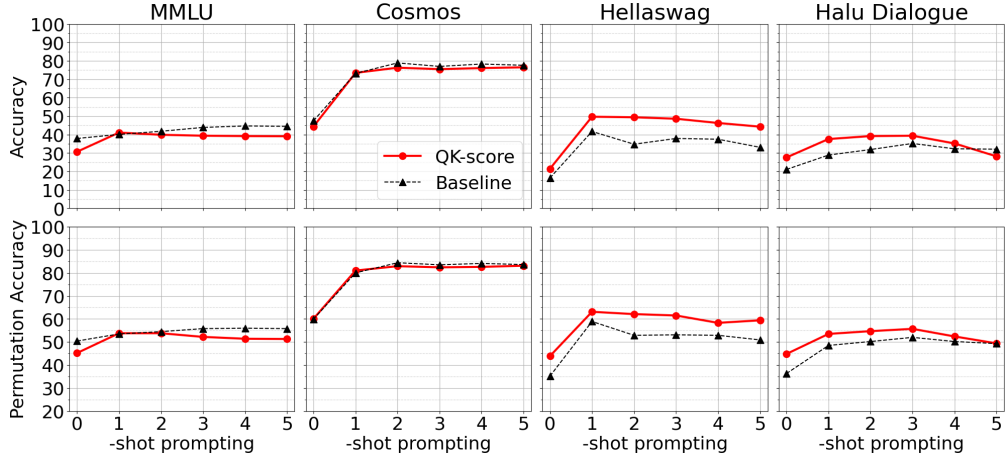


Figure 28: Comparison of different methods for LLaMA-30B (base) on various Q&A datasets.

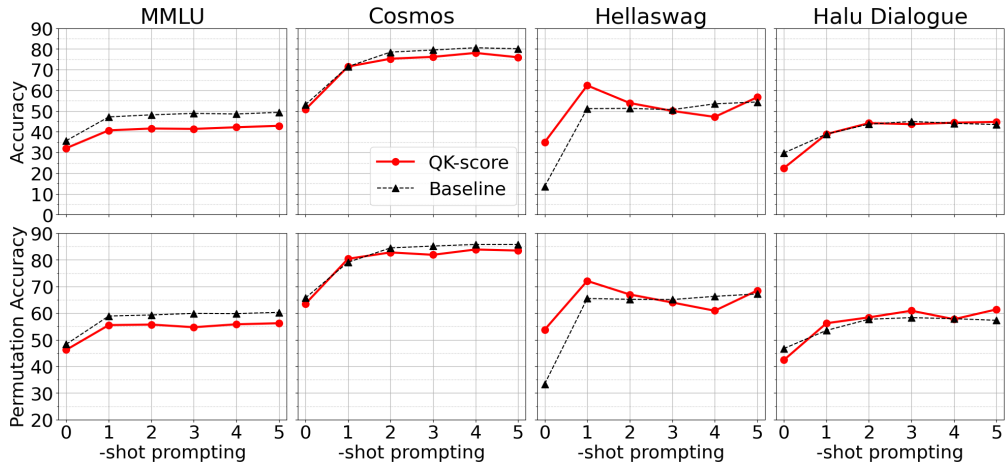


Figure 29: Comparison of different methods for LLaMA-65B (base) on various Q&A datasets.

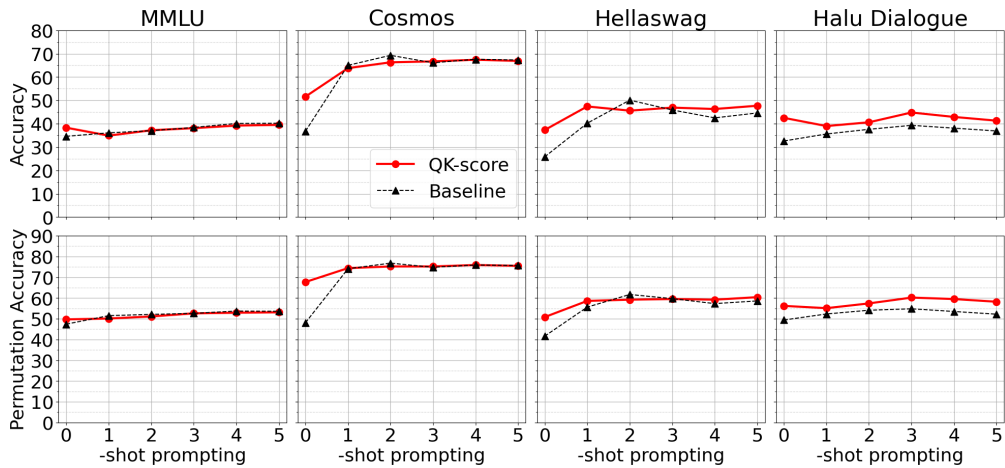


Figure 30: Comparison of different methods for LLaMA2-13B-chat on various Q&A datasets.

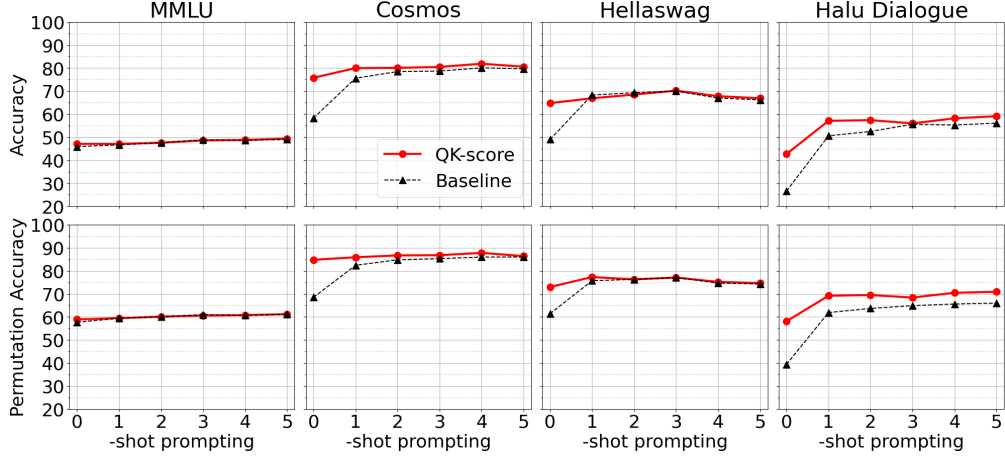


Figure 31: Comparison of different methods for LLaMA2-70B-chat on various Q&A datasets.

