CROSS-ENTROPY LOSS FALLS SHORT: WHAT CROSS-ENTROPY LOSS REVEALS ABOUT LARGE-LANGUAGE-MODEL PERFORMANCE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Scaling laws have made cross-entropy loss a central quantity for understanding and predicting the behavior of large language models, often with an implicit assumption that lower loss indicates better downstream performance. In this work, we revisit the reliability of this assumption on multiple-choice question answering (MCQA). Across diverse model families and datasets, we find that lower loss does not consistently correspond to higher accuracy. A key reason is that loss and accuracy depend on different quantities: cross-entropy loss measures how well a response is modeled, whereas accuracy is determined by whether the correct response has lower loss than its alternatives. We further observe that differences in loss-landscape geometry may make absolute loss values difficult to compare across model architectures and families. We then extend our analysis to an unsupervised setting, using the cross-entropy loss of the question context without ground-truth answers, and observe an even less consistent relationship with model performance. Motivated by the role of relative separation, we also examine loss dispersion across answer choices, but find that greater dispersion does not consistently indicate better performance. Together, our empirical and theoretical analyses reveal when and why loss-based quantities can become unreliable proxies for downstream model performance.

1 INTRODUCTION

Large language models (LLMs), such as GPT (Brown et al., 2020; Achiam et al., 2023), PaLM (Chowdhery et al., 2023), and LLaMA (Touvron et al., 2023), have achieved widespread adoption in both industrial applications and daily life, delivering state-of-the-art performance across diverse domains (Schlegel et al., 2025; Xiang et al., 2025). This success, however, has also created a new practical challenge. To further advance the success of LLMs on general downstream tasks, a fundamental question becomes increasingly important: *what factors determine LLM performance?*

Scaling laws, which empirically show that model performance follows power-law scaling with model size, dataset size, and training compute, have become a reliable tool for extrapolating model performance of large language models (LLMs) during the training process (Kaplan et al., 2020; Hoffmann et al., 2022; Grattafiori et al., 2024). Extending this principle to downstream tasks, prior work has shown that loss-to-loss scaling laws can be used to predict LLM downstream performance during at test time (Brandfonbrener et al., 2024), which is largely influenced by pretraining data (Mayilvahanan et al., 2025). However, these works all assume that lower supervised cross-entropy loss on multiple-choice question-answering datasets indicates better model performance. This naturally raises the question: **does lower loss necessarily imply better model performance for LLMs?**

Besides scaling laws, this assumption has also been used to assess optimization methods (Yu & Zhang, 2023; Chen et al., 2026), support theoretical analyses (Xie et al., 2025), and motivate other studies. However, several works have emphasized that loss and accuracy capture different aspects of performance (Terven et al., 2025; Chen et al., 2026; Schaeffer et al., 2025). Recent work (Schaeffer et al., 2024) further shows that probability mass assigned to incorrect responses contributes to the unpredictability of downstream-task accuracy, but does not examine how this relationship varies across model architectures or evaluation settings. Using loss as a proxy for model performance

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

without sufficient validation may lead to inappropriate model or checkpoint selection and potentially misleading conclusions about training effectiveness. To address this gap, we show that lower loss does not necessarily imply better performance across diverse model families and datasets, and investigate differences in loss-landscape geometry as a possible factor underlying this inconsistency.

Specifically, focusing on the multiple-choice question answering (MCQA) task, we investigate this question by conducting comprehensive experiments across different model families with diverse model sizes, including the OPT (Zhang et al., 2022) family, Cerebras-GPT (Dey et al., 2023) family, OpenELM (Mehta et al., 2024) family, DeepSeek-R1-Distill (DeepSeek-AI, 2025) family, Llama (Touvron et al., 2023) family, and Qwen (Bai et al., 2023; Team, 2024; Yang et al., 2024; Team, 2025) family. We find an inconsistent linear relationship between cross-entropy loss and loss-based accuracy across these diverse model families and evaluation datasets. This inconsistency persists in the unsupervised setting, where ground-truth answers are unavailable and we instead compute cross-entropy loss directly from the question context. Furthermore, loss dispersion, which quantifies the variability among losses assigned to different answer options, also exhibits an inconsistent relationship with model performance. To investigate the underlying causes of these phenomena, we further provide both empirical evidence and theoretical analysis.

Our contributions are summarized as follows:

- We examine the correlation between cross-entropy loss and model performance on MCQA datasets across diverse LLM families and evaluation settings. We show that lower loss does not always correspond to better performance, suggesting limitations in directly using loss as a performance indicator.
- We provide theoretical and empirical analyses to better understand why the relationship between loss values and model performance can vary across model families and datasets.
- We highlight several practical considerations when using loss as a proxy for model performance across different LLMs and evaluation settings.

2 Relation between loss and model performance for LLMs

2.1 Problem setup in the supervised setting

In this section, we focus on exploring the relationship between the accuracy of MCQA datasets and the cross-entropy loss across multiple choices in the supervised setting. In particular, we have access to M pre-trained large language models (LLMs) denoted as $\{f_m\}_{m=1}^{M}$, which parameters are donated as θ_m. For a MCQA dataset $\mathcal{D}$ with N question prompts $\mathcal{X} \in \{x_1, x_2, ..., x_N\}$ and their corresponding multiple-choice space $\mathcal{Y} \in \{A, B, C, D, ...\}$, we can observe the option description of each choice $\{\gamma_A^i, \gamma_B^i, \gamma_C^i, \gamma_D^i, ...\}$, where γ_k^i denotes the answer k's prompt of the i-th question. We define the correct answer to the i-th question as $\mathbf{y_i}$. Let $z_{f_m}(x_i) \in \mathbb{R}^{|\mathcal{Y}|}$ be the logits of f_m on the i^{th} prompt x_i. The predicted probabilities are

$$p_{f_m}(y = j|x_i) = \frac{exp(z_{f_{m,j}}(x_i))}{\sum_{j' \in \mathcal{Y}} exp(z_{f_{m,j'}}(x_i))}.$$

Loss-based accuracy. In this work, we utilize loss-based accuracy as the measurement of the ground-truth accuracy of LLMs. Specifically, we treat the response with the minimum auto-regression cross-entropy loss values as the correct answer, and use it to calculate the accuracy as

$$Acc_{\mathcal{D}}(f_m) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(\hat{y}_i^m = \mathbf{y}_i), \tag{1}$$

where $\mathbf{y}_i$ denotes as the correct answer of the i-th prompt, and $\mathbb{1}(\cdot)$ denotes the the indicator function, and $\hat{y}_i^m$ denotes the predicted answer of the m-th LLM estimated as follows:

$$\hat{y}_i^m = \arg\min_j \frac{1}{|T_j|} \sum_{t=0}^{|T_j|} \ell\big(f_m(x_i, \gamma_{j,<t}^i), \gamma_{j,t}^i\big) = \arg\min_j -\frac{1}{|T_j|} \sum_{t=1}^{|T_j|} \gamma_{j,t}^i \log P_{f_m}(\gamma_{j,t}^i | x_i, \gamma_{j,<t}^i),$$

where T_j denotes the number of tokens for the j-th response.

Table 1: Coefficient of determination (R^2) and Spearman rank correlation (ρ) between loss-based accuracy and **supervised** cross-entropy loss values across model families and evaluation datasets. The detailed numerical results are shown in Appendix D.

