Evaluating the Robustness of LLMs against Label Variants MCQs in Logits Space

Anonymous ACL submission

Abstract

The widespread use of Large Language Models (LLMs) has made the robustness emerge as a critical metric in LLM evaluation. Multiple-choice questions(MCQs) constitute a significant form of LLM evaluation, which underscores the importance of studying the robustness of LLMs to MCQs. While there has been considerable research on the robustness of LLMs, the majority of these studies have been conducted as black-box assessments in the textual space. In this paper, we further evaluated the robustness of LLMs to label variants in the logits space. Our experiments on 3 datasets and 10 models show that LLMs exhibit a significant selection bias towards different choice token sets, meaning that variant choice options can alter the model's confidence in answering questions. In particular, the smaller size models exhibit more pronounced selection bias. We also found that post-training can significantly enhance model robustness to label variants after comparing the base version and the instruct version of different LLMs. The results demonstrate that the evaluation in the logits space can tell us more about LLMs.

1 Introduction

Since the advent of ChatGPT, large language models (LLMs) have undergone rapid development, with their capabilities continuously improving. These LLMs have achieved remarkable results in various fields, such as text generation and mathematical reasoning. Notable LLMs include the closed-source models GPT-4 (OpenAI, 2024) and Gemini (Google, 2024), as well as the open-source models Qwen (An Yang, 2025), LLaMA (Meta, 2024), and DeepSeek (DeepSeek-AI, 2024).

The rapid advancement of LLMs has also introduced several issues, which present new challenges for LLM evaluation. For example, LLMs may experience testing data leakage (Deng et al., 2024; Balloccu et al., 2024) or specifically optimization for particular evaluation datasets, which can lead to inflated scores on the evaluation set, resulting in the invalidation of leaderboards (Zhou et al., 2023; Shi et al., 2024). To address these issues, researchers have proposed a variety of robustness evaluation methods that introduce perturbations to the datasets (Zhu et al., 2024b; Yang et al., 2024). The results indicated that even the state-of-the-art LLMs may exhibit significant performance decline on these datasets, demonstrating the poor robustness of existing LLMs (Wang et al., 2023b; Pezeshkpour and Hruschka, 2024). For instance, several studies have shown that LLMs exhibit position bias or token bias in handling MCQs (Zheng et al., 2024; Li et al., 2024b).

However, while these studies effectively evaluate model capabilities, the majority adopted a black-box approach, $i.e.$, relying solely on score differences in the textual space. In fact, an effective evaluation should not only assess the model's textual score but also facilitate model optimization, and it is widely acknowledged that evaluating in the logits space can offer more insights for the researchers, particularly for the trainers.

In this paper, we investigate the robustness of LLMs against label variants in three MCQs datasets, specifically by replacing the original choice options with different choice token sets. We conducted extensive experiments on various LLMs in the logits space and identified some intriguing conclusions. Our main contributions are as follows:

1. LLMs generally experience performance drops with label variants. An analysis in the logits space indicates that LLMs exhibit selection bias towards different token sets, whereby specific token sets can lead to reduced confidence in their responses.

2. This selection bias towards the choice token set is related to the model size, with smaller models exhibited more pronounced selection bias.

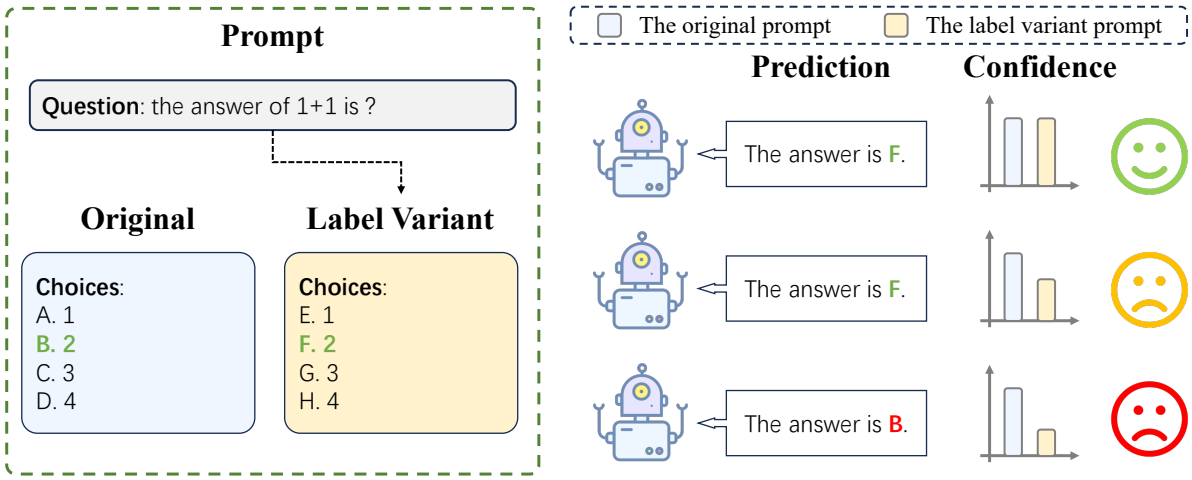

Figure 1: **The overview of our robustness evaluation for the label variant.** For example in the figure, the variant token set $\{E, F, G, H\}$ of choices instead of the original token set $\{A, B, C, D\}$. We consider the LLMs to potentially exhibit the following three scenarios: **(1).** The model can accurately answer the correct option ($i.e.$, 'F'), without a significant drop in the confidence of its choice. **(2).** The model's answer is correct but its confidence has a significant drop. **(3).** The model selects a wrong answer ($e.g.$, 'B') and has a low confidence for it at the same time.

3. LLMs with similar performance in the textual space can be quite different in the logits space, which underscores the necessity of evaluating models in the logits space.

4. By analyzing the performance of the base and instruct models, we found that post-training can enhance the robustness of the models to label variants.

2 Related Work

LLMs are undergoing rapid development, making the effective evaluation of LLMs a critical area of research (Chang et al., 2023). Numerous studies have highlighted the issue of data leakage in current LLMs, leading to unreliable rankings on LLM leaderboards (Deng et al., 2024; Zhou et al., 2023; Balloccu et al., 2024; Shi et al., 2024). This has underscored the need for evaluations that measure the robustness and true capabilities of LLMs.

