Beyond Generation: Leveraging LLM Creativity to Overcome Label Bias in Classification

Xiaoyue Wang and **Xin Liu**
School of Informatics, Xiamen University, China
xiaoyuewang@stu.xmu.edu.cn

Abstract

Large Language Models (LLMs) exhibit impressive capabilities in In-Context Learning (ICL) but are prone to label bias—an undesirable tendency to favor certain answers. Existing calibration methods mitigate bias by leveraging in-domain data, yet such data is often unavailable in real-world scenarios. To address this limitation, we propose SDC (Synthetic Data Calibration), a simple-yet-effective approach that generates synthetic in-domain data from a few in-context demonstrations and utilizes it for calibration. By approximating the benefits of real in-domain data, SDC effectively reduces label bias without requiring access to actual domain-specific inputs. Experimental evaluations on 279 classification and multiple-choice tasks from the SUPER-NATURALINSTRUCTIONS benchmark. The results show that SDC significantly reduces label bias, achieving an average Bias Score reduction of 57.5%, and outperforming all competitive baselines. Moreover, when combined with Leave-One-Out Calibration (LOOC), SDC further improves performance, underscoring its effectiveness and generalizability in enhancing the reliability of LLMs.

1 Introduction

Large Language Models (LLMs) demonstrate impressive capabilities in handling unseen tasks by conditioning on examples of input-output pairs, known as In-Context Learning (ICL) demonstrations. However, recent research reveals that LLMs' predictions exhibit *Label Bias* (Zhao et al., 2021; Chen et al., 2023, 2024), an undesirable tendency to favor certain answers. This phenomenon is influenced by the label distribution in the demonstrations (Min et al., 2022), or by the order of them (Lu et al., 2022; Zheng et al., 2023). Such a bias undermines the reliability of LLM predictions and limits their practical applications, particularly in fields demanding high reliability, i.e finance.

To address label bias, several calibration-based methods have been proposed, each using progressively more information from the target task's input. Contextual Calibration (CC) (Zhao et al., 2021) uses little to no domain-relevant input, instead feeding tokens like `N/A` to estimate and correct for the model's prior predictions. Domain-Context Calibration (DCC) (Fei et al., 2023) refines this idea by sampling random texts directly from the in-domain input, thereby capturing more domain-specific signals in the calibration process. More recently, Leave-One-Out Calibration (LOOC) (Reif and Schwartz, 2024) removes each demonstration in turn to compute a more precise bias estimation, effectively harnessing the original demonstration inputs themselves. Although each of these methods reduces label bias, they also reveal that additional, task-related text (domain, random samples, or full demonstrations) can significantly improve calibration quality.

Motivated by these trends, we first performed a preliminary investigation into how real in-domain inputs help estimate a better prior for calibration. As expected, when in-domain data is available, it yields remarkably accurate estimates of the model's tendency to favor certain labels. However, real in-domain inputs are often unavailable in real-world ICL scenarios, where the model faces entirely unseen tasks with only a handful of example demonstrations. Leveraging the strong generative capabilities of LLMs, we propose using the model itself to create synthetic in-domain data. In this work, we develop **SDC**—Synthetic Data Calibration. SDC leverages LLMs to generate synthetic in-domain data from a few in-context demonstrations. This synthetic data is then used to calibrate model predictions, following the same approach as in our preliminary experiments. By doing so, SDC effectively mitigates label bias without requiring real in-domain data.

We compared the proposed method with

Findings of the Association for Computational Linguistics: ACL 2025, pages 25500–25506
July 27 - August 1, 2025

Method	$\mathbf{p}$	$\mathrm{diag}(\mathbf{p}_{dc})^{-1}\mathbf{p}$	$\mathrm{diag}(\mathbf{p}_{I})^{-1}\mathbf{p}$
Bias Score ↓	0.098	0.060	**0.029**
RSD ↓	0.562	0.385	**0.194**

Table 1: The bias evaluation results for the uncalibrated model predictions, as well as for the model predictions calibrated using $\mathbf{p}_{dc}$ and $\mathbf{p}_I$.

competitive baselines on 279 diverse classification and multiple-choice tasks from SUPER-NATURALINSTRUCTIONS (Wang et al., 2022) using two widely used LLMs: Llama3-7b (AI@Meta, 2024) and Qwen2-7b (Yang et al., 2024). The results show that SDC achieves the best performance among all comparisons, as evidenced by an average 57.5% reduction in Bias Score (Reif and Schwartz, 2024) on two models. Furthermore, when combining the label bias estimated by SDC with LOOC, the model's label bias is further mitigated, achieving state-of-the-art Micro-F1, strongly demonstrating the generalizability and effectiveness of SDC.

2 Preliminaries

Label Bias In-Context Learning (ICL) enables LLMs to solve unseen tasks by prompting them with several demonstrations. Let $\mathcal{C} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{|\mathcal{C}|}, y_{|\mathcal{C}|})\}$ denotes the demonstrations, where x_* and y_* are the input and output, respectively. The model is then expected to predict the answer y for the input x by feeding the concatenation of $\mathcal{C}$ and x into the model, formally: $y = \arg\max_{y \in Y} p(y|x, \mathcal{C})$, where $p(\cdot)$ denotes the probability predicted by the model, Y is the set of all possible output answers.

However, ICL has been shown to exhibit label bias, where the model displays an unexpected preference for certain answers. This bias can be influenced by the order of examples in $\mathcal{C}$ or the token frequency of answers encountered during the LLM's pretraining phase. In this work, we follow Reif and Schwartz (2024) to measure label bias and performance using three metrics: **Bias Score**, **Relative Standard Deviation** of class-wise accuracy (**RSD**), and **Micro-F1**. The first two capture how strongly the model favors certain classes, whereas Micro-F1 evaluates its overall classification performance. Formal definitions and detailed explanations of these metrics are available in Appendix A.1.