Dataset	OPT-family		Cerebras-GPT-family		DeepSeek-family		Qwen-family		Llama-family		OpenELM-family	
	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ
openbookqa	0.9066	-0.7714	0.9787	-1.0000	0.6237	-0.7143	0.6977	-0.8598	0.8303	-0.9000	0.6989	-0.7537
piqa	0.9959	-1.0000	0.9928	-1.0000	0.7041	-0.8286	0.8124	-0.9273	0.8789	-0.9000	0.2661	-0.4058
mmlu	0.9996	-1.0000	0.9121	-0.9643	0.8669	-0.9429	0.2390	-0.5148	0.0431	-0.2000	0.0643	0.0286
hellaswag	0.9996	-1.0000	0.9950	-1.0000	0.9113	-0.9429	0.9211	-0.9818	0.9793	-1.0000	0.5349	-0.7143

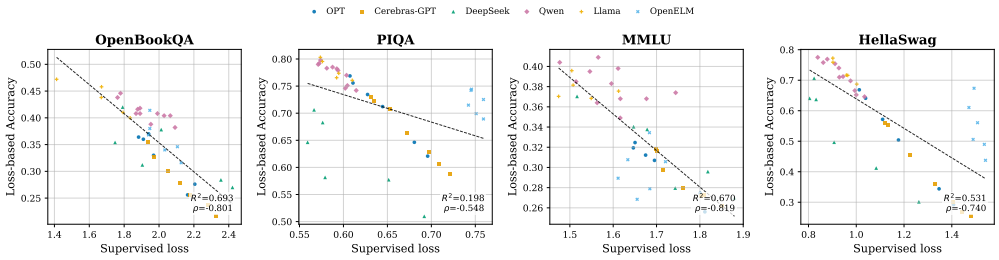


Figure 1: The weak correlation between the supervised cross-entropy loss values and loss-based accuracy across different families, including OPT family, Cerebras-GPT family, DeepSeek family, Qwen family, Llama family and OpenELM family on four datasets (i.e., OpenBookQA, PIQA, MMLU, and HellaSwag).

Supervised cross-entropy loss. To obtain the supervised cross-entropy loss, we first start by calculating the negative log likelihood (NLL) across multiple choices, where NLL of the j-th response can be expressed as:

$$NLL(z_{f_{m,j}}(x_i)) = -\frac{1}{|T_j|}\sum_{t=1}^{|T_j|} logP_{f_m}(\gamma^i_{j,t}|x_i, \gamma^i_{j,<t}).$$

Then the cross-entropy loss value is calculated as:

$$\mathbb{E}[l(\mathcal{D})] = \frac{1}{|\mathcal{Y}|}\sum_{j\in\mathcal{Y}} \mathbb{1}(\mathbf{y_i} = j)\log p_{f_m}(NLL(z_{f_{m,j}}(x_i))),$$

where $\mathbb{1}(\cdot)$ is the indicator function, and p_{f_m} denotes the softmax function across choices.

Goal. This section aims to investigate the relationship between loss-based accuracy and supervised cross-entropy loss, with the objective of evaluating the widely adopted assumption that lower loss values correspond to superior model performance.

2.2 Lower loss does not necessarily imply better performance

In this section, we demonstrate that **the common assumption that lower loss values indicate better model performance does not hold across different model architectures and datasets**. Specifically, we observe a clear linear relationship between cross-entropy loss and model performance on MCQA datasets within certain model families, such as the OPT family (Zhang et al., 2022), whose model sizes range from 125M to 13B parameters, and the Cerebras-GPT family (Dey et al., 2023), which ranges from 111M to 13B parameters. However, this linear trend becomes inconsistent within several other families, including the DeepSeek-R1-Distill family (DeepSeek-AI, 2025), ranging from 1.5B to 70B parameters; the OpenELM family (Mehta et al., 2024), ranging from 270M to 3B parameters; the Qwen family (Bai et al., 2023), covering 3B to 8B models across different versions; and the Llama family (Touvron et al., 2023), covering 3B and 8B models across different versions, on at least one dataset. More experimental results are shown in Appendix D.

Inconsistent linear relation within identical model families. Within the identical model family, models of different sizes vary in parameter count and depth, while sharing the same training strategy

and training data. In this analysis, we investigate the correlation between cross-entropy loss values and model size. In Table 1, we observe a clear linear relationship between model performance and cross-entropy loss within the OPT, the DeepSeek-R1-Distill, and Cerebras-GPT families across all four benchmarks. For example, within the Cerebras-GPT family, coefficient of determination (R^2) reaches 0.995 and Spearman rank correlation (ρ) reaches -1 on Hellaswag (Zellers et al., 2019).

On the other hand, the negative correlation between cross-entropy loss and model performance is inconsistent across a wide range of model families, including the OpenELM, the Qwen, and the Llama family. For example, within the Llama family on MMLU (Hendrycks et al., 2021), the coefficient of determination is only $R^2 = 0.043$, and the Spearman rank correlation is $\rho = -0.2$.

Weak linear relation across different model families. Across diverse model families, models differ substantially in their tokenization schemes, training strategies, training datasets, and architectural designs. Now, we investigate the correlation between cross-entropy loss values and loss-based accuracy across different model families. Figure 1 presents scatter plots of loss-based accuracy against supervised cross-entropy loss across different model families on the same dataset. The figures show a negative linear trend between the two metrics across diverse model families; however, this relationship is not sufficiently strong or consistent to support the conclusion that lower loss values reliably imply better model performance. For instance, on PIQA (Bisk et al., 2020), the coefficient of determination is only $R^2 = 0.198$, and the Spearman rank correlation is $\rho = 0.548$.

Unreliable ranking estimation of model performance across datasets and model architectures. A qualified metric for measuring model performance should be consistently comparable across diverse datasets and model architectures in practical model selection scenarios. Figure 2 shows that although supervised cross-entropy loss is negatively associated with model performance, this relationship is not consistent across diverse datasets and model architectures: lower loss values do not necessarily imply better performance. In particular, when loss-based accuracy is low, cross-entropy loss struggles to accurately estimate the relative ranking of model performance. Therefore, although lower loss may indicate better model performance in certain cases, this negative correlation does not always hold. For a similar cross-entropy loss, loss-based accuracy can vary substantially. For example, DeepSeek-R1-Distill-Llama-70B (DeepSeek-AI, 2025) achieves a cross-entropy loss of 1.748 on OpenBookQA (Mihaylov et al., 2018), with an accuracy of 0.354, whereas DeepSeek-R1-Distill-Llama-8B achieves a similar loss of 1.743 on MMLU (Hendrycks et al., 2021), but with a markedly lower accuracy of 0.280.

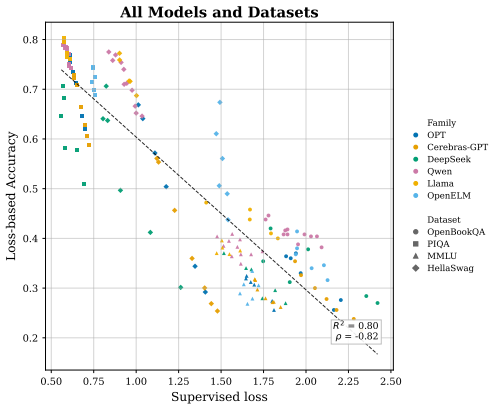


Figure 2: Scatter plot of loss-based accuracy versus supervised cross-entropy loss across all models and datasets, which illustrates that a lower loss unnecessarily indicates better model performance.

3 Extension to Unsupervised Settings

Annotations for downstream-task datasets are often expensive and difficult to obtain. Therefore, it is important to examine whether lower loss indicates better performance in an unsupervised setting, where ground-truth answers for MCQA datasets are unavailable. Specifically, we replace supervised cross-entropy loss with an unsupervised counterpart, computed as the next-token cross-entropy of the question context for each MCQA instance, with details provided in Appendix B.

3.1 The Inconsistent Linear Trend Phenomenon

In this section, we further demonstrate that **lower loss values do not necessarily indicate better model performance across different model architectures and datasets**. Specifically, we observe a clear linear trend between model performance on MCQA tasks and the corresponding loss values

Table 2: Coefficient of determination (R^2) and Spearman rank correlation (ρ) between loss-based accuracy and **unsupervised** cross-entropy loss values across model families and evaluation datasets. The detailed numerical results are shown in Appendix D.