Recently, there is a significant body of research focused on evaluating the robustness of LLMs. (Wang et al., 2023a) and (Chai et al., 2024) respectively investigated the impact of word-level and token-level perturbations on the performance of LLMs. The results indicate that LLMs are not robust to such perturbations. (Zhu et al., 2024a) developed a prompt benchmark to evaluate the impact of different prompts on LLMs. (Shi et al., 2023) investigated the impact of distracting information on the performance of LLMs, and the results indicate that LLMs are highly susceptible to interference from irrelevant context. (Li et al., 2024a; Hong et al., 2024) applied perturbations to mathematical and code datasets to evaluate the robustness of LLMs in mathematical and coding.

MCQs serve as a crucial assessment format for evaluating the LLMs, due to their ease of construction and assessment (Robinson and Wingate, 2023). Several studies have been conducted on the robustness of LLMs in handling MCQs. (Zheng et al., 2024; Alzahrani et al., 2024; Pezeshkpour and Hruschka, 2024) showed that LLMs commonly exhibit token bias and position bias in MCQs, tending to select specific tokens or option in specific position (such as "A" or the option in the first position). (Li et al., 2024b) found that the positional preferences in LLMs remain consistent across datasets, demonstrating through a self-constructed dataset. (Balepur et al., 2024) found that LLMs possess the capability to answer based solely on the options provided, and suggested that this approach is a stronger baseline than a majority baseline for MCQs.

While substantial research focuses on LLM robustness in MCQs, most studies employ a black-box approach, analyzing model behaviour within the token space, which offers limited insights. Inspired by existing methods, our study systematically analyzes the robustness of LLMs against different choice tokens and interprets the anomalous behaviors of models from the logits space.

3 Methodology

3.1 Preliminaries

The Robustness Benchmark. Constructing the robustness benchmark is one of the key steps for the robustness evaluation. Let $\mathcal{D} = \{(\boldsymbol{p}_i, g_i)\}_{i=1}^{N}$ denotes the original benchmark ($i.e.$, testing set), where $\boldsymbol{p}_i \in \mathcal{P}$ is i-th prompt, and g_i is its ground-truth label. To assess the robustness of model $\mathcal{M}$, the evaluators will generate the robustness benchmark $\hat{\mathcal{D}} = \{(\hat{\boldsymbol{p}}, \hat{g}) | \hat{\boldsymbol{p}} = F_P(\boldsymbol{p}), \hat{g} = F_G(g), (\boldsymbol{p}, g) \in \mathcal{D}\}\}$ with the textual perturbations, where F_P and F_G respectively are the specific functions to perturb the prompts and the labels.

Evaluator's Goal. A common goal for the evaluators is to examine the impact of prompt perturbations on model performance. Recent research(Chang et al., 2024) have introduced lots of evaluation methods from different perspectives in the textual space. In this paper, we further expand the evaluation space to the logits space. Specifically, we focus on the label variant method on the MCQ benchmarks following the existing works (Alzahrani et al., 2024).

3.2 Label Variant

Label variant is one of the useful methods to assess the robustness on the MCQ benchmarks. Specifically, this method perturbs the prompt by modifying the choice token set $\mathcal{T}_c$ the choice options ($i.e.$, the choice options in the question) and changes its ground-truth label to the corresponding variant version at the same time. For example, existing work(Alzahrani et al., 2024; Meta, 2024) modified the original choice token set $\{A, B, C, D\}$ to the ordered numerical token set $\{1, 2, 3, 4\}$, and the common language independent token set $\{\$, \&, \#, @\}$ whose order is not implicitly relative. To further examine the models' robustness for the label variant, we additionally constructed the benchmark with $\mathcal{T}_c = \{E, F, G, H\}$, as we consider this token set has a closer distance to the original token set. Our intuition is that the model will exhibit different performance of the robustness on the different similarities between the variant token set and the original token set. The example for the label variant prompt can be seen in Figure 2.

Prompt 3.1: original choice token set *{A,B,C,D}*

Question: q
Choices:\nA. c_1 \nB. c_2 \nC. c_3 \nD. c_4
Answer: B

Prompt 3.2: variant choice token set *{E,F,G,H}*

Question: q
Choices:\nE. c_1 \nF. c_2 \nG. c_3 \nH. c_4
Answer: F

Prompt 3.3: variant choice token set *{1,2,3,4}*

Question: q
Choices:\n1. c_1 \n2. c_2 \n3. c_3 \n4. c_4
Answer: 2

Prompt 3.4: variant choice token set *{$, &, #, @}*

Question: q
Choices:\n$. c_1 \n&. c_2 \n#. c_3 \n@. c_4
Answer: &

Figure 2: **The example for the original prompt and the variant prompt**.

3.3 Logits Space Evaluation

We define the confidence of the token in the predicted text as the highest value in its corresponding logit ($i.e.$, the probabilities vector that the model predicted). Considering the predicted text always contains the tokens that are not the final answer ($e.g.$, CoT), similar to existing textual evaluation methods(Chang et al., 2024), we employ the post-processing function $f(\boldsymbol{t})$ to extract the final answer option token, where $\boldsymbol{t}$ is the predicted text. Please note there is not a one-to-one correspondence between the option character and its token, in fact, we will map the answer option to its corresponding token after the post-processing.

For label variant method, one of the limitations of the textual evaluation method is that once inference can't accurately reflect the robustness of the model. In other words, in spite of the model can output the correct answer option in the inference stage, the confidence of the answer option may have a significant drop. As mentioned in Section 3.1, to achieve an accurate assessment of with robustness with lower costs, we further utilized the logit of the answer option token (dubbed 'answer logit') to assess the robustness of the models from the two following perspective.

the Confidence of the Answer option (AC). The model's confidence in the answer token can reflect its robustness. We defined AC as the maximum value in the answer logit vector. The AC can be formulated as follows:

$$AC = C(y) = \max(\mathbf{v}_{ans}) \tag{1}$$

where $C(y)$ denotes the confidence of the answer token y, and $\mathbf{v}_{ans}$ is the answer logit vector. For example, by comparing the original dataset's AC to the variant dataset, we can evaluate the confidence changes brought by the label variant.

the Inner Confidence Distance (ICD). We futher introduce the Inner Confidence Distance (ICD), which denotes the distance between the variant token and the original token in the answer logit. The ICD can be formulated as follows:

$$ICD = \begin{cases} C(F_G(y)) - C(y), & \text{if } y \in \mathcal{T}_c^{org}, \\ C(y) - C(F_G^{-1}(y)), & \text{if } y \in \mathcal{T}_c^{var}. \end{cases}$$
$$y = f(\boldsymbol{t}) \tag{2}$$

where $C(x)$ denotes the confidence of the token x in the answer logit and F_G^{-1} is the inverse process of F_G. $\mathcal{T}_c^{org}$ is the original token set $\{A, B, C, D\}$, and $\mathcal{T}_c^{var}$ is the variant version ($e.g.$, $\{E, F, G, H\}$). If the answer option belong to the variant token set $\mathcal{T}_c^{var}$, the ICD will be positive; otherwise, it will be negative. The robust model should confidently provide a variant option as an answer. In other words, the higher the ICD, the robust the model.