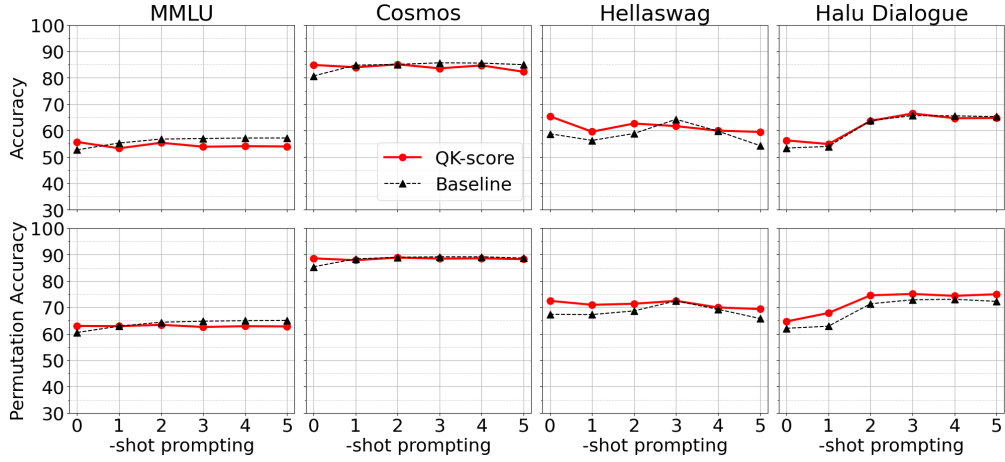


Figure 32: Comparison of different methods for LLaMA3-8B-instruct on various Q&A datasets.

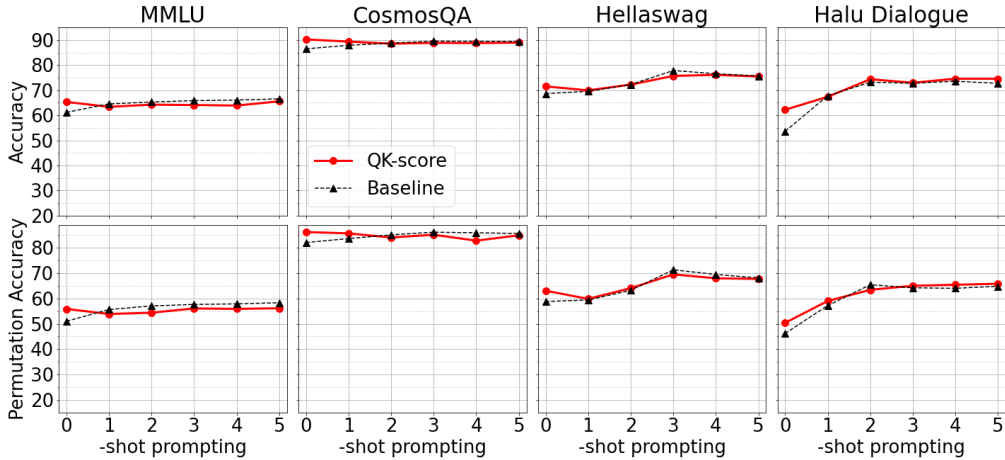


Figure 33: Comparison of different methods for LLaMA3.1-8B-instruct on various Q&A datasets.

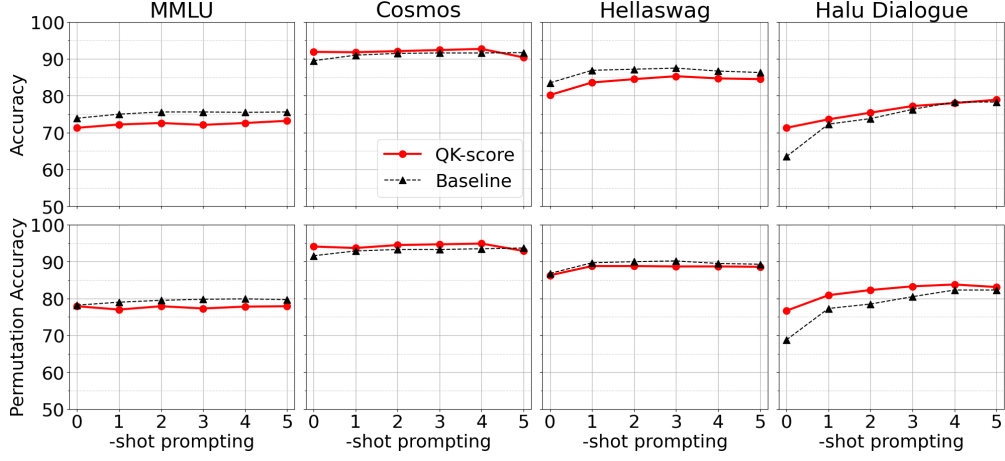


Figure 34: Comparison of different methods for LLaMA3-70B-instruct on various Q&A datasets.

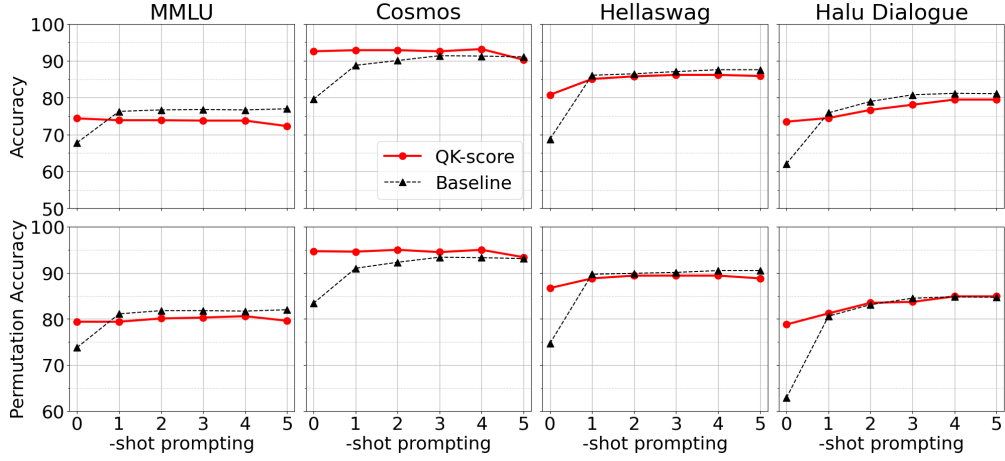


Figure 35: Comparison of different methods for LLaMA3.1-70B-instruct on various Q&A datasets.

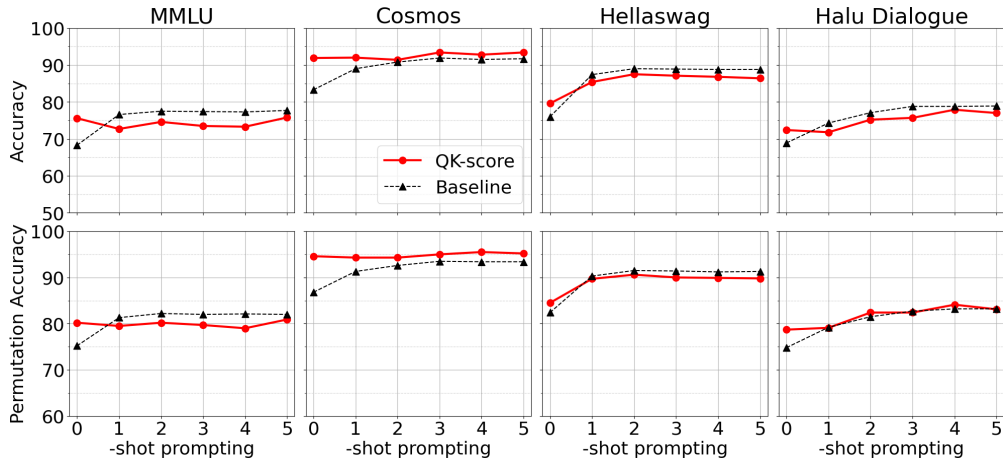


Figure 36: Comparison of different methods for LLaMA3.3-70B-instruct on various Q&A datasets.