Dataset	OPT		Cerebras-GPT		DeepSeek		Qwen		Llama		OpenELM	
	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ	R^2	ρ
OpenBookQA	0.7290	-0.7714	0.9324	-0.9643	0.8387	-0.9429	0.3987	-0.7960	0.1815	0.6000	0.0808	0.1429
PIQA	0.8551	-0.8286	0.9562	-1.0000	0.8251	-0.9429	0.1156	-0.4360	0.0416	0.0000	0.0607	-0.3143
MMLU	0.9813	-0.9429	0.8621	-0.9643	0.8650	-0.9429	0.0454	-0.0460	0.0424	-0.1000	0.0171	-0.0857
HellaSwag	0.9884	-1.0000	0.8508	-1.0000	0.9784	-0.9429	0.2177	-0.3550	0.5114	-0.7000	0.0619	0.1429

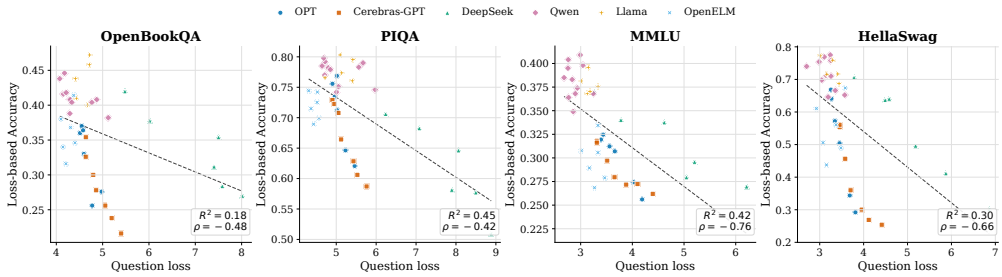


Figure 3: The weak correlation between the cross-entropy loss values on question prompts and loss-based accuracy across different families, including OPT family, Cerebras-GPT family, DeepSeek family, Qwen family, Llama family and OpenELM family on four datasets (i.e., OpenBookQA, PIQA, MMLU, and HellaSwag).

within certain model families under several evaluation conditions, including the OPT family (Zhang et al., 2022), with parameter scales ranging from 125 million to 13 billion; the Cerebras-GPT family (Dey et al., 2023), ranging from 111 million to 13 billion parameters; and the DeepSeek-R1-Distill family (DeepSeek-AI, 2025), ranging from 1.5 billion to 70 billion parameters. However, this linear trend does not consistently hold across all model families. We observe weaker or inconsistent relationships for the OpenELM family (Mehta et al., 2024), ranging from 270 million to 3 billion parameters; the Qwen family (Bai et al., 2023), covering models from 3 billion to 8 billion parameters across different versions; and the Llama family (Touvron et al., 2023), covering 3 billion and 8 billion parameter models across different versions.

3.1.1 Inconsistent relationship between cross-entropy loss values of question prompts and accuracy

We explore the correlation between the loss values and the model performance on multiple-choice QA datasets across diverse model sizes and model families. For certain model families including the OPT family, the Cerebras-GPT family, and the DeepSeek family, we find a strong linear relationship between auto-regressive loss values on question prompts and loss-based accuracy across both model sizes and model families.

Inconsistent linear relation within the identical model family. Within the identical model family, models of different sizes vary in parameter count and depth, while sharing the same training strategy and training data. In this analysis, we investigate the correlation between loss values on question prompts and model size. As shown in Table 2, within certain model families such as the OPT, the Cerebras-GPT, and the DeepSeek family, there is a strong negative linear relationship between auto-regressive cross-entropy loss and model performance. For instance, the DeepSeek family achieves an R^2 of 0.9784 and an ρ of -0.9429 on HellaSwag (Zellers et al., 2019).

However, this pattern does not hold across a broader range of model families. We observe that for other prominent families, such as the Qwen, Llama, and OpenELM families, the linear correlation between loss and model performance breaks down. In particular, for the OpenELM family, the coefficient of determination R^2 remains at most 0.1 and Spearman's rank correlation ρ stays below 0.35 across all four datasets. A similar trend is observed in the Qwen family, where the highest R^2 reaches only 0.39 and the highest ρ is -0.8. In both cases, the linear relationship is substantially weaker than that observed in the previously discussed model families.

Weak linear relation across different model families. Across diverse model families, models differ substantially in their tokenization schemes, training strategies, training datasets, and architectural designs. Now, we investigate the correlation between cross-entropy loss values and loss-based accuracy across different model families. Figure 3 presents the scatter plot of loss-based accuracy against cross-entropy loss on question prompts for all the models. Although a negative correlation between cross-entropy loss and model performance is observable, the relationship is weak overall, with the highest coefficient of determination R^2 reaching only 0.45. Furthermore, this negative correlation is primarily driven by a subset of model families exhibiting strong negative correlations, including the OPT, Cerebras-GPT, and DeepSeek families.

Inconsistent linear relation across different datasets. Beyond the unreliable relationship observed across diverse model families, cross-entropy loss is also **not** consistently aligned with model performance for the same model evaluated across different datasets. For example, Table 11 shows that for OPT-6.7B, the loss-based accuracy values on OpenBookQA, PIQA, MMLU, and HellaSwag are 0.3600, 0.7557, 0.3246, and 0.6409, respectively, whereas the corresponding unsupervised cross-entropy loss values are 4.5128, 4.9053, 3.4354, and 3.2663. These four observations yield only moderate explanatory power, with $R^2 = 0.0901$ and $\rho = 0.4000$, indicating that cross-entropy loss alone does not provide a robust or sufficient proxy for downstream task performance.

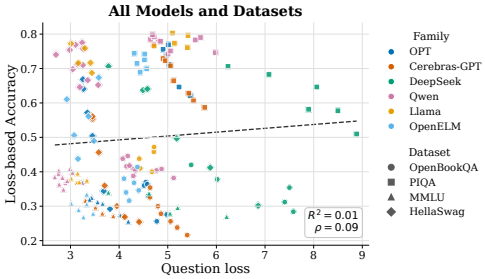


Figure 4: Scatter plot of loss-based accuracy versus unsupervised cross-entropy loss across all models and datasets, which illustrates that a lower loss unnecessarily indicates better model performance.

No trend across datasets and model architectures. Accuracy serves as a robust supervised metric for assessing model performance, offering a measure of classification capability on balanced datasets that generalizes consistently across diverse model architectures and datasets. However, Figure 4 demonstrates that loss values do not necessarily correlate with accuracy when examined across all datasets and model architectures. Consequently, the widely held prior assumption that lower loss values indicate better model performance is not generally valid.

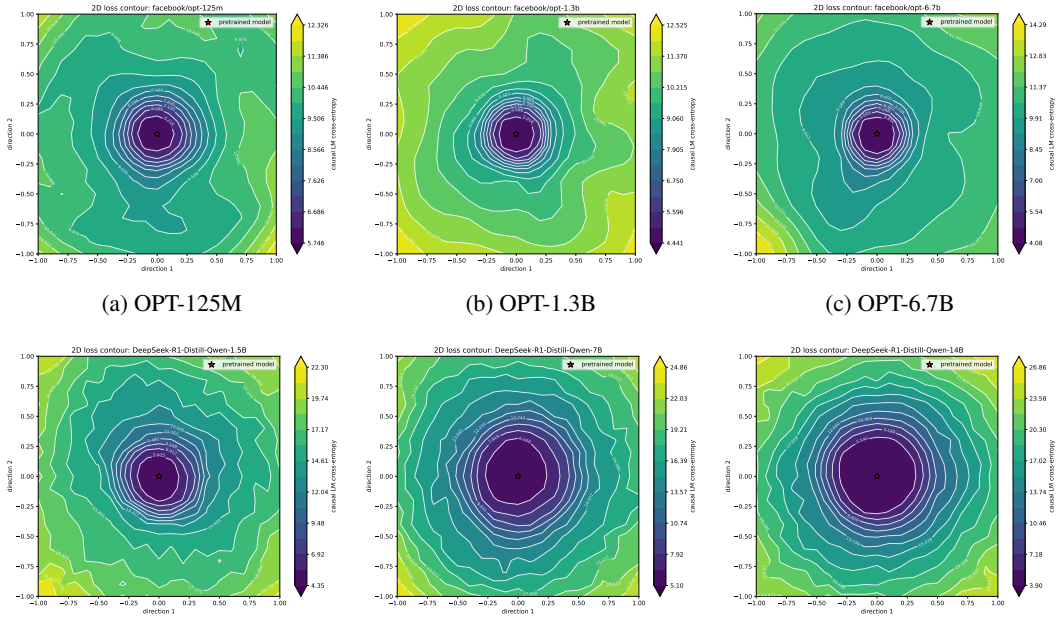
4 Understand the Unreliable Relationship in Both Supervised and Unsupervised Settings

We have empirically demonstrated that lower cross-entropy loss values do not always correspond to better performance for diverse model families and datasets in both supervised and unsupervised settings. Now, we investigate the underlying reasons for the counterintuitive results.