4 Experimental Setup

In this section, we describe the setting of our experiments. We provide detailed information on the benchmarks used for evaluation, the LLMs evaluated, and the evaluation methods employed in our experiments.

4.1 Benchmarks

To investigate the impact of label variants, we selected the current mainstream benchmarks, all of which are in the form of multiple-choice questions. To eliminate the impact of language on the results, we included datasets in both English and Chinese. The three datasets we ultimately selected are as follows:

MMLU Massive Multitask Language Understanding (MMLU) is a comprehensive English knowledge benchmark that covers a wide range of domains, with questions in multiple-choice format (Hendrycks et al., 2021).

HellaSwag Hellaswag is a dataset for commonsense natural language inference, designed to challenge machines to identify the most likely continuation of an event description (Zellers et al., 2019).

C-Eval C-Eval is a comprehensive Chinese knowledge dataset that encompasses a wide range of categories and varying levels of difficulty. The questions in the dataset are presented in a multiple-choice format (Huang et al., 2023).

4.2 Evaluation Metric.

Following the existing works(Li et al., 2024b), we first utilized the textual score to evaluate the performance of the model. We adopt the Ratio of the model's Answer to the Original options (AOR) to further evaluate the predicted text. Moreover, as we mentioned in Section3.3, we also utilized the AC and the ICD to evaluate the robustness of the models. Notably, in general, we adopted their average values for assessment.

4.3 Considered LLMs

Since our method relies on the logit in models, we considered two series of open-source models, including Qwen and LLaMA. For the Qwen series model, We evaluated both the 7B and 72B models for versions 1.5, 2.0 and 2.5. As for LLaMA series, we assessed both the 8B and 70B models for versions 3.0 and 3.1. Furthermore, we conducted comparisons on the pre-trained (base) version and the instruction-tuned (instruct) version. More details can be found in Appendix A.1.

4.4 Performance Evaluation

We evaluate each model on MMLU's testing set with 14042 prompts and CEval's validation set which contains 1346 prompts. For each prompt on these two datasets, we use 5-shot settings. For HellaSwag, we use its testing set with 10042 propmts and select zero-shot setting for assessment. For each model, we set the temperature to 0 to eliminate the impact of randomness, and use vLLM(Kwon

Model	MMLU			CEval			Hellaswag		
	EFGH	1234	$&#@	EFGH	1234	$&#@	EFGH	1234	$&#@
Qwen1.5-7B	-2.63	-1.71	-5.63	-13.38	-3.77	-8.94	-0.65	-1.95	-12.17
Qwen2-7B	-1.77	-0.75	-4.89	-1.98	-1.49	-4.07	-1.82	-1.03	-9.41
Qwen2.5-7B	-1.72	-0.52	-3.35	-1.58	-0.24	-5.01	+0.03	+1.08	-4.02
Qwen1.5-72B	-1.28	-0.7	-1.66	-12.59	-0.24	-3.99	-1.07	-0.15	-2.97
Qwen2-72B	-1.31	-0.61	-1.52	-11.11	+0.58	-0.63	-1.14	-4.09	-2.55
Qwen2.5-72B	-0.90	-0.97	-2.29	-1.59	-0.61	-4.46	-0.62	-0.69	-1.29
LLaMA3-8B	-1.02	-1.09	-2.74	-1.85	-0.11	-2.89	+0.14	-2.29	-9.28
LLaMA3.1-8B	-1.4	-0.43	-4.78	-0.21	-0.56	-2.74	-3.46	-2.32	-12.45
LLaMA3-70B	-1.17	-0.91	-2.75	+0.49	+0.03	-2.08	-0.51	-1.32	-2.14
LLaMA3.1-70B	-0.73	-0.86	-3.55	+0.67	-0.28	-2.54	+0.97	-0.76	-0.33

Table 1: **The textual score gaps of different models on the variant choice token sets.** The red color denotes the score drop is bigger than 5.0. We found that the label variants generally lead to a score decline in the models' performance, even with similar token sets (*e.g.*, $\{E, F, G, H\}$). Different models always exhibited varying sensitivities to specific token sets across different datasets (*e.g.*, Qwen series on CEval-$\{E, F, G, H\}$ and LLaMA series on MMLU-$\{\$, \&, \#, @\}$).

et al., 2023) to accelerate inference. All the experiments were conducted on NVIDIA A100 GPUs.

5 Results & Analysis

In this section, we will present the key findings from our extensive experiments (detailed in Section 3) and provide further analysis of some intriguing observations from the logits space. More comprehensive details and results of the experiments can be found in the Appendix A.

5.1 The label variant is detrimental to the model's performance.

Following the existing work(Alzahrani et al., 2024; Meta, 2024), we first reviewed the performance of the instruction-tuned model's robustness in the textual space, and additionally evaluated the token set $\{E, F, G, H\}$. As shown in Table 1, most LLMs exhibited varying degrees of performance drops on the testing sets modified with different token sets. Moreover, we observed that the models are differently sensitive to the different label variant datasets. For example, the large size Qwen series models and LLaMA series models have more significant score drops than their small size versions on the Hellaswag with the token set $\{\$, \&, \#, @\}$. Despite the token set $\{E, F, G, H\}$ having a higher similarity with the original token set $\{A, B, C, D\}$, the score drops of some Qwen series models are still bigger than 11%.

As we mentioned in Section 3.3, to avoid the limitations of the textual score, we further examined the distribution of AC for the models. As shown in Figure 3, we found that for most cases, even those without significant score drops in the textual space, the AC distribution of the token set $\{\$, \&, \#, @\}$ tends to be more dispersed, while others are concentrated around a high probability. The complete AC distribution can be found in Appendix A.2. Based on this observation, we hypothesized that LLMs might exhibit a selection bias towards different token sets. These findings preliminarily demonstrated that the label variant methods can lead to the robustness performance drops for LLMs. More details will be discussed in the following parts.

Models	ABCD	$&#@	Gap
Qwen2-7B	0.90	0.65	**-0.25**
Qwen2-72B	0.93	0.79	-0.14
LLaMA3-8B	0.91	0.57	**-0.34**
LLaMA3-70B	0.98	0.70	-0.28

Table 2: **The AC gap between the token set $\{\$, \&, \#, @\}$ and the token set $\{A, B, C, D\}$ for different size models on the HellaSwag dataset.** The results indicate that smaller models exhibit a significantly larger AC gap compared to their larger counterparts, reflecting a more pronounced token set selection bias in the smaller models.