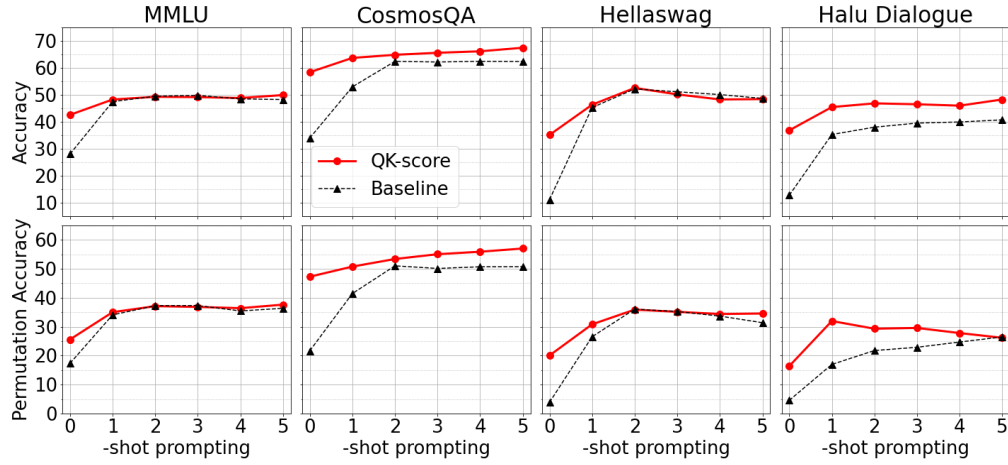


Figure 37: Comparison of different methods for DeepSeek-R1 distilled on Qwen2.5-7B on various Q&A datasets.

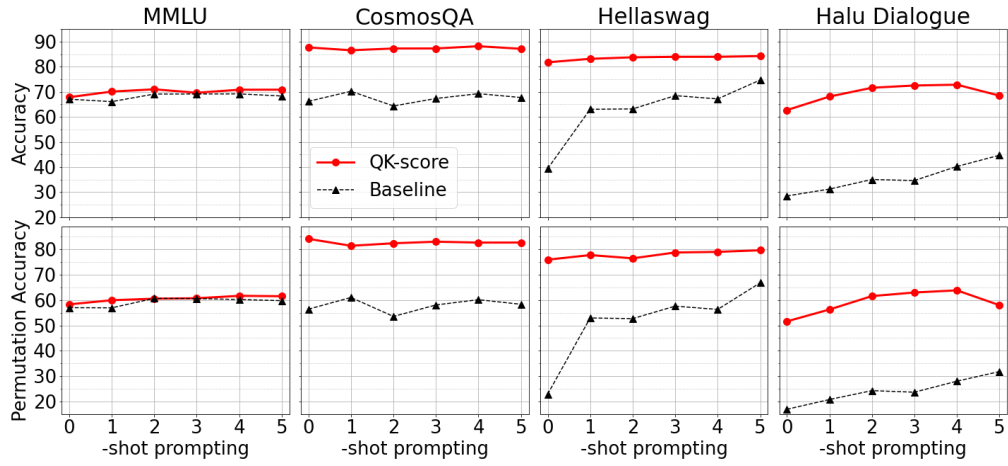


Figure 38: Comparison of different methods for DeepSeek-R1 distilled on Qwen2.5-14B on various Q&A datasets.

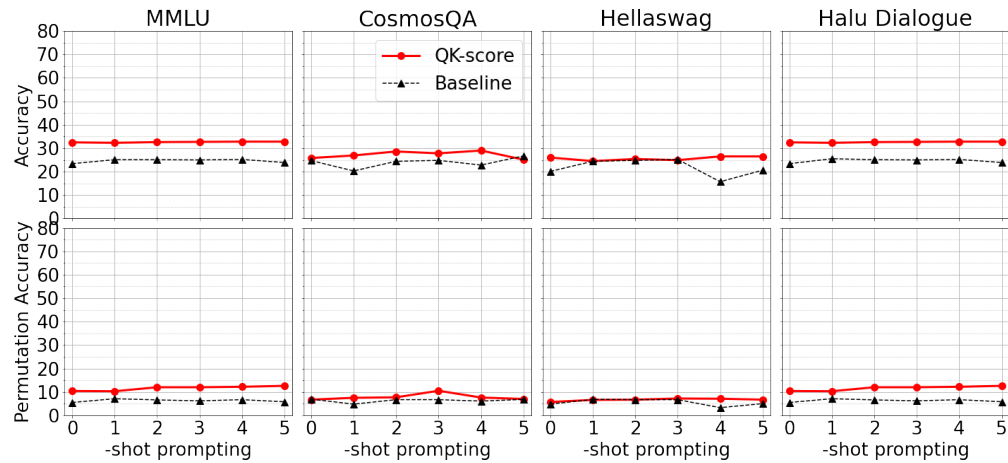


Figure 39: Comparison of different methods for Dolly V2-3B on various Q&A datasets.

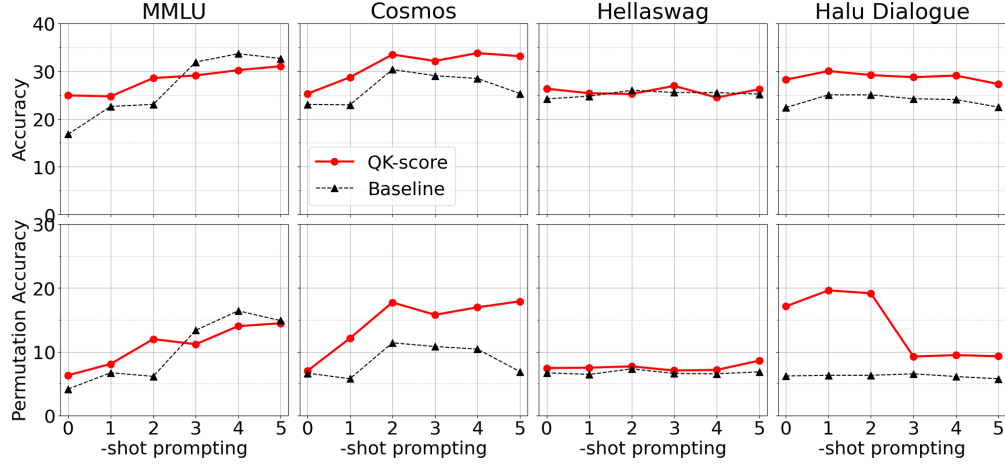


Figure 40: Comparison of different methods for Gemma-2B on various Q&A datasets.