There is a gap between accuracy and cross-entropy loss. From a metrics perspective, lower loss values do not necessarily indicate superior model performance, as loss functions and accuracy quantify different aspects of model predictions. Specifically, since the pre-trained parameters are obtained by maximum likelihood estimation with respect to the training data (Ciampiconi et al., 2026), cross-entropy loss quantifies how well a given token sequence is predicted under the learned model distribution. In contrast, loss-based accuracy evaluates whether the candidate response assigned the lowest loss corresponds to the correct answer. These two quantities therefore capture distinct aspects of model behavior, inducing a nontrivial gap between token-level likelihood and candidate-level correctness.

Theoretically, let $L^*(x)$ denote the cross-entropy loss of the correct response, and let $L_k(x)$ denote the cross-entropy loss of the k-th candidate response. Loss-based accuracy is then determined by whether the correct response achieves the minimum loss among all candidates:

$$\mathrm{Acc}(x) = \mathbb{1}\left[L^*(x) < \min_{k \neq *} L_k(x)\right].$$



(a) OPT-125M (b) OPT-1.3B (c) OPT-6.7B

(d) DeepSeek-Qwen-1.5B (e) DeepSeek-Qwen-7B (f) DeepSeek-Qwen-14B

Figure 5: 2-D loss-landscape visualizations for the OPT and DeepSeek-R1-Distill-Qwen model families. The top row shows OPT models with increasing sizes, and the bottom row shows DeepSeek-R1-Distill-Qwen models with increasing sizes.

Equivalently, defining the loss margin as

$$\Delta(x) = \min_{k \neq *} \{L_k(x) - L^*(x)\},$$

we have

$$\mathrm{Acc}(x) = \mathbb{1}[\Delta(x) > 0].$$

This formulation shows that a small correct-response loss $L^*(x)$ alone does not guarantee correctness: the model is correct only when $L^*(x)$ is smaller than the losses assigned to all incorrect candidates.

From a distributional perspective, different model architectures across diverse model families map variant downstream-task datasets into different distribution spaces. Thus, loss values may correspond to different regions of the loss landscape, rendering direct comparisons across such losses potentially meaningless (Chen et al., 2024).

Why does estimation performance degrade across model families? Figures 1 and 3 show that model-performance estimation is less accurate across model families than within a single family. **One possible explanation is that different model families exhibit substantially different loss-landscape geometries.** To investigate this hypothesis, we visualize the 2-D loss landscapes (Li et al., 2018) of the OPT and DeepSeek-Qwen families in Figure 5. The results reveal clear family-dependent differences. Within the OPT family, the contours around the pretrained solution are relatively compact and smooth, and their spatial extent does not increase consistently with model size. This observation contrasts with the findings of Liu et al. (2023), who report that flatter models tend to achieve better downstream performance. In contrast, the DeepSeek-Qwen models exhibit a larger dynamic range and a more pronounced expansion of the surrounding loss landscape as the model size increases from 1.5B to 7B and 14B parameters. The smaller DeepSeek-Qwen model also shows more irregular contour structures, whereas the larger models exhibit broader basins around the pretrained solution. These differences may make direct comparisons across models unreliable.

When can loss be used to predict model performance? Here, we identify two possible conditions under which loss can be used to predict model performance: 1) the training objective is aligned

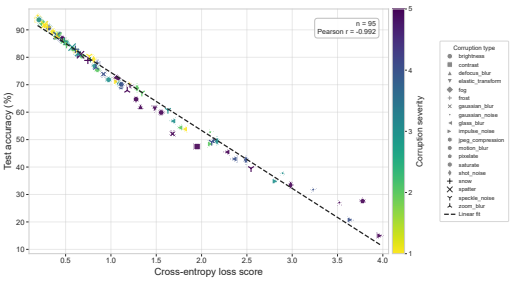


Figure 6: Classification accuracy versus cross-entropy loss on CIFAR-10C with ResNet-18. Lower loss does not necessarily indicate better model performance under small perturbations.

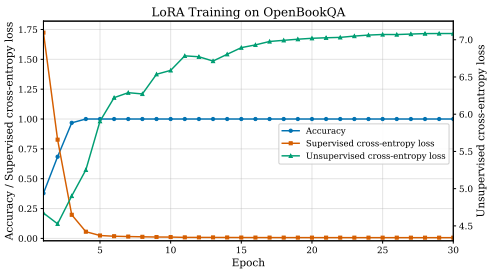


Figure 7: Training dynamics of loss-based accuracy, supervised cross-entropy loss, and unsupervised cross-entropy loss during LoRA fine-tuning.

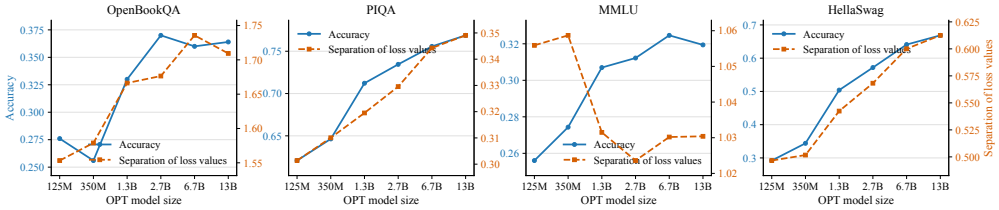


Figure 8: The inconsistent relationship between the separability of loss values across multiple choices and model performance within the OPT family. The model performance is measured by the loss-based accuracy shown as Eq. 1, while the separability of loss values is gauged as Eq. 2.

with the evaluation metric; and 2) different model architectures map downstream-task datasets into sufficiently similar loss distribution spaces, making cross-model loss comparisons meaningful.

To verify this assumption, we conduct a simple experiment in which perturbations of varying severity are applied to the original CIFAR-10 images, formulating the dataset, CIFAR-10C. We then examine the relationship between classification accuracy and cross-entropy loss using ResNet-18 pre-trained on the original CIFAR-10 training data. Figure 6 shows a strong negative linear correlation between cross-entropy loss and accuracy. Since these perturbations only moderately shift the data distribution, the resulting loss values are likely to remain within comparable regions of the same loss landscape. In addition, the cross-entropy loss of classification tasks is aligned with the pre-training process. Consequently, lower loss is more likely to indicate better model performance.

This also explains why loss values can be used to predict model performance during training. In Figure 7, we conduct a LoRA fine-tuning experiment (Hu et al., 2022) on OpenBookQA using DeepSeek-R1-Distill-Qwen-1.5B. We observe that, before convergence, supervised cross-entropy loss decreases as accuracy increases, whereas unsupervised cross-entropy loss increases. This phenomenon can be explained by the fact that the fine-tuning objective and the accuracy metric are aligned: fine-tuning encourages the model to assign lower cross-entropy loss to the correct answer and relatively higher loss to incorrect answers, thereby reducing supervised cross-entropy loss and improving accuracy before convergence. More fundamentally, fine-tuning makes the supervised cross-entropy loss distributions more comparable across models, whereas the unsupervised cross-entropy loss distributions remain distinct.

Why does the supervised setting exhibit a stronger negative correlation than the unsupervised setting? It is evident that the negative linear relationship is stronger in the supervised setting than in the unsupervised setting. A possible explanation is that the training objective is not aligned with unsupervised cross-entropy loss, which measures how well the model models the question context rather than whether it ranks the correct answer above the distractors. Consequently, question-context loss can predict performance only under much stronger assumptions, such as when better question modeling is consistently coupled with better answer discrimination across the compared models. Otherwise, lower question-context loss does not necessarily imply higher MCQA accuracy.

5 THE FAILURE OF LOSS DISPERSION ACROSS MULTIPLE CHOICES TO ESTIMATE MODEL PERFORMANCE

There is another intuition about the unsupervised cross-entropy loss that a well-informed model should assign substantially lower cross-entropy loss to the correct (or high-quality) answer than to incorrect (or low-quality) ones. However, in this section, we theoretically and empirically demonstrate that the separation of loss values across multiple choices is unreliable. The relative experimental setup is shown in Appendix C.

Experimental analysis. The numerical results are presented in Figure 8. Within each model family, we observe an inconsistent trend between the separability of loss values and model performance. For example, the linear relationship can be negative for certain model families (e.g., DeepSeek-R1-distillation) and datasets (e.g., MMLU), while it remains positive for the OPT family on OpenbookQA. It indicates the failure of loss separability across multiple choices in selecting the optimal model architecture. Additional detailed experimental results are provided in Appendix D.