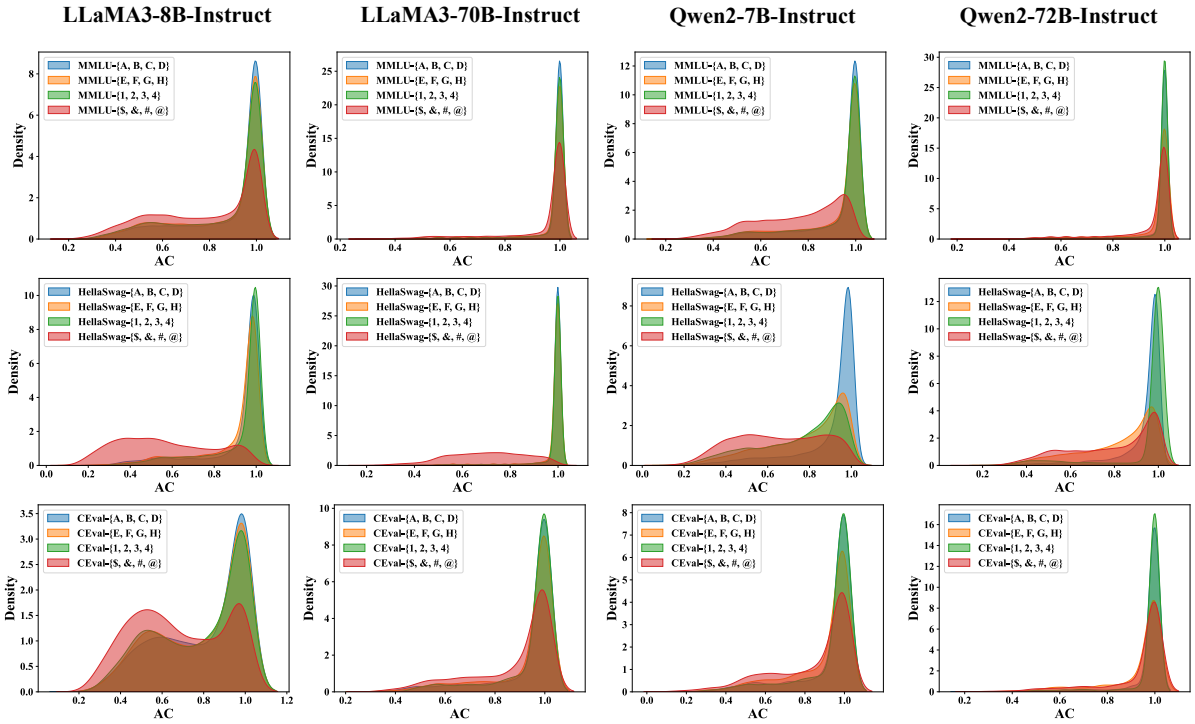

Figure 3: **The AC distributions**. We utilized Kernel Density Estimation (KDE) to illustrate the AC distributions across various instruction-tuned models and datasets. The higher density near 1.0 on the x-axis, the better the model's performance on the dataset. For example, in most cases, the original token set $\{A, B, C, D\}$ exhibited higher ACs than others.

5.2 The relationship between model size and selection bias

Based on the findings in Section 5.1, we further investigated the relationship between model size and selection bias in the logits space. We calculated the gap between the performance on the token set $\{\$, \&, \#, @\}$ and the token set $\{A, B, C, D\}$ for different size models across various datasets. As shown in Table 2, the results indicated that both Qwen and LLaMA exhibit a consistent pattern with respect to model size on the HellaSwag. Specifically, smaller models show a larger performance gap compared to their larger counterparts within the same series, suggesting that smaller models exhibit more pronounced selection bias.

The underlying causes for the more pronounced selection bias observed in smaller models remain uncertain. As can be observed from the Figure 4, the behaviour of Qwen and LLaMA is consistent, despite variations in their training data. We hypothesize that this phenomenon may be attributed to two main factors: the quantity of training data and model capacity (number of parameters), as larger models typically benefit from more extensive training data, and powerful model capacity.

Furthermore, we observed that despite large-sized models exhibiting similar scores and score gaps in the textual space, there are significant differences when viewed from the perspective of the logits space. As illustrated in Figure 5.2, we used the HellaSwag dataset as an example to compare the scores and response confidence levels of the LLaMA and Qwen large-size models. It is evident that although the performance of LLaMA and Qwen is comparable, the AC of LLaMA is notably lower than that of Qwen. As the version of the Qwen series is updated from 1.5 to 2.5, both their scores and AC progressively increase, whereas LLaMA shows a slight decline from 3 to 3.1. This analysis in the logits space inspired us more compared to the textual space. Therefore, we recommend that, in addition to observing score differences, the differences in the logits space should also be considered when evaluating or training models.

5.3 The LLMs still answered the original options

In this section, we discuss why the Qwen series models exhibited significant performance drops in the CEval dataset with the token set $\{E, F, G, H\}$.

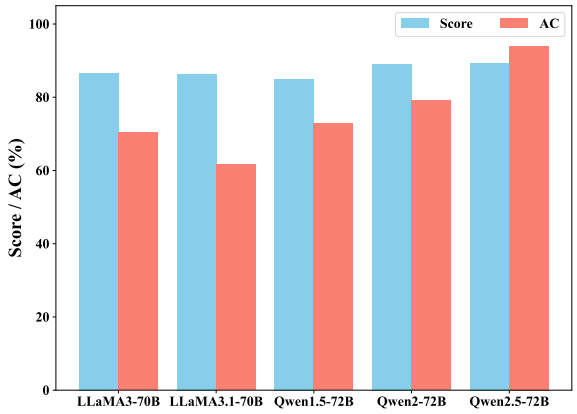

Figure 4: **The scores and the average ACs of LLaMA 70B and Qwen 72B instruct models on the Hellaswag with the token set** $\{\$, \&, \#, @\}$**.** While the score performance of LLaMA and Qwen models is comparable, there are notable differences in the average AC values. Specifically, Qwen2.5 version exhibited the highest average AC, indicating a high level of confidence in its responses.

