A Appendix

A.1 Pretrained Large Language Models

Neural autoregressive language model (LMs) are designed for next token prediction to predict the probability distribution over the next token after a sequence of tokens input, and pre-trained LMs show their superior performance since they are trained on various programming languages and a large-scale curated dataset. Training large natural LMs are very expansive and time-consuming process since they always have billions of parameters, which limits the development of LMs. Fortunately, many pre-trained LMs are open access or limited access, which promotes researchers to pool their time and makes the resources to collectively achieve a higher impact. EleutherAI makes the GPT-J [22] and GPT-Neox [23] public available on Hugging Face. GPT-3 [1] is limited access in OpenAI which can be used by researchers for a fee, and another large open-science open-access multilingual language model named Bloom [2] is provided by BigScience.

A.2 Open Access Models

Table 4: Pretrained language models

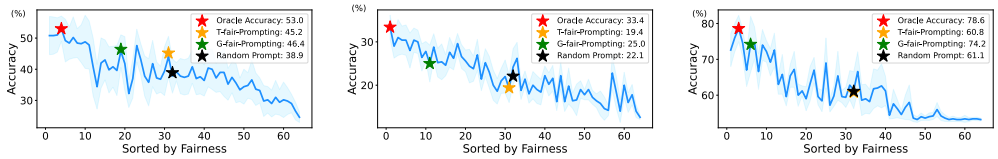
Model	Params	Provider	Access
GPT-2	124 M	Hugging Face	OPEN
GPT-Medium	335 M	Hugging Face	OPEN
GPT2-Large	774 M	Hugging Face	OPEN
GPT-XL	1.5 B	Hugging Face	OPEN
GPT-3 (ada)	350 M	OPENAI	LIMITED
GPT-3 (babbage)	1.3 B	OPENAI	LIMITED
GPT-3 (curie)	6.7 B	OPENAI	LIMITED
GPT-3 (davinci)	175 B	OPENAI	LIMITED
GPT-J	6 B	EleutherAI	OPEN
GPT-NeoX	20 B	EleutherAI	OPEN
Bloom	176 B	BigScience	OPEN
LLaMA	7 B	Meta	OPEN
	13 B	Meta	OPEN
	33 B	Meta	OPEN
	65 B	Meta	OPEN

A.3 Additional Figures on Different Settings

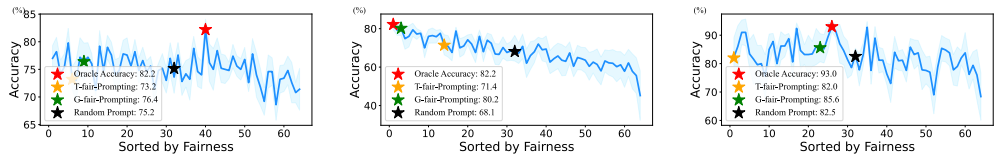
In additional to the Fig. 2, we shows the performance on different models for enumerating all candidates, note that the shadow indicates the half value of standard deviation for clear presentation since the variance is very high for LLMs.

A.4 Accuracy Varies with demonstrations

Accuracy Varies with Example Amount Demonstrations play an important role in imparting task-related information to language models through in-context learning. Then, the question arises - does a larger number of demonstrations necessarily equate to better performance? To answer this question, we evaluated performance in terms of accuracy by gradually increasing the number of demonstrations. We set $\rho = \Gamma(x_1, y_1) \oplus \cdots \oplus \Gamma(x_k, y_k)$, where $k = 1, \cdots, n$, and demonstrations are erased with k decreasing from n to 1. Intuitively, accuracy would vary highly across different numbers of demonstrations, and the phenomenon is observed in Fig. 6a. To our surprise, however, erasing some demonstrations can result in a better performance. Removing some demonstrations can perform better and sometimes GPT-3 achieves best accuracy when there is only a few demonstrations remaining. This highlights the importance of considering the appropriate number of demonstrations.

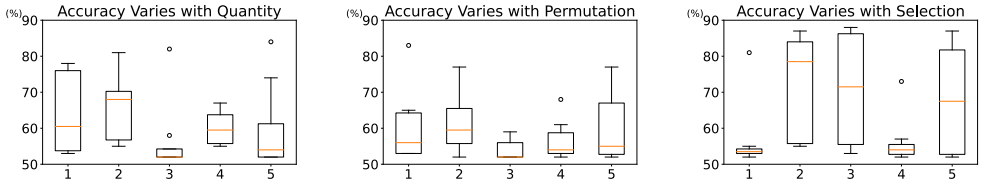


(a) AGNews (GPT2-XL 1.5B) (b) TREC (GPT2-XL 1.5B) (c) RTE (GPT2-XL 1.5B)



(d) AGNews (LLaMA 33B) (e) TREC (LLaMA 33B) (f) SST-2 (LLaMA 33B)

Figure 5: Accuracy is highly consistency with fairness and greedy search can find a good prompt, where "Random" and "Oracle" indicates the average accuracy of all prompts and the upper-bound performance according to fairness.