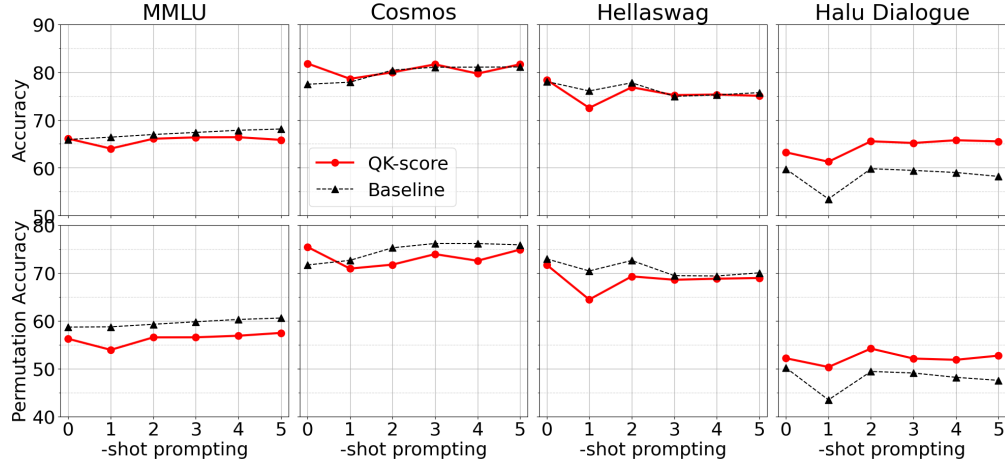


Figure 41: Comparison of different methods for Phi-3.5-mini (instruct tuned) on various Q&A datasets.

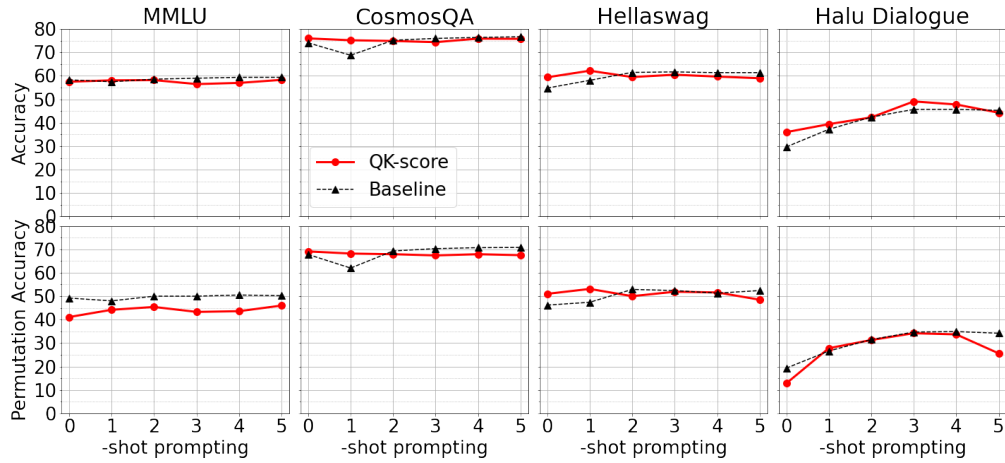


Figure 42: Comparison of different methods for Qwen-1.5B on various Q&A datasets.

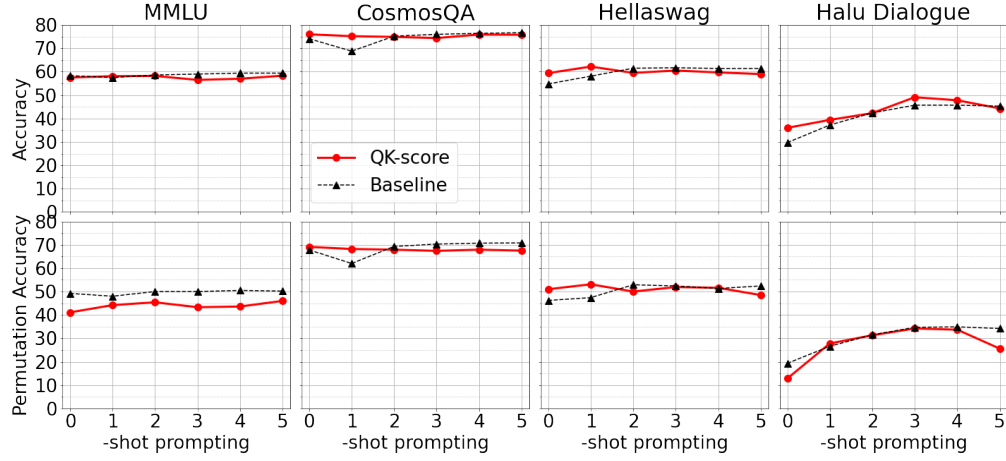


Figure 43: Comparison of different methods for Qwen-1.5B-Instruct on various Q&A datasets.

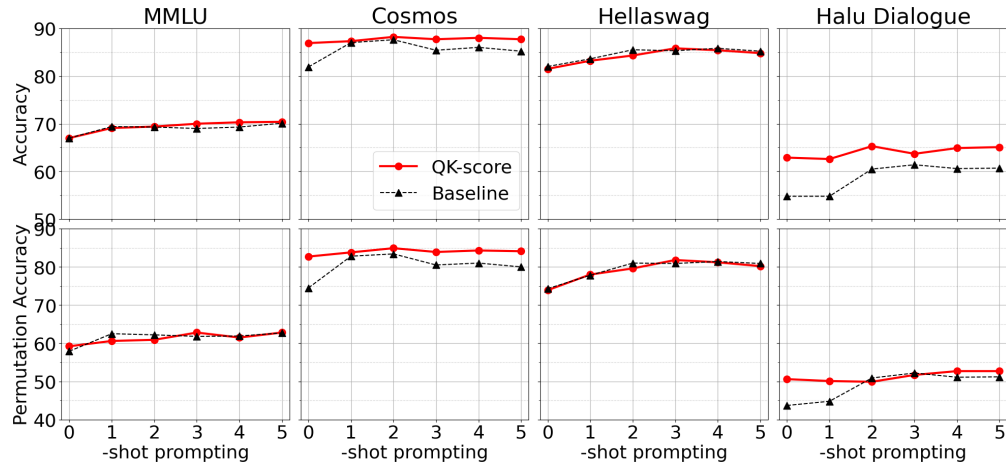


Figure 44: Comparison of different methods for Qwen-7B on various Q&A datasets.

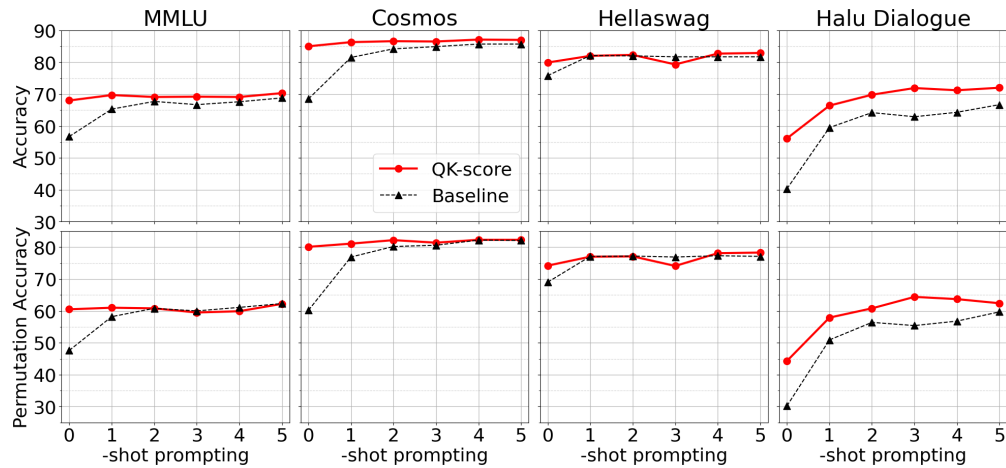


Figure 45: Comparison of different methods for Qwen-7B-Instruct on various Q&A datasets.