Model	Base		Instruct	
	AOR	ICD	AOR	ICD
Qwen1.5-7B	11.89	35.75	19.91	30.41
Qwen2-7B	43.83	14.13	1.71	84.35
Qwen2.5-7B	3.79	59.68	0.97	93.12
Qwen1.5-72B	22.81	30.41	14.56	64.62
Qwen2-72B	23.03	40.18	16.64	63.12
Qwen2.5-72B	7.43	63.57	0.30	96.45
LLaMA3-8B	0.00	42.26	0.07	78.99
LLaMA3.1-8B	0.00	47.91	0.00	65.81
LLaMA3-70B	0.15	61.67	0.07	92.41
LLaMA3.1-70B	0.00	58.72	0.07	73.83

Table 3: **The AOR and ICD for different models on CEval dataset with the token set** $\{E, F, G, H\}$**.** In most cases, the instruct model exhibited the lower AOR and the higher ICD than its base version. Moreover, Qwen and LLaMA exhibit markedly different behaviors. $e.g.$ Both Qwen2.5 instruct and LLaMA 3.1 instruct have near-zero AORs, but Qwen2.5 demonstrates higher ICDs.

We conducted a detailed analysis of the model outputs and found the models continued to generate the original options ($i.e.$, the token set $\{A, B, C, D\}$) even though their real token set had been altered. An example is shown in Figure 6. Table 3 shows the proportion of cases where the model outputs the original options ($i.e.$, AOR). To further explore the reason why the models answered the original options, we examined the ICD between both original and variant in the logit space. The main findings are as follows.

ICD and AOR are negatively correlated. We compared the AOR and the average ICDs for different versions' instruct models. As shown in Table 3, with the iteration of the LLMs, in most cases, their AORs demonstrate a gradual decline and their ICDs increase at the same time. Notably, Qwen2-72B-Instruct's AOR has actually increased even in contrast to its older version. Figure 5 also illustrates that the density of the ICD around -1.0 has increased in Qwen2-72B-Instruct compared to its pre-trained version. These results indicate a certain degree of negative correlation between AOR and ICD. Moreover, we also finded the LLaMA3.1 instructed models (both 8B and 70B) have a significant drop compared to the 3.0 versions. This conclusion is consistent with the findings that we presented in Section 5.2.

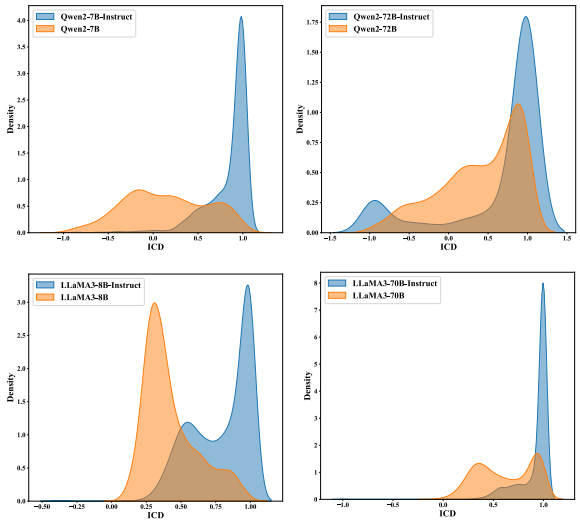

Figure 5: **The ICD distributions**. In all cases, instruct models exhibited the better performance than their pre-trained versions. $i.e.$, the density of the ICDs round 1.0 significantly increased.

Efficient post-training is benefit for the robustness. To further demonstrate the relationship between ICD and AOR, we take a closer look at the differences between the base models and the instruct models for Qwen and LLaMA series model. As shown in Table 3, for the Qwen series models, although the AOR metric of instruct model decreases with model version iterations, the AOR of the base models remains relatively high with low average ICDs. Post-training is a common method for enhancing model capabilities. As shown in

Prompt:以下是中国关于计算机网络考试的单项选择题，请选出其中的正确答案。\n使用位填充方法，以01111110为位首flag，数据为01101111111111111111110010，求问传送时要添加几个0____\nE. 1\nF. 2\nG. 3\nH. 4\n答案:

Ground-truth: G

Response: C

Figure 6: **A badcase for the CEval dataset with the token set $\{E, F, G, H\}$.** The ground-truth label is 'G', but the model answer the 'C'.

Figure 5, compared with the base models, all the instruction-tuned models have a significantly improvement in the average ICD. For example, the average ICD of Qwen2-72B-Instruct is 23% higher than its pre-trained version. Notably, LLaMA3.1 has lower average ICD than its 3.0 version both in pre-trained model and post-trained model, which is consistent with the findings in the second section, but the post-training still provided 15% improvement. These results illustrate the efficient post-training can make a significant improvement for the robustness of the models. Similar conclusions were obtained on other datasets and models as well, more details can be found in Appendix A.3.

6 Conclusion

In this paper, we have investigated the robustness of Large Language Models (LLMs) against label variants in MCQs datasets. Extensive experiments on various LLMs and datasets show that LLMs generally experience a decline in performance when the choice tokens are varied. An analysis in the logits space revealed that LLMs exhibit a selection bias towards different token sets, which can lead to reduced confidence in their responses. This finding indicates that even state-of-the-art models do not rely solely on the content of the question when answering MCQs, underscoring the importance of improving model robustness and conducting robustness evaluations. Further analysis indicates that the selection bias towards the choice token set is related to the model size, specifically, smaller models exhibit more pronounced selection bias compared to larger models. We hypothesize that this is the combined effects of the training data and model capacity. By comparing the performance of base and instruct models, we demonstrated that post-training can significantly enhance the robustness of LLMs to label variants. This finding has important implications for improving the reliability and robustness of LLMs in practical applications. Furthermore, our study highlights that models with comparable performance in textual space can exhibit significant disparities in the logits space. This observation points to the necessary for a more nuanced evaluation approach, incorporating logits space analyses to capture the full spectrum of model behavior. Overall, our findings contribute valuable insights into the robustness of LLMs to label variants in MCQs dataset, and underscore the importance of evaluating LLM from both textual and logits perspectives.

7 Limitations

Our study requires extracting the answer option from the predicted text and relies on the logits output by the model. Consequently, our method has the following limitations.

Our current evaluation is limited to assessing the logit space of a single token, thus it can only be assessed on the multiple-choice question (MCQ) datasets. In scenarios where the ground-truth label of the prompt comprises multiple tokens, our method will be ineffective. Furthermore, in more complex situations where the ground-truth label is not only one, *e.g.*, mathematical expressions always have multiple valid variants, our method is currently incapable of handling such variability. Additionally, even within MCQ datasets, our approach might encounter failures when the models hava a bad instruction following.