Why loss dispersion is unreliable. This inconsistent relationship arises because the dataset-level standard deviation of loss values across multiple choices does not guarantee an upper bound on the prediction error of the corresponding dataset. Specifically, although the standard deviation of loss values can provide an upper bound on prediction error at the individual data-point level, this property does not necessarily extend to the dataset level. We formalize this observation as follows.

Theorem 5.1. *(Upper bound of prediction error at the data-point level) For each question x, let $\mathcal{Y}(x) = \{y_1, y_2, \ldots, y_K\}$ be a set of K candidate answers, exactly one of which y^* is correct with its corresponding probability $p^*(\theta; x)$. Let $\sigma(x) = \mathrm{std}_{k\in[K]}\big(L_k(\theta; x)\big)$, denoting the standard deviation of loss values across multiple choices, and $\mathrm{Error}(\theta) = 1 - p^*(\theta; x)$ for any K candidate answers meeting $K \geq 2$, we have:*

$$\mathrm{Error}(\theta) \leq (K-1)\exp\left(-\frac{K}{\sqrt{K-1}}\,\sigma(x)\right).$$

Theorem 5.1 demonstrates that the standard deviation of loss values across multiple choices provides an upper bound on the prediction error at the data-point level. The detailed proof is given in Appendix E.1. However, since $\mathbb{E}\big[(K-1)\exp\left(-\frac{K}{\sqrt{K-1}}\,\sigma(x)\right)\big] \geq (K-1)\exp\left(-\frac{K}{\sqrt{K-1}}\,\mathbb{E}\big(\sigma(x)\big)\right)$, where $\mathbb{E}(\sigma(x))$ equal to Eq. 2, it shows that the separability of loss values across multiple choices cannot guarantee the negative correlation with model performance at the dataset level.

6 CONCLUSION

This work investigates whether lower cross-entropy loss consistently indicates better model performance. Focusing on multiple-choice question answering, we examine this relationship in supervised and unsupervised settings and report three main observations. First, lower cross-entropy loss does not always correspond to higher accuracy. Second, the relationship is more consistent within a model family than across different families. Third, supervised cross-entropy loss shows a stronger relationship with accuracy than loss computed from question contexts. We further examine differences in loss-landscape geometry as a possible factor affecting cross-model loss comparison, and find that loss dispersion across answer choices is likewise not consistently associated with better performance. Taken together, these observations suggest that the usefulness of loss as a performance proxy depends on both how the loss is defined and which models and datasets are being compared.

Potential limitations and future directions. Several directions may further improve the reliability of loss-based performance estimation. One promising direction is to develop methods that make loss values more comparable across different model architectures and datasets. It is also worth exploring whether combining absolute loss with relative signals, such as answer margins, uncertainty, or distributional statistics, can provide a more stable indicator of model performance. Extending these analyses beyond MCQA to open-ended generation and reasoning-intensive tasks may further clarify the broader applicability of loss-based evaluation.

AI USE STATEMENT

In this work, we used generative AI tools for polishing English writing and create or modify scientific figures or images. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, or assisting in the writing of proofs, and cleaning and reformatting dataset are not applicable to this work. Additionally, we used generative AI tools for editing a research paper to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, 2020.

David Brandfonbrener, Nikhil Anand, Nikhil Vyas, Eran Malach, and Sham Kakade. Loss-to-loss prediction: Scaling laws for all datasets. *arXiv preprint arXiv:2411.12925*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. In *International Conference on Learning Representations*, volume 2024, pp. 19199–19225, 2024.

Huanran Chen, Huaqing Zhang, Xiao Li, Yinpeng Dong, Ke Shen, and Jun Zhu. Nexus: Same pretraining loss, better downstream generalization via common minima. *arXiv preprint arXiv:2604.09258*, 2026.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240): 1–113, 2023.

Peter Christen, David J Hand, and Nishadi Kirielle. A review of the f-measure: its history, properties, criticism, and alternatives. *ACM Computing Surveys*, 56(3):1–24, 2023.

Lorenzo Ciampiconi, Adam Elwood, Marco Leonardi, Ashraf Mohamed, and Alessandro Rozza. A survey and taxonomy of loss functions in machine learning. *AI*, 7(4):128, 2026.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Weijian Deng and Liang Zheng. Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15069–15078, 2021.

Weijian Deng, Stephen Gould, and Liang Zheng. What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning (ICML)*, pp. 2579–2589, 2021.

Nolan Dey, Gurpreet Gosal, Hemant Khachane, William Marshall, Ribhu Pathria, Marvin Tom, Joel Hestness, et al. Cerebras-gpt: Open compute-optimal language models trained on the cerebras wafer-scale cluster. *arXiv preprint arXiv:2304.03208*, 2023.

Saurabh Garg, Sivaraman Balakrishnan, Zachary C Lipton, Behnam Neyshabur, and Hanie Sedghi. Leveraging unlabeled data to predict out-of-distribution performance. *arXiv preprint arXiv:2201.04234*, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Devin Guillory, Vaishaal Shankar, Sayna Ebrahimi, Trevor Darrell, and Ludwig Schmidt. Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1134–1144, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, DDL Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 10, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 1(2):3, 2022.

Yiding Jiang, Vaishnavh Nagarajan, Christina Baek, and J Zico Kolter. Assessing generalization of sgd via disagreement. *arXiv preprint arXiv:2106.13799*, 2021.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. Same pre-training loss, better downstream: Implicit bias matters for language models. In *International Conference on Machine Learning*, pp. 22188–22214. PMLR, 2023.

Yuzhe Lu, Yilong Qin, Runtian Zhai, Andrew Shen, Ketong Chen, Zhenlin Wang, Soheil Kolouri, Simon Stepputtis, Joseph Campbell, and Katia Sycara. Characterizing out-of-distribution error via optimal transport. *arXiv preprint arXiv:2305.15640*, 2023.

Omid Madani, David Pennock, and Gary Flake. Co-validation: Using model disagreement on unlabeled data to validate classification algorithms. *Advances in Neural Information Processing Systems (NeurIPS)*, 17, 2004.

Prasanna Mayilvahanan, Thaddäus Wiedemer, Sayak Mallick, Matthias Bethge, and Wieland Brendel. Llms on the line: Data determines loss-to-loss scaling laws. *arXiv preprint arXiv:2502.12120*, 2025.

Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, and Mohammad Rastegari. OpenELM: An Efficient Language Model Family with Open Training and Inference Framework. *arXiv.org*, April 2024. URL https://arxiv.org/abs/2404.14619v1.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Emmanouil Platanios, Hoifung Poon, Tom M Mitchell, and Eric J Horvitz. Estimating accuracy from unlabeled data: A probabilistic logic approach. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.

Emmanouil Antonios Platanios, Avinava Dubey, and Tom Mitchell. Estimating accuracy from unlabeled data: A bayesian approach. In *International Conference on Machine Learning (ICML)*, pp. 1416–1425. PMLR, 2016.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *2007 15th European signal processing conference*, pp. 606–610. IEEE, 2007.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier ai models with scale remained elusive? In *ICML 2024 Next Generation of AI Safety Workshop*, 2024.

Rylan Schaeffer, Hailey Schoelkopf, Brando Miranda, Gabriel Mukobi, Varun Madan, Adam Ibrahim, Herbie Bradley, Stella Biderman, and Sanmi Koyejo. Why has predicting downstream capabilities of frontier ai models with scale remained elusive? In *International Conference on Machine Learning*, pp. 53177–53266. PMLR, 2025.

Katja Schlegel, Nils R Sommer, and Marcello Mortillaro. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*, 3(1):80, 2025.

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Juan Terven, Diana-Margarita Cordova-Esparza, Julio-Alejandro Romero-González, Alfonso Ramírez-Pedraza, and Edgar A Chavez-Urbiola. A comprehensive survey of loss functions and metrics in deep learning. *Artificial Intelligence Review*, 58(7):195, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Weijie Tu, Weijian Deng, Dylan Campbell, Yu Yao, Jiyang Zheng, Tom Gedeon, and Tongliang Liu. Ranked from within: Ranking large multimodal models without labels. In *Forty-second International Conference on Machine Learning*, 2025.

Lai Wei, Zhiquan Tan, Chenghai Li, Jindong Wang, and Weiran Huang. Diff-erank: A novel rank-based metric for evaluating large language models. *Advances in Neural Information Processing Systems*, 37:39501–39521, 2024.

Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, 638(8051):769–778, 2025.

Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Weijian Deng, Jianfeng Zhang, and Bo An. Mano: Exploiting matrix norm for unsupervised accuracy estimation under distribution shifts. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2024.

Renchunzi Xie, Ambroise Odonnat, Vasilii Feofanov, Ievgen Redko, Jianfeng Zhang, and Bo An. Leveraging gradients for unsupervised accuracy estimation under distribution shift. *Transactions on Machine Learning Research Journal*, 2025.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

Yaodong Yu, Zitong Yang, Alexander Wei, Yi Ma, and Jacob Steinhardt. Predicting out-of-distribution error with the projection norm. *arXiv preprint arXiv:2202.05834*, 2022.

Yueyao Yu and Yin Zhang. Why" classic" transformers are shallow and how to make them go deep. *arXiv preprint arXiv:2312.06182*, 2023.

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

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

A Related Works

Leveraging loss values to predict model performance. Many research areas use loss values as a proxy for model performance. For example, scaling-law studies in both pre-training (Kaplan et al., 2020; Hoffmann et al., 2022; Grattafiori et al., 2024) and downstream settings (Brandfonbrener et al., 2024) examine how model size, dataset size, and training compute scale with model performance, often relying on cross-entropy loss to estimate performance under the assumption that lower loss indicates better performance. Similarly, research on optimization methods commonly assumes that lower loss reflects better model performance when evaluating optimization efficiency (Yu & Zhang, 2023; Chen et al., 2026). This assumption is also used in theoretical analyses, where loss is often adopted as a continuous proxy for model performance (Xie et al., 2025). However, limited research has examined whether loss can serve as a reliable proxy for model performance.

Evaluation of large language models (LLMs). Evaluation of LLMs is a rapidly evolving field spanning a wide range of downstream tasks. The most direct and reliable approach to assessing LLM performance involves collecting or utilizing task-specific datasets with human annotations (Bisk et al., 2020; Lin et al., 2022; Cobbe et al., 2021; Mihaylov et al., 2018) and applying supervised evaluation metrics such as accuracy, F1 score (Christen et al., 2023), BLEU (Papineni et al., 2002), or ROUGE (Lin, 2004). However, obtaining high-quality ground-truth labels for downstream tasks is often time-consuming, expensive, and labor-intensive in practice. Consequently, there has been growing interest in developing unsupervised or label-free methods to estimate model performance. While several recent studies have explored unsupervised performance estimation for LLMs, the area remains relatively underexplored. For example, Diff-eRank (Wei et al., 2024) estimates model quality by measuring "noise reduction" through the effective rank (eRank) (Roy & Vetterli, 2007) of internal representations, comparing the current model to its initialization. However, this approach requires access to intermediate features and is therefore difficult to apply to many commercial or black-box LLMs. A more accessible line of work relies only on model outputs, using logits or token probabilities to construct uncertainty-based scores. A representative example is Ranked from Within (Tu et al., 2025), which proposes metrics based on token-level entropy and average negative log-likelihood. While simple and useful, these token-level measures summarize uncertainty within a single output sequence and do not leverage the comparative structure available across multiple candidate answers. As a result, they can be unstable for long-form generations and may not directly reflect whether a model can reliably distinguish correct answers from plausible alternatives, especially in complex reasoning or open-ended tasks.

Unsupervised performance estimation. Prior to the rise of LLMs, research on unsupervised performance estimation focused primarily on traditional deep learning models, such as ResNet families (He et al., 2016), to predict classification accuracy on specific test datasets using only unlabeled test samples. To provide a comprehensive overview of this field, we introduce the two main existing settings relevant to this topic. The first prominent setting is the **data-centric** paradigm, which focuses on estimating a model's classification accuracy across different test datasets, including scenarios involving distribution shifts between training and test data. The primary research direction in unsupervised performance estimation for traditional models focuses on exploring the negative relationship between distribution discrepancy and model performance from the space of features (Deng & Zheng, 2021), parameters (Yu et al., 2022; Xie et al., 2025) and labels (Lu et al., 2023). Another widely adopted approach involves designing estimation scores based on the softmax outputs of test samples (Guillory et al., 2021; Jiang et al., 2021; Garg et al., 2022; Xie et al., 2024). These methods rely heavily on the assumption that the model is well-calibrated. Some works draw inspiration from unsupervised learning techniques, such as agreement across multiple classifiers (Jiang et al., 2021; Madani et al., 2004; Platanios et al., 2016; 2017) or auxiliary prediction tasks like image rotation prediction (Deng et al., 2021). However, these metrics are primarily designed to evaluate the quality or difficulty of datasets, and therefore cannot be used to select the optimal model architecture from a large pool of candidate models.

B Preliminaries in the unsupervised setting (Section 3.1)

In this section, we consider a more practical setting in which ground-truth labels for MCQA datasets are unavailable, since annotation information is often costly to obtain. Specifically, we replace

supervised cross-entropy loss with an unsupervised counterpart, computed as the next-token cross-entropy of the question context for each MCQA instance:

$$\mathbb{E}_{x\sim\mathcal{D}}[l(f_m(x_{<t}), x_t)] = -\frac{1}{|T_x|}\sum_{t=1}^{|T_x|} x_t \log P_{f_m}(x_t|x_{<t}),$$

where x_t denotes the t-th token of the question context, and $|T_x|$ denotes the number of tokens in the question context.

Similarly, we continue to use the loss-based accuracy defined in Eq. 1 as the ground-truth measure of model performance. **This section aims to examine the relationship between loss-based accuracy and unsupervised cross-entropy loss.**

C Experimental setup of Section5

Besides directly using loss values, we also explore the correlation between the separability of loss values across multiple choices and the loss-based accuracy. Specifically, this separability is measured by the standard deviation of the auto-regressive cross-entropy losses across multiple choices, which can be expressed as:

$$\mathcal{S}(f_m) = \frac{1}{N}\sum_{i=1}^{N} std(\mathcal{L}_A^{i,f_m}, \mathcal{L}_B^{i,f_m}, .., \mathcal{L}_K^{i,f_m}), \tag{2}$$

where $\mathcal{L}_k^{i,f_m} = -\sum_{t=1}^{|\gamma_k|} \gamma_{k,t}^i \log P_{f_m}(\gamma_{k,t}^i|x_i, \gamma_{k,<t})$, denoting the autoregressive cross-entropy loss of the k^{th} choice for the question x_i. It is worth noting that we only consider option-relevant information to formulate the score.

D More numerical results

MCQA datasets. To evaluate the effectiveness of loss values in predicting the accuracy of LLMs under the multiple-choice setting, we use the validation sets of OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2020), MMLU (Hendrycks et al., 2021), and HellaSwag (Zellers et al., 2019) as benchmark datasets. OpenBookQA includes 500 validation questions with 4 candidate answers per question; PIQA contains 1,838 validation questions with 2 candidate answers per question; MMLU contains 1,531 validation questions across 57 subjects with 4 candidate answers per question; and HellaSwag includes 10,042 validation examples with 4 candidates per example.

Evaluation metrics To measure the linear relationship between model performance and proposed scores, we use the absolute values of coefficients of determination (R^2) and Spearman correlation coefficients (ρ) as the evaluation metrics. R^2 reflects predictive power of the scores, while ρ evaluates their ability to recover the true performance ranking.

In Table 3, 4, 5, 6, 7, 8, we provide the detailed results about the loss-based accuracy (Acc.) and supervised cross-entropy loss (SL) across diverse model families and datasets. In Table 9, 10, 11, 12, 13, 14, we provide the detailed numerical results about the loss-based accuracy (Acc.) and unsupervised cross-entropy loss (QL) for different model families and datasets.