(a) Varying amount of examples (b) Permutation (c) Select different examples

Figure 6: ICL suffers from high instability due to variations in example amount, example order, and example selection.

Example Order The performance of a model is sensitive to the order of the demonstrations, as has been discussed in [4]. Even when the demonstrations are the same, different permutations of the demonstrations can result in vastly different outcomes. As there are $n!$ possible permutations, we introducing a strategy of permuting the demonstrations by circularly shifting the index of the demonstrations. The demonstration can be represented as $\rho = \Gamma(x_{k+1}, y_{k+1}) \oplus \cdots \oplus \Gamma(x_n, y_n) \oplus \Gamma(x_1, y_1) \oplus \cdots \oplus \Gamma(x_k, y_k)$.As shown in Fig. 6b, the accuracy varies highly with permutation which consistent with the observations in [4].

Example Selection In this paper, we find which demonstrations are selected is influence the model extremely. This scenario can be described as selecting k demonstrations in n training samples. In Fig. 6c, we only select one example for demonstration to ablate the impact of demonstrations order, and the accuracy also varies highly with different example selected. In this work, we only detail evaluate the proposed probing method on the erasing demonstrations and permutation, although our method improves by 20% in the setting of example selection on SST-2 (GPT2-XL), because selecting k demonstrations on a set with n training samples can't be regarded as $k-$shot learning in the strict sense.

A.5 Relationship between with- and without-calibration

• **G-fair-Prompting without post-calibration outperforms random demonstrations after post-calibration.** Based on Table 2, it is apparent that G-fair-Prompting outperforms random selection prior to post-calibration. This leads to a natural question: do prompts with better performance before calibration also indicate better performance after calibration proposed by Zhao et al. [18]? To investigate the relationship between performance with- and without-calibration, we calculated the Pearson correlation coefficient between the accuracy with- and without-calibration $Pearson(acc_{w/o}, acc_{with})$. A positive coefficient value suggests that a prompt with high accuracy before calibration has a

Table 5: Accuracy for different prompting strategies (averaged on $5_{0,\cdots,4}$ different seeds).

Model	Dataset	Random	Diversity	Similarity	Ours Top-2	Ours Top-4	Ours Greedy
GPT2-XL (1.5B)	SST-2	$61.1_{6.1}$	–	–	$60.8_{11.4}$	$65.8_{8.7}$	$74.2_{12.0}$
	AGNews	$38.9_{11.4}$	–	–	$45.2_{12.5}$	$37.2_{11.2}$	$46.4_{11.9}$
	TREC	$22.1_{5.7}$	–	–	$19.4_{8.9}$	$28.2_{9.2}$	$25.0_{7.4}$
	RTE	$53.2_{6.9}$	–	–	$54.0_{7.5}$	$53.6_{5.9}$	$56.4_{2.2}$
LLaMA (7B)	AGNews	$64.5_{10.0}$	$66.4_{9.1}$	–	$66.0_{11.7}$	$69.2_{5.5}$	$63.8_{5.7}$
	TREC	$49.5_{10.4}$	$51.4_{9.6}$	–	$48.4_{10.5}$	$38.6_{15.2}$	$61.3_{4.8}$
	CoLA	$60.4_{10.6}$	$63.8_{8.7}$	–	$58.2_{7.8}$	$61.6_{6.5}$	$36.4_{3.6}$
LLaMA (13B)	AGNews	$72.2_{7.7}$	$78.4_{3.5}$	–	$73.6_{9.0}$	$74.2_{4.3}$	$75.2_{2.8}$
	TREC	$46.4_{16.5}$	$48.0_{16.0}$	–	$51.0_{16.6}$	$39.2_{23.3}$	$61.4_{12.1}$
	CoLA	$67.7_{2.9}$	$67.2_{2.4}$	–	$67.0_{2.0}$	$67.2_{1.6}$	$67.0_{2.0}$

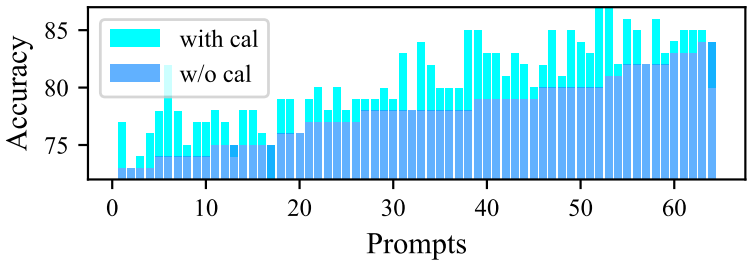


Figure 7: Illustration of accuracy relationship between with- and without calibration when $Pearson(\cdot)$ is positive.

higher likelihood of achieving higher accuracy after calibration than other prompts. We take the topic classification task on LLaMA(65B) for illustration to show the relationship between with- and without calibration when $Pearson$ is positive in Fig.7. Table 6 presents the Pearson correlation coefficient on accuracy of permutation and G-fair-Prompting after calibration. The majority of Pearson correlation coefficients were found to be positive, indicating that prompts with better performance before calibration have more potential to perform well after calibration. Furthermore, our results on the LLaMA family reveal that the larger the model, the stronger the correlation between performance with- and without-calibration. For instance, the value of the Pearson correlation coefficient increases from 0 to 0.7 as the model size increases.