Our approach is heavily reliant on the logits output by the model, therefore, currently, our evaluation can only be applied to the open-source models. This limitation precludes us from further assessing the closed-source models like GPT-4. Moreover, our evaluation does not directly reveal the underlying reasons for variations performance in the logit space. In future work, we will further investigate how the training dataset and internal model features influence the logits and the overall model's performance.

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. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–

13805, Bangkok, Thailand. Association for Computational Linguistics.

Baosong Yang An Yang. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do LLMs answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Yekun Chai, Yewei Fang, Qiwei Peng, and Xuhong Li. 2024. Tokenization falling short: On subword robustness in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1582–1599, Miami, Florida, USA. Association for Computational Linguistics.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *Preprint*, arXiv:2307.03109.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

Google. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Pengfei Hong, Navonil Majumder, Deepanway Ghosal, Somak Aditya, Rada Mihalcea, and Soujanya Poria. 2024. Evaluating llms' mathematical and coding competency through ontology-guided interventions. *Preprint*, arXiv:2401.09395.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Qintong Li, Leyang Cui, Xueliang Zhao, Lingpeng Kong, and Wei Bi. 2024a. GSM-plus: A comprehensive benchmark for evaluating the robustness of LLMs as mathematical problem solvers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2961–2984, Bangkok, Thailand. Association for Computational Linguistics.

Wangyue Li, Liangzhi Li, Tong Xiang, Xiao Liu, Wei Deng, and Noa Garcia. 2024b. Can multiple-choice questions really be useful in detecting the abilities of llms? *Preprint*, arXiv:2403.17752.

Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Joshua Robinson and David Wingate. 2023. Leveraging large language models for multiple choice question answering. In *The Eleventh International Conference on Learning Representations*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *CoRR*, abs/2302.00093.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. *Preprint*, arXiv:2310.16789.

Haoyu Wang, Guozheng Ma, Cong Yu, Ning Gui, Linrui Zhang, Zhiqi Huang, Suwei Ma, Yongzhe Chang, Sen Zhang, Li Shen, Xueqian Wang, Peilin Zhao, and Dacheng Tao. 2023a. Are large language models really robust to word-level perturbations? *Preprint*, arXiv:2309.11166.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023b. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *Preprint*, arXiv:2302.12095.

Zeyu Yang, Zhao Meng, Xiaochen Zheng, and Roger Wattenhofer. 2024. Assessing adversarial robustness of large language models: An empirical study. *Preprint*, arXiv:2405.02764.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. Don't make your llm an evaluation benchmark cheater. *Preprint*, arXiv:2311.01964.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. 2024a. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. *Preprint*, arXiv:2306.04528.

Kaijie Zhu, Qinlin Zhao, Hao Chen, Jindong Wang, and Xing Xie. 2024b. Promptbench: A unified library for evaluation of large language models. *Preprint*, arXiv:2312.07910.

A More experiment details

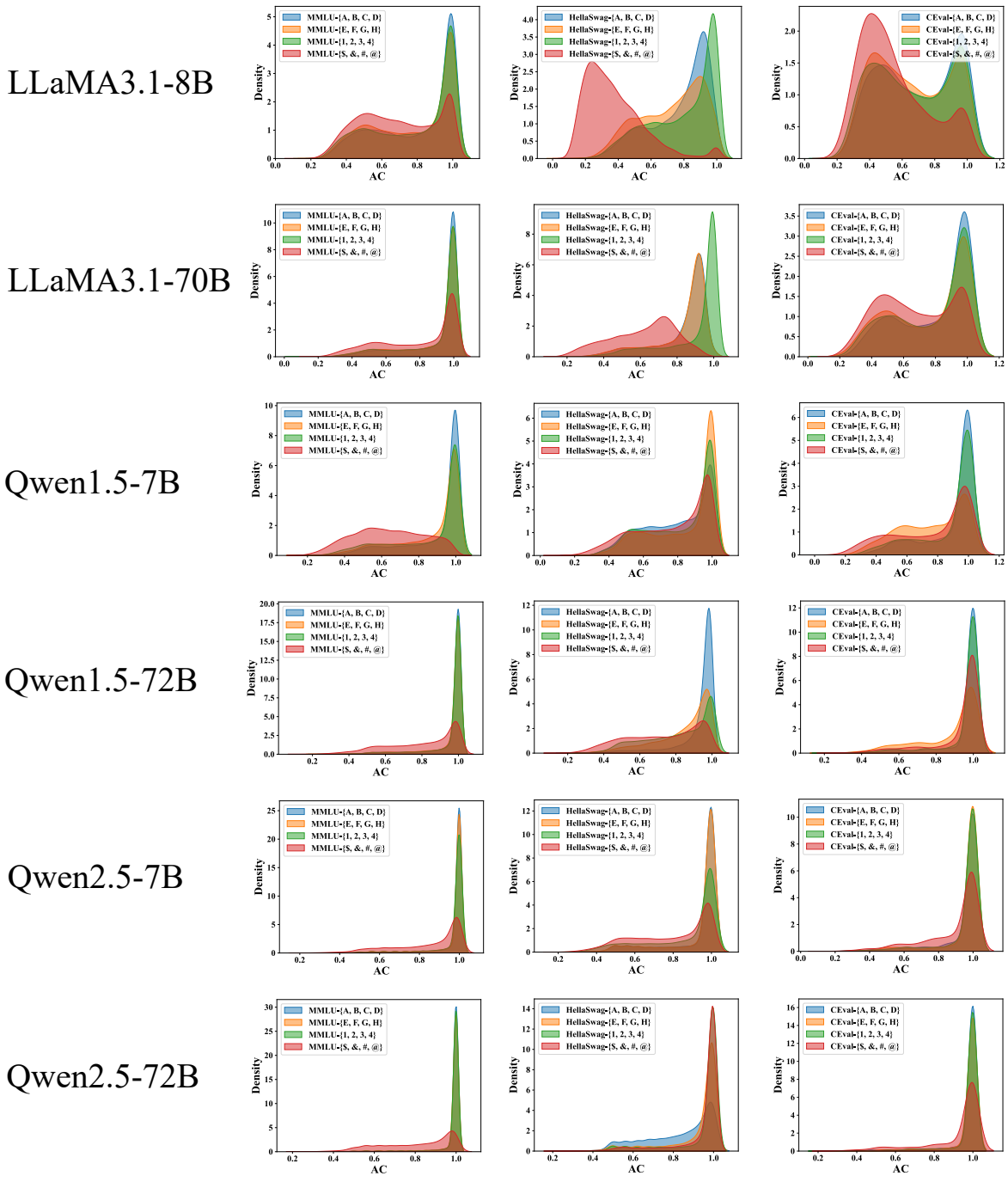

Figure 7: **The AC distributions**. Supplementary for Figure 3.