E Proofs

E.1 Upper bound of prediction error

Proof. We define the model's predicted answer is $\hat{y}(x;\theta) = \arg\min_k L_k(\theta;x)$, and its implied probability over candidates is the softmax of the negative losses $p_k(\theta;x) = \frac{\exp(-L_k(\theta;x))}{\sum_{j=1}^{K}\exp(-L_j(\theta;x))}$. Thus, we have:

$$\text{Error}(\theta) = 1 - p^*(\theta;x) = 1 - \frac{exp(-L^*(\theta;x))}{\sum_{k=1}^{K} exp(-L_k(\theta;x))} = \frac{\sum_{k\neq *}\exp(-\delta_k)}{1+\sum_{k\neq *}\exp(-\delta_k)},$$

Table 3: Loss-based accuracy (Acc.) and supervised cross-entropy loss (SL) for the OPT model family across evaluation datasets.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	SL	Acc.	SL	Acc.	SL	Acc.	SL
OPT-125M	0.2760	2.2051	0.6208	0.6958	0.2560	1.8115	0.2921	1.4072
OPT-350M	0.2560	2.1641	0.6464	0.6807	0.2743	1.8008	0.3440	1.3467
OPT-1.3B	0.3300	1.9688	0.7122	0.6445	0.3070	1.6953	0.5040	1.1768
OPT-2.7B	0.3700	1.9404	0.7345	0.6274	0.3122	1.6748	0.5718	1.1104
OPT-6.7B	0.3600	1.9102	0.7557	0.6108	0.3246	1.6504	0.6409	1.0391
OPT-13B	0.3640	1.8828	0.7688	0.6074	0.3194	1.6465	0.6686	1.0127

Table 4: Loss-based accuracy (Acc.) and supervised cross-entropy loss (SL) for the Cerebras-GPT model family.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	SL	Acc.	SL	Acc.	SL	Acc.	SL
Cerebras-GPT-111M	0.2160	2.3262	0.5871	0.7207	0.2619	1.8506	0.2539	1.4785
Cerebras-GPT-256M	0.2380	2.2812	0.6061	0.7090	0.2724	1.8105	0.2688	1.4424
Cerebras-GPT-590M	0.2560	2.1797	0.6284	0.6978	0.2717	1.8027	0.3003	1.4033
Cerebras-GPT-1.3B	0.2780	2.1211	0.6643	0.6729	0.2796	1.7617	0.3598	1.3291
Cerebras-GPT-2.7B	0.3000	2.0527	0.7078	0.6528	0.2972	1.7139	0.4564	1.2266
Cerebras-GPT-6.7B	0.3260	1.9707	0.7291	0.6313	0.3181	1.6982	0.5535	1.1318
Cerebras-GPT-13B	0.3540	1.9355	0.7225	0.6348	0.3161	1.7012	0.5608	1.1221

where for each incorrect candidate $k \neq *$, $\delta_k = L_k(\theta; x) - L^*$.

Let $\Delta = \min_{k \neq *} \delta_k$, since every $\delta_k \geq \Delta$, we have $\exp(-\delta_k) \leq \exp(-\Delta)$. Therefore, $\sum_{k \neq *} \exp(-\delta_k) \leq (K-1)\exp(-\Delta)$. Thus, we have:

$$1 - p^*(\theta; x) = \frac{\sum_{k \neq *} \exp(-\delta_k)}{1 + \sum_{k \neq *} \exp(-\delta_k)} \leq \sum_{k \neq *} \exp(-\delta_k) \leq (K-1)\exp(-\Delta).$$

Lemma 1. *Defining* $\sigma(x) = \mathrm{std}(\mathbf{L}) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(L_k - \bar{L})^2}$, $\bar{L} = \frac{1}{K}\sum_k L_k$. *For any fixed value of* $\sigma(x)$, *the error probability* $1 - p^*(\theta; x)$ *is maximum when* $\delta_1 = \delta_2 = \cdots = \delta_{K-1} = \Delta$.

Proof. As $1 - p^*(\theta; x) = \frac{\sum_{k \neq *} \exp(-\delta_k)}{1 + \sum_{k \neq *} \exp(-\delta_k)}$, and is strictly increasing function of $s = \sum_{k \neq *} exp(-\delta_k)$. Therefore, showing that s is larger when the δ_k are unequal immediately proves the claim. □

As $(exp(-\delta_k))' = -exp(-\delta_k) < 0$, it shows the function $exp(-\delta_k)$ is strictly decreasing. As $(exp(-\delta_k))'' = exp(-\delta_k) > 0$, it shows that the function $exp(-\delta_k)$ is strictly convex. Therefore, when $\delta_k = \Delta = \min_{k \neq *} \delta_k$, $k = 1, 2, ..., K-1$, s obtains the largest value, so that $1 - p^*(\theta; x)$ obtains the largest value.

When we consider the worst-case configuration for a fixed question x, given the correct answer has loss $L^* = \min_k L_k(\theta; x)$, and all $K-1$ incorrect answers have the same loss $L_k = L^* + \Delta$, we have:

Table 5: Loss-based accuracy (Acc.) and supervised cross-entropy loss (SL) for the DeepSeek-R1-Distill model family.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	SL	Acc.	SL	Acc.	SL	Acc.	SL
DeepSeek-R1-Distill-Qwen-1.5B	0.2700	2.4212	0.5098	0.6919	0.2698	1.8798	0.3014	1.2622
DeepSeek-R1-Distill-Qwen-7B	0.2840	2.3550	0.5773	0.6515	0.2959	1.8184	0.4121	1.0835
DeepSeek-R1-Distill-Llama-8B	0.3120	1.9043	0.5816	0.5791	0.2796	1.7431	0.4967	0.9073
DeepSeek-R1-Distill-Qwen-14B	0.3780	2.0125	0.6828	0.5764	0.3403	1.6470	0.6371	0.8319
DeepSeek-R1-Distill-Qwen-32B	0.4200	1.7915	0.7062	0.5666	0.3703	1.5166	0.7062	0.8236
DeepSeek-R1-Distill-Llama-70B	0.3540	1.7479	0.6464	0.5592	0.3377	1.6784	0.6407	0.8048

Table 6: Loss-based accuracy (Acc.) and supervised cross-entropy loss (SL) for the Llama model family.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	SL	Acc.	SL	Acc.	SL	Acc.	SL
Llama-3.2-3B	0.4100	1.7939	0.7737	0.5947	0.3684	1.5493	0.7159	0.9643
Llama-3.2-3B-Instruct	0.4000	1.8342	0.7606	0.6100	0.3756	1.6124	0.6872	1.0019
Llama-3.1-8B	0.4380	1.6690	0.8030	0.5737	0.3815	1.5075	0.7724	0.9020
Llama-3.1-8B-Instruct	0.4580	1.6702	0.7954	0.5761	0.3958	1.5045	0.7589	0.9025
Llama-2-7b-chat-hf	0.4720	1.4125	0.7655	0.5922	0.3703	1.4741	0.7167	0.9586

$$\mathrm{Var}(\mathbf{L}) = \frac{1}{K}\left[\Big(L^* - \bar{L}\Big)^2 + (K-1)\Big((L^* + \Delta) - \bar{L}\Big)^2\right]$$

$$= \frac{1}{K}\left[(-\frac{K-1}{K}\Delta)^2 + (K-1)(\frac{1}{K}\Delta)^2\right]$$

$$= \frac{K-1}{K^2}\Delta^2.$$

Therefore, we have:

$$\sigma(x) = \frac{\sqrt{K-1}}{K}\Delta \quad \Longrightarrow \quad \Delta = \frac{K}{\sqrt{K-1}}\,\sigma(x).$$

Finally, we have proved that:

$$\mathrm{Error}(\theta) \leq (K-1)\exp(-\frac{K}{\sqrt{K-1}}\sigma(x)).$$

□

Table 7: Loss-based accuracy and supervised cross-entropy loss for the Qwen model family.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	SL	Acc.	SL	Acc.	SL	Acc.	SL
Qwen1.5-7B	0.4080	1.9912	0.7820	0.5925	0.3490	1.6163	0.7400	0.9273
Qwen2.5-1.5B-Instruct	0.3880	1.9531	0.7420	0.6145	0.3680	1.6165	0.6460	1.0346
Qwen2.5-3B	0.4160	1.8777	0.7700	0.6035	0.3640	1.5630	0.6980	0.9759
Qwen2.5-3B-Instruct	0.4080	1.8911	0.7790	0.5938	0.3830	1.5960	0.7120	0.9447
Qwen2.5-7B	0.4380	1.7621	0.7850	0.5876	0.3850	1.5143	0.7540	0.9112
Qwen2.5-7B-Instruct	0.4460	1.7792	0.7920	0.5723	0.3950	1.5462	0.7750	0.8388
Qwen2-7B-Instruct	0.4180	1.8910	0.7980	0.5739	0.4040	1.4766	0.7690	0.8787
Qwen3-14B	0.4080	1.8682	0.7900	0.5713	0.4090	1.5654	0.7580	0.8618
Qwen3-4B	0.3820	2.0918	0.7460	0.6021	0.3680	1.6768	0.6520	1.0000
Qwen3-4B-Instruct-2507	0.4040	2.0645	0.7510	0.6045	0.3740	1.7441	0.6660	0.9941
Qwen3-8B	0.4040	2.0293	0.7830	0.5811	0.3980	1.6113	0.7100	0.9292