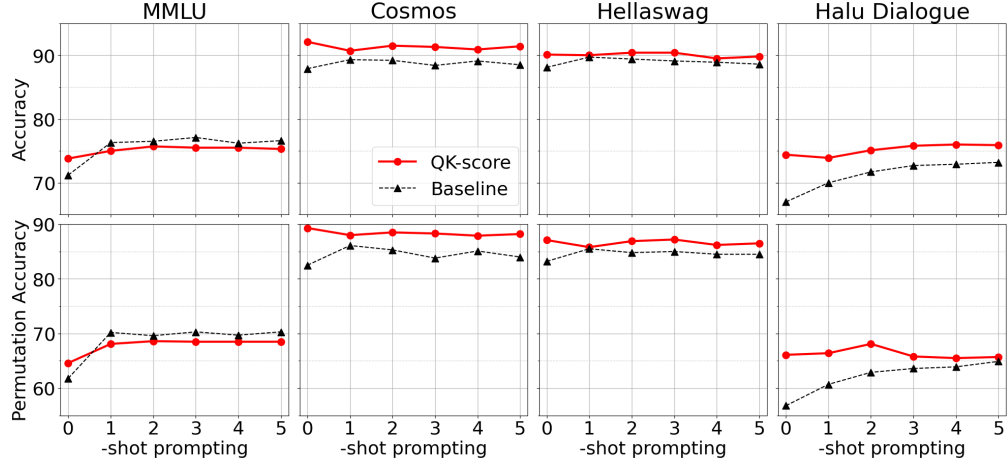


Figure 46: Comparison of different methods for Qwen-14B on various Q&A datasets.

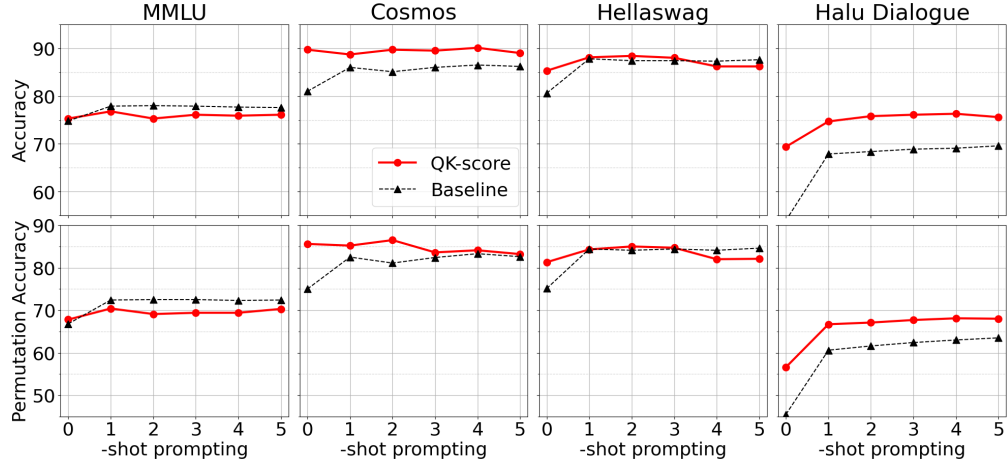


Figure 47: Comparison of different methods for Qwen-14B-Instruct on various Q&A datasets.

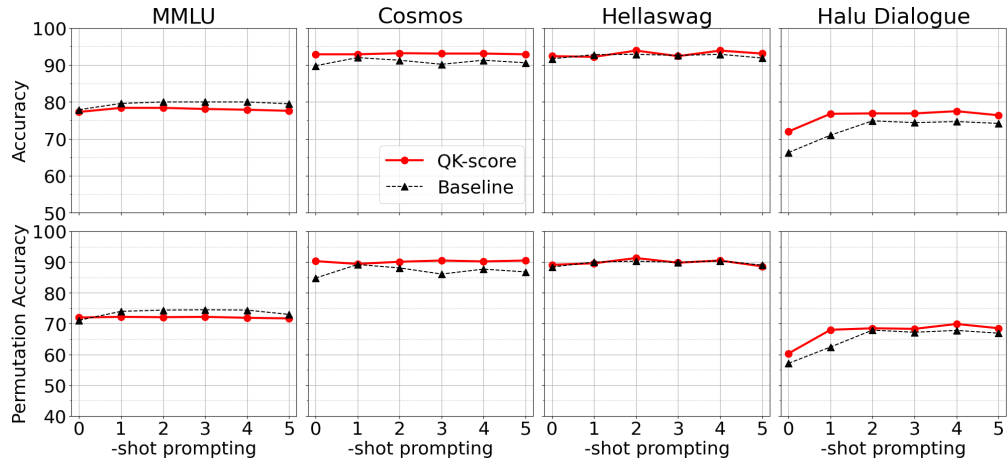


Figure 48: Comparison of different methods for Qwen-32B on various Q&A datasets.

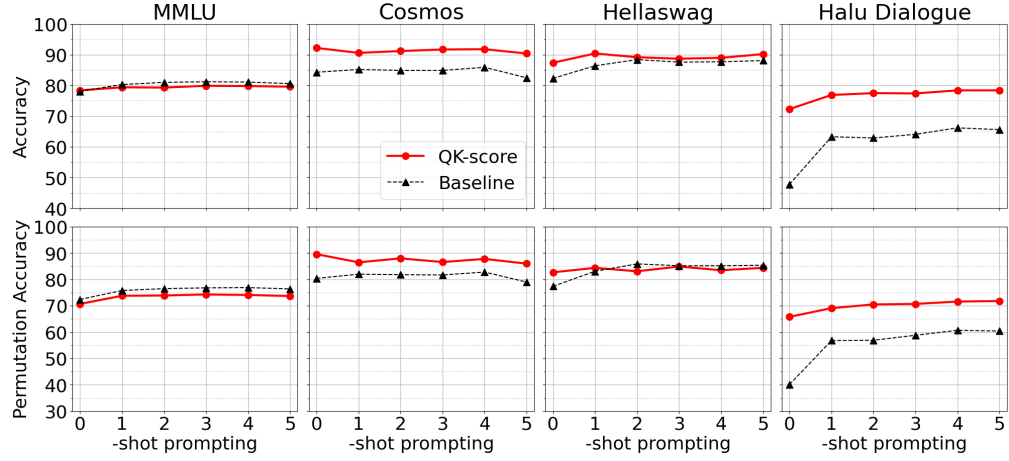


Figure 49: Comparison of different methods for Qwen-32B-Instruct on various Q&A datasets.