A.1 The score details of conducted experiments

The performance of different models across various choice token sets and datasets is shown in Table 4.

A.2 A Comprehensive Distribution of ACs

As shown in Figure 7, in most cases, the label variant will lead to a AC drop, particularly within the token set $\{\$, \&, \#, @\}$. Interestingly, the AC of Qwen2.5-72B-Instruct on HellaSwag with the variant token sets is even higher than that of the original token set $\{A, B, C, D\}$.

Model	MMLU				CEval				Hellaswag			
	ABCD	EFGH	1234	$&#@	ABCD	EFGH	1234	$&#@	ABCD	EFGH	1234	$&#@
Qwen1.5-7B	61.97	59.33	60.26	56.33	72.18	58.81	68.42	63.25	67.44	66.79	65.49	55.27
Qwen2-7B	70.89	69.12	70.14	66.0	83.0	81.02	81.5	78.92	78.91	77.08	77.89	69.5
Qwen2.5-7B	74.32	72.6	73.8	70.96	78.76	77.18	78.52	73.75	81.17	81.2	82.25	77.16
Qwen1.5-72B	77.36	76.08	76.66	75.7	84.21	71.62	83.97	80.22	87.78	86.72	87.63	84.83
Qwen2-72B	82.76	81.45	82.15	81.24	88.94	77.83	89.52	88.32	91.42	90.29	87.33	88.87
Qwen2.5-72B	84.21	83.31	83.25	81.92	89.3	87.71	88.69	84.84	90.6	89.98	89.91	89.32
LLaMA3-8B	68.27	67.26	67.19	65.53	53.86	52.01	53.75	50.97	74.64	74.78	72.35	65.36
LLaMA3.1-8B	69.24	67.83	68.81	64.46	55.64	55.42	55.08	52.9	75.32	71.84	73.0	62.87
LLaMA3-70B	80.99	79.81	80.08	78.24	66.97	67.46	67.0	64.89	88.71	88.2	87.39	86.57
LLaMA3.1-70B	82.28	81.55	81.42	78.73	69.14	69.82	68.86	66.6	86.53	87.5	85.77	86.2

Table 4: **The performance of different models across various choice token sets and datasets.**

A.3 A Comprehensive Distribution of ICDs

As shown in Figure 8, Figure 9, and Figure 10, in all cases, the pre-trained base model exhibited an increase in ICD after post-training, $i.e.$, the density of ICD around 1.0 increased.

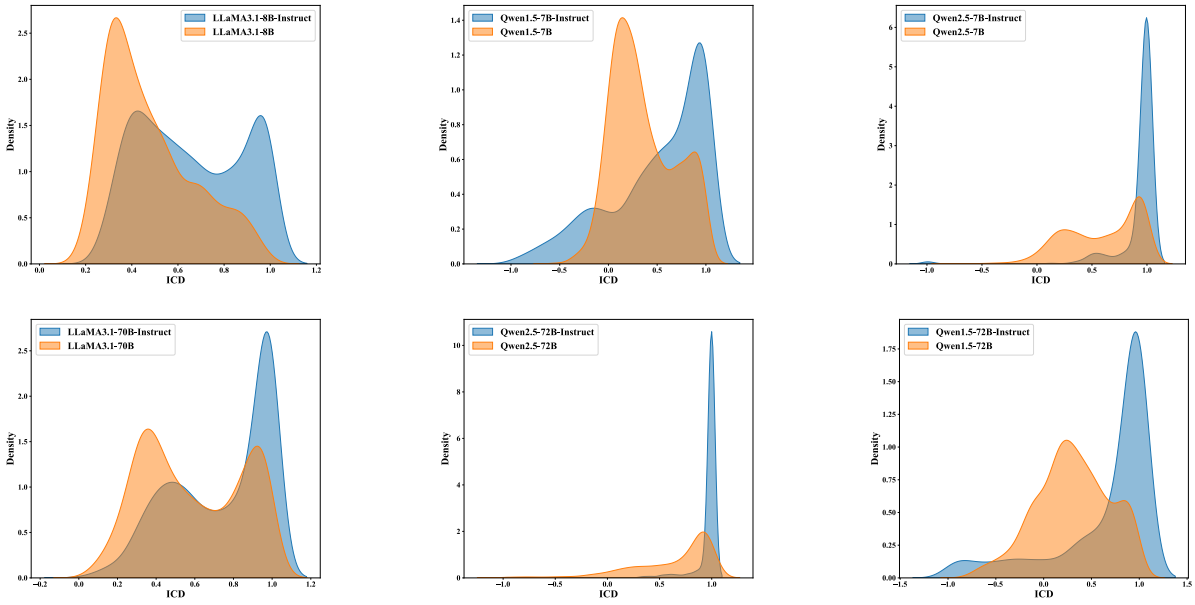

Figure 8: **The ICD distributions**. Supplementary for Figure 5 on the CEval dataset with the token set $\{E, F, G, H\}$.