Table 8: Loss-based accuracy and supervised cross-entropy loss for the OpenELM model family.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	SL	Acc.	SL	Acc.	SL	Acc.	SL
OpenELM-270M	0.3160	2.1270	0.6893	0.7593	0.2685	1.6553	0.4375	1.5410
OpenELM-450M	0.3400	2.0332	0.7149	0.7422	0.2894	1.6113	0.5059	1.4893
OpenELM-1.1B	0.3800	1.9473	0.7443	0.7456	0.3076	1.6338	0.6105	1.4717
OpenELM-270M-Instruct	0.3460	2.1055	0.6991	0.7510	0.2789	1.6846	0.4897	1.5361
OpenELM-450M-Instruct	0.3680	1.9395	0.7252	0.7593	0.3057	1.7207	0.5606	1.5078
OpenELM-1.1B-Instruct	0.4140	1.9473	0.7421	0.7451	0.3344	1.6836	0.6737	1.4932

Table 9: Loss-based accuracy and unsupervised cross-entropy loss for the OPT model family across evaluation datasets.

Model	**OpenBookQA**		**PIQA**		**MMLU**		**HellaSwag**	
	Acc.	**QL**	**Acc.**	**QL**	**Acc.**	**QL**	**Acc.**	**QL**
OPT-125M	0.2760	4.9765	0.6208	5.4614	0.2560	4.1889	0.2921	3.8128
OPT-350M	0.2560	4.7746	0.6464	5.2320	0.2743	4.0202	0.3440	3.6836
OPT-1.3B	0.3300	4.5956	0.7122	5.0230	0.3070	3.6607	0.5040	3.4512
OPT-2.7B	0.3700	4.5491	0.7345	4.9627	0.3122	3.5583	0.5718	3.3457
OPT-6.7B	0.3600	4.5128	0.7557	4.9053	0.3246	3.4354	0.6409	3.2663
OPT-13B	0.3640	4.5778	0.7688	5.0156	0.3194	3.3923	0.6686	3.2547

Table 10: Loss-based accuracy and unsupervised cross-entropy loss for the Cerebras-GPT model family across evaluation datasets.

Model	**OpenBookQA**		**PIQA**		**MMLU**		**HellaSwag**	
	Acc.	**QL**	**Acc.**	**QL**	**Acc.**	**QL**	**Acc.**	**QL**
Cerebras-GPT-111M	0.2160	5.4000	0.5871	5.7601	0.2619	4.3949	0.2539	4.4127
Cerebras-GPT-256M	0.2380	5.1946	0.6061	5.5296	0.2724	4.1001	0.2688	4.1168
Cerebras-GPT-590M	0.2560	5.0566	0.6284	5.4317	0.2717	3.8806	0.3003	3.9555
Cerebras-GPT-1.3B	0.2780	4.8603	0.6643	5.1230	0.2796	3.6551	0.3598	3.7028
Cerebras-GPT-2.7B	0.3000	4.7970	0.7078	5.0546	0.2972	3.5138	0.4564	3.5805
Cerebras-GPT-6.7B	0.3260	4.6353	0.7291	4.9040	0.3181	3.3102	0.5535	3.4541
Cerebras-GPT-13B	0.3540	4.6371	0.7225	4.9525	0.3161	3.3125	0.5608	3.4520

Table 11: Loss-based accuracy and unsupervised cross-entropy loss for the DeepSeek model family across evaluation datasets.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	QL	Acc.	QL	Acc.	QL	Acc.	QL
DeepSeek-Qwen-1.5B	0.2700	8.0135	0.5098	8.8832	0.2698	6.2103	0.3014	6.8589
DeepSeek-Qwen-7B	0.2840	7.5827	0.5773	8.4982	0.2959	5.2033	0.4121	5.8739
DeepSeek-Llama-8B	0.3120	7.4048	0.5816	7.8965	0.2796	5.0468	0.4967	5.1834
DeepSeek-Qwen-14B	0.3780	6.0171	0.6828	7.0807	0.3403	3.7791	0.6371	4.4912
DeepSeek-Qwen-32B	0.4200	5.4812	0.7062	6.2385	0.3703	3.2525	0.7062	3.7818
DeepSeek-Llama-70B	0.3540	7.5019	0.6464	8.0634	0.3377	4.6167	0.6407	4.5818

Table 12: Loss-based accuracy and unsupervised cross-entropy loss for the Qwen model family across evaluation datasets.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	QL	Acc.	QL	Acc.	QL	Acc.	QL
Qwen1.5-7B	0.4080	4.2830	0.7820	4.8010	0.3490	2.8550	0.7400	2.7020
Qwen2.5-1.5B-Inst.	0.3880	4.2930	0.7420	5.0050	0.3680	2.8960	0.6460	3.1910
Qwen2.5-3B	0.4160	4.1510	0.7700	4.7100	0.3640	2.7680	0.6980	3.0470
Qwen2.5-3B-Inst.	0.4080	4.3020	0.7790	4.8530	0.3830	2.8360	0.7120	3.1460
Qwen2.5-7B	0.4380	4.0740	0.7850	4.6450	0.3850	2.6790	0.7540	2.9860
Qwen2.5-7B-Inst.	0.4460	4.1770	0.7920	4.7200	0.3950	2.7540	0.7750	3.2410
Qwen2-7B-Inst.	0.4180	4.2140	0.7980	4.6850	0.4040	2.7600	0.7690	3.1040
Qwen3-14B	0.4080	4.8720	0.7900	5.6720	0.4090	2.9930	0.7580	3.2520
Qwen3-4B	0.3820	5.1210	0.7460	5.9760	0.3680	3.2440	0.6520	3.5700
Qwen3-4B-Inst.-2507	0.4040	4.3350	0.7510	5.0550	0.3740	2.9360	0.6660	3.3550
Qwen3-8B	0.4040	4.7670	0.7830	5.5650	0.3980	3.0430	0.7100	3.2560

Table 13: Loss-based accuracy and unsupervised cross-entropy loss for the Llama model family across evaluation datasets.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	QL	Acc.	QL	Acc.	QL	Acc.	QL
Llama-3.2-3B	0.4100	4.4342	0.7737	5.1370	0.3684	3.1322	0.7159	3.1615
Llama-3.2-3B-Inst.	0.4000	4.6749	0.7606	5.4031	0.3756	3.3366	0.6872	3.4401
Llama-3.1-8B	0.4380	4.4135	0.8030	5.0981	0.3815	3.0119	0.7724	3.0119
Llama-3.1-8B-Inst.	0.4580	4.7147	0.7954	5.4160	0.3958	3.1476	0.7589	3.2967
Llama-2-7B-chat	0.4720	4.7192	0.7655	4.7193	0.3703	3.1814	0.7167	3.4111

Table 14: Loss-based accuracy and unsupervised cross-entropy loss for the OpenELM model family across evaluation datasets.

Model	OpenBookQA		PIQA		MMLU		HellaSwag	
	Acc.	QL	Acc.	QL	Acc.	QL	Acc.	QL
OpenELM-270M	0.3160	4.2071	0.6893	4.4322	0.2685	3.2675	0.4375	3.1544
OpenELM-450M	0.3400	4.1428	0.7149	4.3675	0.2894	3.1677	0.5059	3.0756
OpenELM-1.1B	0.3800	4.1001	0.7443	4.3099	0.3076	3.0145	0.6105	2.9246
OpenELM-270M-Inst.	0.3460	4.4032	0.6991	4.5669	0.2789	3.4656	0.4897	3.4830
OpenELM-450M-Inst.	0.3680	4.3070	0.7252	4.5208	0.3057	3.3343	0.5606	3.3704
OpenELM-1.1B-Inst.	0.4140	4.3756	0.7421	4.5267	0.3344	3.3347	0.6737	3.5811