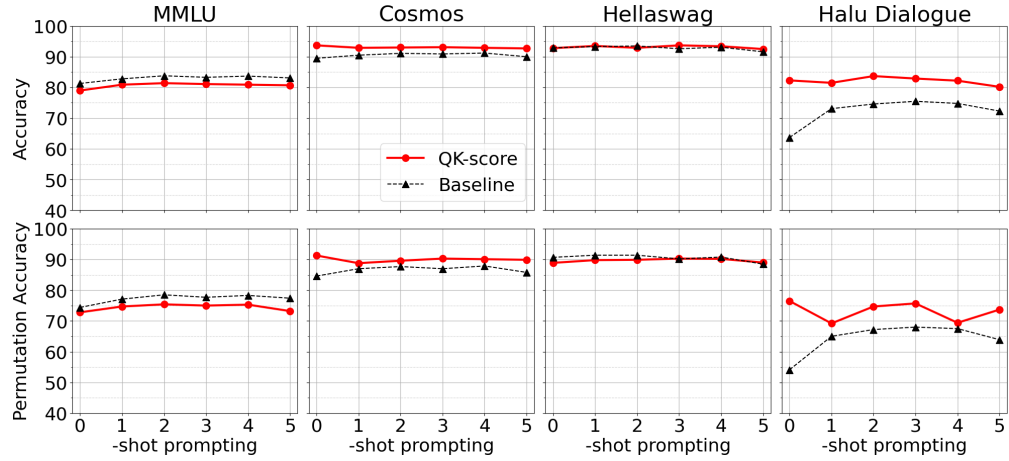


Figure 50: Comparison of different methods for Qwen-72B on various Q&A datasets.

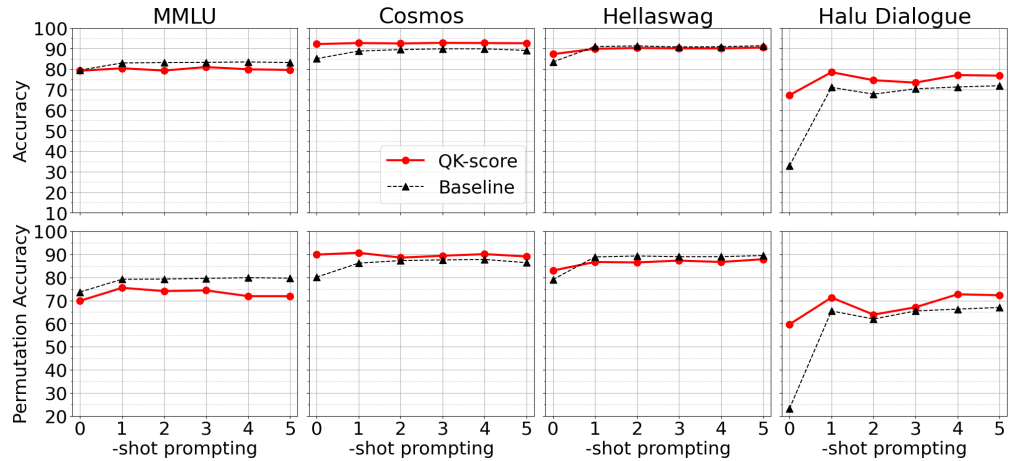


Figure 51: Comparison of different methods for Qwen-72B-Instruct on various Q&A datasets.

N Comprehensive results for experiments on Qwen-2.5 and other model families

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

- Figures 37 and 38 contain results for DeepSeek-R1 distilled on Qwen2.5-7B and -14B.
- Figure 39 contains results for Dolly V2-3B
- Figure 40 contains results for Gemma-2B
- Figure 41 contains results for Phi-3.5-Instruct
- Figure 42 contains results for Qwen-2.5-1.5B, and Figure 43 for its instruct-tuned version
- Figure 44 contains results for Qwen-2.5-7B, and Figure 45 for its instruct-tuned version
- Figure 46 contains results for Qwen-2.5-14B, and Figure 47 for its instruct-tuned version.
- Figure 48 contains results for Qwen-2.5-32B, and Figure 49 for its instruct-tuned version.
- Figure 50 contains results for Qwen-2.5-72B, and Figure 51 for its instruct-tuned version.

We observe that the accuracy and permutation accuracy plots for the baseline and QK-scores of models around 3B in size — specifically, Dolly-v2-3B (Figure 39) and Phi-3.5-mini (Figure 41) — mostly show behavior similar to LLaMA-2 models ranging from 7B (Figure 3) to 13B (Figure 23). For these models, the QK-score is usually higher than the baseline score in zero-shot setups, and the QK-score and baseline often show convergence in few-shot setups, though at times they remain at a similar distance from each other. A similar trend is observed for the Qwen 2.5 models, with sizes of 7B and 14B, especially for instruct versions. The QK-score is also typically better than the baseline for Gemma-2B (Figure 40), although the plots for this model exhibit some unusual patterns in certain cases.

However, we observe a deviation from the general trend for both the base and instruct versions of the smallest model, Qwen 2.5-1.5B. Specifically, the QK-score and baseline scores are unusually close to each other in few-shot setups, and in several cases, the QK-score performs worse than the baseline in zero-shot setups. We hypothesize that this may be due to these models being overly fine-tuned for multiple-choice question answering (MCQA), which alters the baseline behavior in zero-shot setups. This hypothesis is supported by the fact that these models outperform LLaMA2-7B and other larger models in baseline setups.

Intrigued by these differences, we conducted an additional analysis of the behaviour of individual heads in Qwen 2.5-1.5B. We confirmed that Qwen 2.5-1.5B-base contains several heads that consistently perform well across both real and synthetic datasets, such as heads (20, 4) and (21, 11). In this regard, it remains similar to the LLaMA models. For a detailed discussion of individual head performance on the SSD dataset, see Appendix J.

O Results for QK-score on fine-tuned QA models

We have finetuned the LLaMA-2-7B on each dataset and tested how our method performs after fine-tuning. We trained LoRA adapters and merged them with the model. The results are in the Table 11. To test the models, we used the subset of data the model did not see during the train.

P Results for cloze prompting

Cloze-style evaluation (or cloze prompting) (Robinson and Wingate, 2023) has been widely used for evaluating language models, but it has certain drawbacks. These include the "probability stealing" effect, where the correct answer's probability is spread across different surface forms (Wiegreffe et al., 2023), (Alzahrani et al., 2024). Cloze prompting is also sensitive to prompt phrasing and may lead to overfitting to training patterns. Although multi-choice prompting (MCP) addresses some of these issues, it introduces its own biases, such as position and label biases, and is sensitive to

		MMLU		CosmosQA		HaluDialogue		HellaSwag	
		SFT	LLaMA	SFT	LLaMA	SFT	LLaMA	SFT	LLaMA
0-shot	Baseline	0.493	0.267	0.863	0.311	0.923	0.211	0.352	0.265
	QK	0.488	0.336	0.840	0.414	0.905	0.371	0.393	0.330
1-shot	Baseline	0.476	0.391	0.836	0.393	0.474	0.309	0.716	0.288
	QK	0.472	0.407	0.810	0.5	0.858	0.366	0.800	0.371
2-shot	Baseline	0.477	0.431	0.643	0.591	0.823	0.342	0.795	0.306
	QK	0.477	0.421	0.685	0.615	0.872	0.406	0.803	0.404
3-shot	Baseline	0.483	0.437	0.670	0.569	0.621	0.361	0.681	0.33
	QK	0.470	0.405	0.692	0.593	0.867	0.423	0.786	0.427
4-shot	Baseline	0.478	0.441	0.820	0.579	0.730	0.345	0.726	0.361
	QK	0.470	0.420	0.802	0.615	0.877	0.453	0.789	0.457
5-shot	Baseline	0.487	0.438	0.827	0.547	0.746	0.356	0.678	0.346
	QK	0.482	0.427	0.810	0.61	0.880	0.428	0.786	0.423

Table 11: Accuracy on supervised fine-tuned LLaMA2-7B on the same dataset the model was fine-tuned

sample order in few-shot settings. Additionally, smaller models often struggle with the required output format [Alzahrani et al., 2024], [Khatun and Brown, 2024].