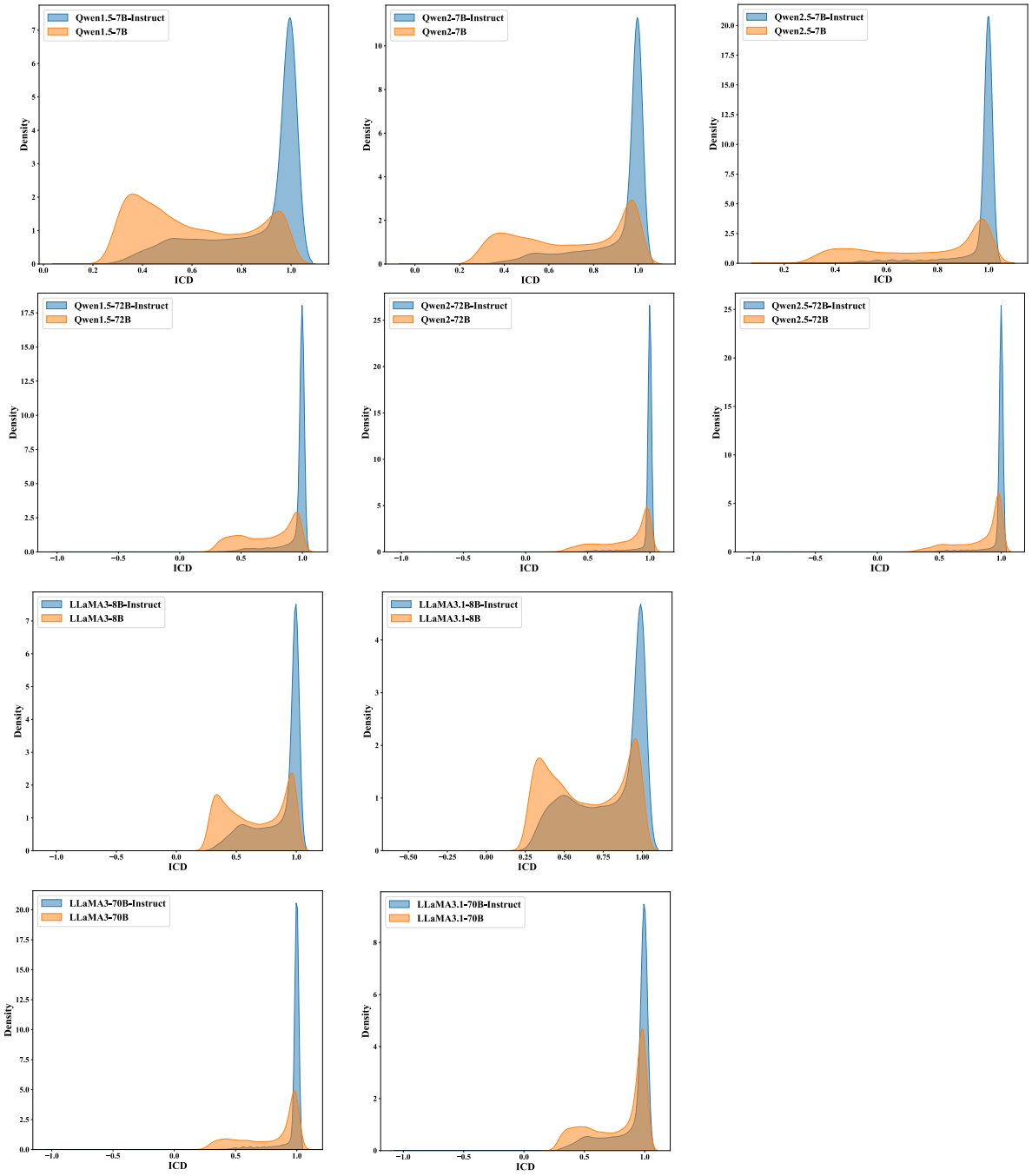

Figure 9: **The ICD distributions**. Supplementary for Figure 5 on the CEval dataset with the token set $\{1, 2, 3, 4\}$.

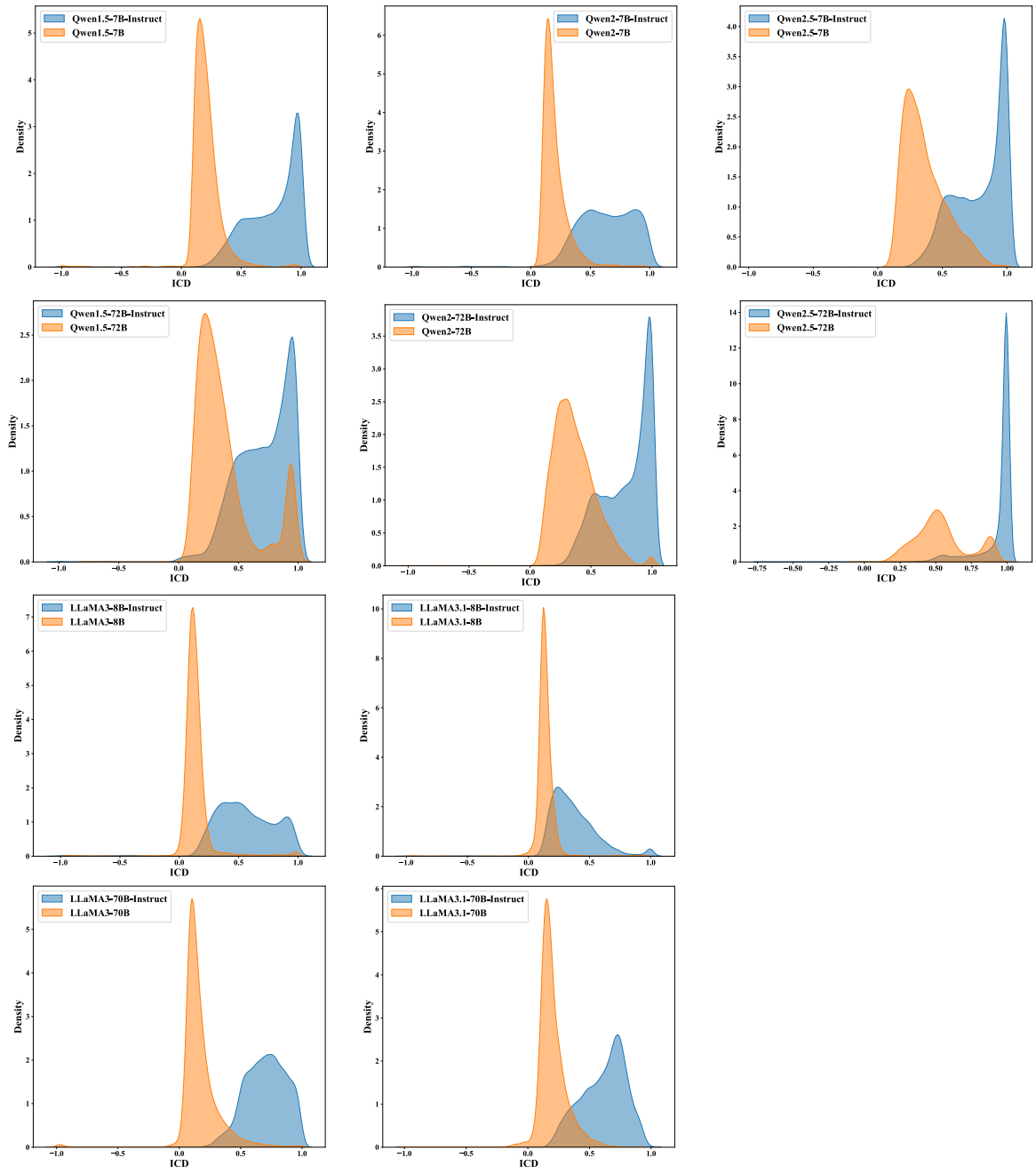

Figure 10: **The ICD distributions**. Supplementary for Figure 5 on the CEval dataset with the token set {$, &, #, @}.