Theorem A.1. *Suppose the performance of the model under certain prompts with- and without-calibration is positively correlated, i.e.,* $Pearson(acc_{w/o}, acc_{with}) > 0$*, if we can assure* $\mathbb{E}(acc_{w/o}^{Selected}) > \mathbb{E}(acc_{w/o}^{Random})$*, then we have* $\mathbb{E}(acc_{with}^{Selected}) > \mathbb{E}(acc_{with}^{Random})$.

Table 6: Pearson's r between the with- and without-calibration.

Dataset	BLOOM 176B	LLaMA 7B	LLaMA 13B	LLaMA 33B	LLaMA 65B
TREC	0.1274	0.1551	0.2959	0.3090	0.5151
AGNews	0.3875	−0.0471	0.3044	0.6953	0.7100
CoLA	0.4050	0.3592	0.5193	0.3611	0.8012

As analysed in Theorem A.1, if we can find a prompt with high accuracy before calibration, we have a higher likelihood of achieving higher accuracy after calibration than random selection. Our approach consistently identifies an appropriate prompt, as evidenced by the results in Table 2. Moreover, the

performance of the model exhibits a positive correlation with and without calibration under certain prompts, as illustrated in Table 6. Therefore, our method is more likely to enhance calibration performance.

A.6 Diversity and Similarity

Diversity indicates the demonstrations are selected according to the diversity of the embeddings of all samples [12]. Specifically, we select the 4 most diverse samples as demonstrations by k-means clustering on training set consists of 16 samples.

Similarity indicates the demonstrations are selected according to the similarity of the embeddings of all samples [15]. Specifically, we select the 4 most similar samples as demonstrations according to Euclidean distance from the training set consists of 16 samples. Note that demonstrations for different test samples are different for best performance. A larger training set may result in a better performance, but we set the size of training set as 16 for a fair comparison.

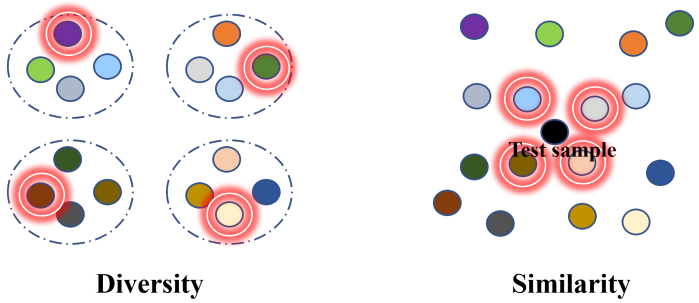


Figure 8: Illustration for Diversity and Similarity.

A.7 Details for experiments

In Fig. 1, we show the high instability due to high variations in demonstrations selection and order for both with- and without-calibration [18]. Specifically, for left two figures, we sample 4 different demonstrations randomly from dataset AGNews, and estimate the influence of demonstration selection. On the other hand, the right two figures show the instable performance due to permutation when the demonstrations are fixed. Specifically, we randomly sample 4 demonstrations and estimate the performance with all possible orders.

A.8 Complexity of different strategies

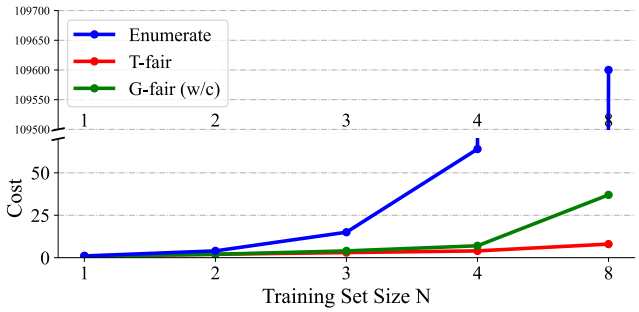


Figure 9: Computational cost. T-fair and G-fair indicate T-fair-Prompting and G-fair-Prompting respectively, and "w/c" indicates the worst case.

A.9 Limitations.

While the proposed strategy does not require modification of the original inference process of the Language Model (LM), it still necessitates the logits or probabilities of the next token. As a result, it may be necessary to approximate the probabilities in some completely black-box LM services, such as GPT-4.