As large language models (LLMs) have advanced, the format of Question Answering tasks has shifted from cloze prompting to multiple-choice formulations [Gu et al., 2024], [OpenAI, 2024], which aligns with the focus of this study. By addressing the limitations of both cloze and MCQA prompting, our method aims to provide a more reliable and insightful model evaluation.

Our method addresses several of the issues mentioned above by separating option selection from text generation within the language model. Compared to cloze and multi-choice prompting, our approach is less sensitive to answer format and wording. It also reduces common biases in MCP, such as those related to option position or the label, by disregarding the most biased attention heads. Due to the differing nature of biases, our method and cloze prompting offer complementary insights. We plan to explore how these two methods can be combined in future work.

A comparison of cloze prompting and our method is presented in Table I2. We observe that QK-score performs similarly to cloze prompting on the MMLU dataset, which requires short, knowledge-based answers, with only slight degradation in the zero-shot setting. The same holds for CosmosQA, where the model is expected to answer questions based on context. However, for datasets that assess the model’s ability to continue given text snippets, QK-score significantly underperforms. This finding aligns with our understanding of QK-score as a method that separates semantic decision-making from text generation. On datasets like HellaSwag and HaluDialogue, the correct answer is often determined by the consistency of the text snippet rather than by factual accuracy or commonsense reasoning. We believe that the QK-score is more heavily influenced by the latter.

In summary, our approach achieves comparable performance while offering complementary insights into model behavior, making it a valuable alternative to traditional cloze prompting.

	MMLU		CosmosQA		HaluDialogue		HellaSwag	
	Cloze	QK	Cloze	QK	Cloze	QK	Cloze	QK
0-shot	0.38	0.35	0.49	0.46	0.42	0.40	0.52	0.38
1-shot	0.40	0.39	0.51	0.50	0.45	0.42	0.52	0.35
2-shot	0.39	0.40	0.48	0.51	0.46	0.42	0.53	0.38
3-shot	0.39	0.40	0.48	0.57	0.45	0.37	0.53	0.43
4-shot	0.39	0.39	0.53	0.54	0.46	0.39	0.53	0.43
5-shot	0.42	0.41	0.52	0.54	0.44	0.40	0.54	0.44

Table 12: Comparison of cloze prompting and our method with QK-Score. All experiments were ran on 4-optioned examples.

952 Q Analysis of Attention Map

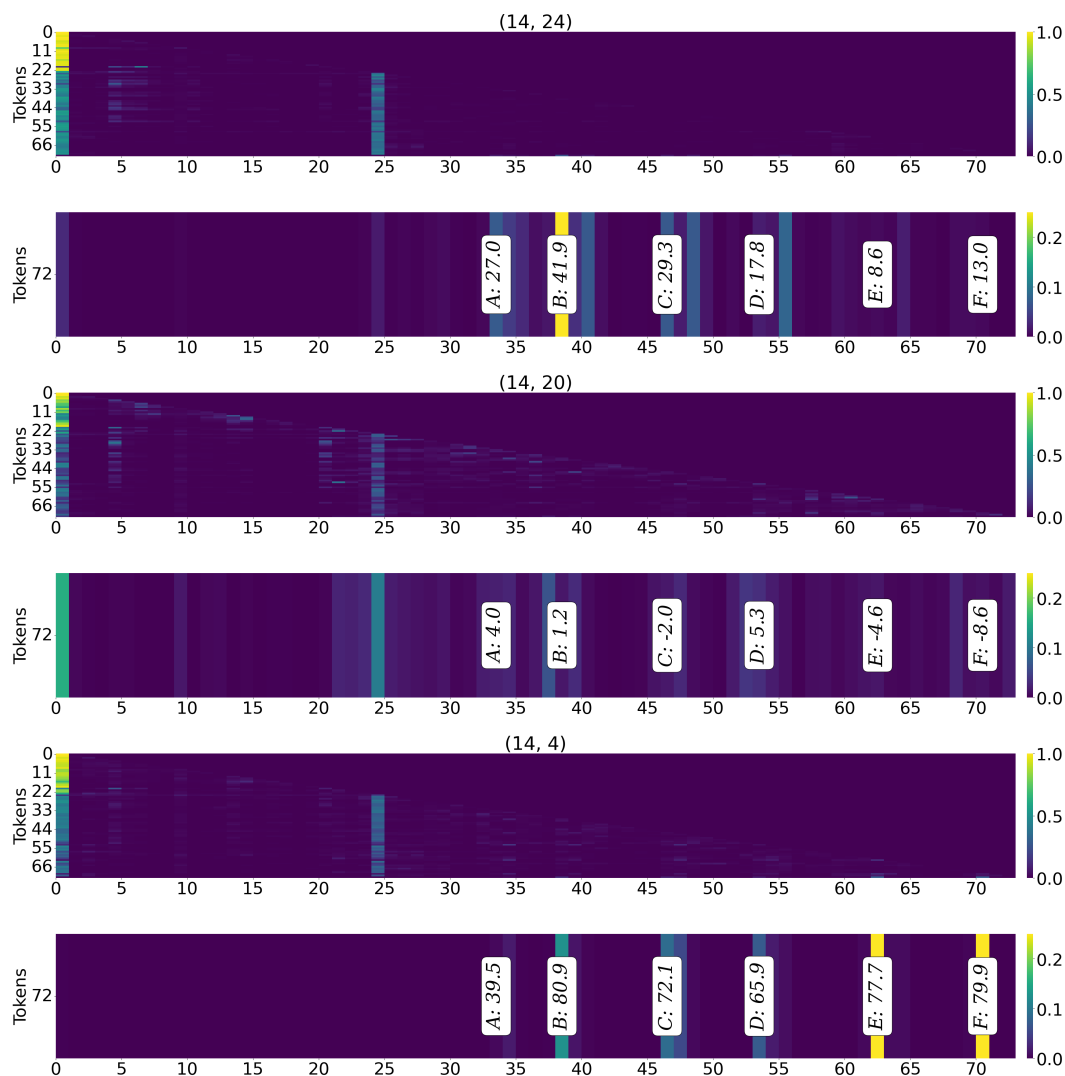


Figure 52: Attention maps of (14, 24), (14, 20) and (14,4) pairs (Head, Layer) for 0-shot setting for MMLU example: Question: What singer appeared in the 1992 baseball film 'A League of Their Own'? \nOptions: \nA. Brandy.\nB. Madonna.\nC. Garth Brooks.\nD. Whitney Houston.\nE. I don't know.\nF. None 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.

953 R Hardware and running time info

954 The most computationally intense (per sample) part of the complete setup is the calibration when
 955 QK-scores are computed for all heads. Our main experiments with LLaMA3.1-8B (and other models
 956 with size up to 8B) were performed on 16xIntel Xeon Gold 6151 CPU @ 3.00GHz with 2x Nvidia
 957 Tesla V100 GPU acceleration and 28xIntel(R) Core(TM) i7-14700K CPU with 2x NVIDIA GeForce
 958 RTX 4090 acceleration.

959 Calibration on MMLU dataset (500 samples) took between 1.9 minutes for 0-shot and 2.8 minutes
 960 for 5-shot promptings.

961 Calibration on CosmosQA dataset (500 samples) took between 4.4 minutes for 0-shot and 6 minutes
 962 for 5-shot promptings.

963 Calibration on the Hellaswag dataset (500 samples) took between 4.5 minutes for 0-shot and 6.1
 964 minutes for 5-shot promptings.

965 Calibration on Halu Dialogue dataset (500 samples) took between 4.6 for 0-shot and 6.5 for 5-shot
 966 promptings.

967 On average, evaluation of a single sample is almost 2 times faster than calibration on it.

968 Total running time of our main experiments for LLaMA3.1-8B (calibration and validation) took
 969 approximately 3.8/5.7/6.6/7.1 hours for MMLU/CosmosQA/Hellaswag/Halu Dialogue respectively.

970 Experiments with larger models ($\geq 14B$) were performed on 8xIntel Xeon Gold 6338 CPU @
 971 2.00GHz with 4x Nvidia A100 80 GB GPU accelertion. For example, for the largest LLaMA3.1-
 972 70B, full computations of our main experiments (0-, . . . 5-shot promptings with 500 samples for
 973 calibration and 9, 500 for evaluation in each) took $\approx 13/21/23/23$ hours for MMLU/CosmosQA/Hel-
 974 laswag/Halu Dialogue respectively.

975 **S Licenses of the used assets**

976 Here is the list of licenses for the assets (models and datasets) that we used in this work.

977 **Datasets**

- 978 • MMLU, Hellaswag, HaluDialogue: [MIT license](#)
- 979 • CosmosQA: [CC-BY-4.0 license](#)

980 **Models**

- 981 • DeepSeek-R1, Dolly V2, Phi-3.5: [MIT license](#)
- 982 • Gemma, Gemma 3: [Gemma Terms of Use](#)
- 983 • LLaMA: [LLAMA COMMUNITY LICENSE AGREEMENT](#)
- 984 • LLaMA2: [LLAMA 2 COMMUNITY LICENSE AGREEMENT](#)
- 985 • LLaMA3: [LLAMA 3 COMMUNITY LICENSE AGREEMENT](#)
- 986 • LLaMA3.1: [LLAMA 3.1 COMMUNITY LICENSE AGREEMENT](#)
- 987 • LLaMA3.3: [LLAMA 3.3 COMMUNITY LICENSE AGREEMENT](#)
- 988 • Qwen2.5: [Apache license 2.0](#)

T NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: Main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations of our method in the last paragraph of Section 7 (Analysis).

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: This paper does not include any theorem or other explicit theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide all the details of the experiment settings in Sections 4.5 and in the appendix.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

1094 Question: Does the paper provide open access to the data and code, with sufficient instruc-
1095 tions to faithfully reproduce the main experimental results, as described in supplemental
1096 material?

1097 Answer: [Yes]

1098 Justification: We use previously published datasets (with proper citations). For some of
1099 the experiments we created a small synthetic dataset; we provide it alongside our code in
1100 supplementary materials.

1101 Guidelines:

- 1102 • The answer NA means that paper does not include experiments requiring code.
- 1103 • Please see the NeurIPS code and data submission guidelines ([https://nips.cc/
1104 public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1105 • While we encourage the release of code and data, we understand that this might not be
1106 possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not
1107 including code, unless this is central to the contribution (e.g., for a new open-source
1108 benchmark).
- 1109 • The instructions should contain the exact command and environment needed to run to
1110 reproduce the results. See the NeurIPS code and data submission guidelines ([https:
1111 //nips.cc/public/guides/CodeSubmissionPolicy](https://nips.cc/public/guides/CodeSubmissionPolicy)) for more details.
- 1112 • The authors should provide instructions on data access and preparation, including how
1113 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- 1114 • The authors should provide scripts to reproduce all experimental results for the new
1115 proposed method and baselines. If only a subset of experiments are reproducible, they
1116 should state which ones are omitted from the script and why.
- 1117 • At submission time, to preserve anonymity, the authors should release anonymized
1118 versions (if applicable).
- 1119 • Providing as much information as possible in supplemental material (appended to the
1120 paper) is recommended, but including URLs to data and code is permitted.

1121 6. Experimental setting/details

1122 Question: Does the paper specify all the training and test details (e.g., data splits, hyper-
1123 parameters, how they were chosen, type of optimizer, etc.) necessary to understand the
1124 results?

1125 Answer: [Yes]

1126 Justification: This details are specified in Section 5.3 (‘Experimental Setup’), as well as in
1127 the Appendices (for corresponding experiments).

1128 Guidelines:

- 1129 • The answer NA means that the paper does not include experiments.
- 1130 • The experimental setting should be presented in the core of the paper to a level of detail
1131 that is necessary to appreciate the results and make sense of them.
- 1132 • The full details can be provided either with the code, in appendix, or as supplemental
1133 material.

1134 7. Experiment statistical significance

1135 Question: Does the paper report error bars suitably and correctly defined or other appropriate
1136 information about the statistical significance of the experiments?

1137 Answer: [No]

1138 Justification: Our main experiments are performed in deterministic setup; information
1139 on train/test split is provided. Results of the experiments that are non-deterministic (e.g.,
1140 selection of random heads for ablation studies) are provided with statistic bars.

1141 Guidelines:

- 1142 • The answer NA means that the paper does not include experiments.
- 1143 • The authors should answer “Yes” if the results are accompanied by error bars, confi-
1144 dence intervals, or statistical significance tests, at least for the experiments that support
1145 the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We provide this information in Appendix [R](#)

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: Research conducted in this paper, in every aspect, conforms with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our work is a foundational research on certain aspects of internal mechanisms within LLMs. There is no direct societal impact of our work that we can clearly identify.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: With this paper we do not release data or models that can be misused in a harmful way.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite all used datasets and models in References. We list licenses for all the assets that we used in Appendix [S](#).

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: With this paper, we release the synthetic QA dataset properly described in Section 5.1 and Appendix I.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- 1300 • Depending on the country in which research is conducted, IRB approval (or equivalent)
1301 may be required for any human subjects research. If you obtained IRB approval, you
1302 should clearly state this in the paper.
- 1303 • We recognize that the procedures for this may vary significantly between institutions
1304 and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the
1305 guidelines for their institution.
- 1306 • For initial submissions, do not include any information that would break anonymity (if
1307 applicable), such as the institution conducting the review.

1308 16. **Declaration of LLM usage**

1309 Question: Does the paper describe the usage of LLMs if it is an important, original, or
1310 non-standard component of the core methods in this research? Note that if the LLM is used
1311 only for writing, editing, or formatting purposes and does not impact the core methodology,
1312 scientific rigorousness, or originality of the research, declaration is not required.

1313 Answer: [NA]

1314 Justification: This paper involves LLMs only as objects of study. Its core methods were
1315 developed without any usage of LLMs.

1316 Guidelines:

- 1317 • The answer NA means that the core method development in this research does not
1318 involve LLMs as any important, original, or non-standard components.
- 1319 • Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>)
1320 for what should or should not be described.