Progressive Use of Task Input in Previous Studies. Several calibration-based methods have been proposed to estimate and correct the model's prior preference over possible labels, each one exploiting progressively more domain-relevant input:

Contextual Calibration (CC) (Zhao et al., 2021) uses minimal domain information to estimate the prior, simply replacing the real input with a placeholder token (`N/A`). Formally, $p_{CC}(y^i) = p\big(y^i|[\texttt{N/A}], \mathcal{C}\big)$.

Domain-Context Calibration (DC) (Fei et al., 2023) samples text *from in-domain data* rather than using `N/A`, thus incorporating more task-related content. This process is described by $p_{dc}(y^i) = \frac{1}{|M|}\sum_{m=1}^{M} p\big(y^i|[\text{random text}]_m, \mathcal{C}\big)$, where M is the number of selected random text.

Leave-One-Out Calibration (LOOC) (Reif and Schwartz, 2024) goes further by exploiting the *original demonstration inputs* themselves. It excludes each (x, y) from $\mathcal{C}$ in turn, forms $\mathcal{C}_{-k}$, and computes the label-wise probability $p_{LOOC}(y^i)$ over these reduced contexts. Repeating for all labels yields the overall prior $\mathbf{p}_{LOOC}$.

In-Context Calibration (ICC) (Jang et al., 2024) extends LOOC by introducing perturbation-based priors. Specifically, it estimates label priors using demonstrations with shuffled tokens, aiming to decorrelate the input-label pairs. While ICC improves over LOOC, its calibration quality depends on the extent to which the perturbations remove spurious correlations without losing task semantics. We include ICC as a competitive baseline in our experiments.

In every case, the model's final output probabilities $\mathbf{p}$ are rescaled by $\mathrm{diag}(\mathbf{p}_*)^{-1}$, where $\mathbf{p}_*$ is the respective prior from one of the above approaches.

Mitigating Label Bias using In-Domain Data Inspired by the observation that richer domain-specific input often yields a more accurate prior, we examine an *idealized* scenario where complete in-domain data $\mathcal{X}^I = \{x_1^I, \ldots, x_{|\mathcal{X}^I|}^I\}$ is available. In this case, we directly average model predictions over *all* in-domain inputs:

$$p_t(y^i) = \frac{1}{|\mathcal{X}^I|} \sum_{x_j^I \in \mathcal{X}^I} p\big(y^i|x_j^I, \mathcal{C}\big). \qquad (1)$$

The estimated prior becomes $\mathbf{p}_I = \big[p_I(y^1), \ldots, p_I(y^{|Y|})\big]$, and we can obtain the calibrated model prediction $\mathrm{diag}(\mathbf{p}_I)^{-1}\mathbf{p}$.

Empirical Setup and Observations. We instantiate this scenario using the Llama3-7b model and evaluate on 279 classification and multiple-choice tasks from SUPER-NATURALINSTRUCTIONS (Wang et al., 2022). We compare the result with DC, where the estimate prior represented as $\mathbf{p}_{dc}$. Table 1 reports the average Bias Score and RSD for the uncalibrated model and for the calibrated predictions under both $\mathbf{p}_{dc}$ and $\mathbf{p}_{I}$. Notably, leveraging the full in-domain dataset (i.e., $\mathbf{p}_{I}$) leads to a marked reduction in Bias Score and RSD, confirming that richer domain content significantly improves the model's prior estimation. However, since complete in-domain data is often unavailable in real-world ICL, we next explore utilize the LLM to *generate* domain-relevant data for calibration.

3 SDC: Synthetic Data Calibration

Building on our findings that domain-relevant input greatly improves calibration (Section 2), we now address the more realistic setting in which *real in-domain data* is unavailable. We propose SDC, a method that leverages the strong generative capability of LLMs to *create* synthetic in-domain input from just a few demonstration examples.

The key intuition behind SDC is that LLMs, when prompted with demonstrations, can generate diverse synthetic data that capture essential patterns of the target domain. By calibrating predictions with this synthetic data, we approximate the benefits of real in-domain data without its availability. In SDC, the LLM is prompted with `In:`x_1`,Out:`y_1`...In:`$x_{|\mathcal{C}|}$`,Out:`$y_{|\mathcal{C}|}$`,In:` to generate synthetic data. [1] By sampling outputs from the model, we can collect a set of unlabeled synthetic in-domain data, $\mathcal{X}^s = \{x^s_1, \ldots, x^s_{|\mathcal{X}^s|}\}$. We then follow Eq. 1 to estimate the model's prediction prior.

$$p_s(y^i) = \frac{1}{|\mathcal{X}^s|} \sum_{x^s_i \in \mathcal{X}^s} p(y^i | x^s_i, \mathcal{C}), \qquad (2)$$

and calibrate the model prediction $\mathbf{p}$ via $\mathrm{diag}(\mathbf{p}_s)^{-1}\mathbf{p}$, where $\mathbf{p}_s$=$[p_s(y^1), ..., p_s(y^{|Y|})]$.

By doing this, SDC only need a few in-domain demonstrations serve merely as seeds to guide the LLM in generating synthetic data. Unlike methods that rely on in-domain input, these demonstrations enable the production of a diverse synthetic set that approximates domain characteristics and is used solely for prior estimation and calibration.

Metric	Llama3-7b	Qwen2-7b
Micro-F1 (↑)		
Original LM	0.562	0.579
CC	0.581	0.583
DC*	0.610	0.609
LOOC	0.654	0.662
ICC	0.660	–
SDC	0.663	0.667
SDC + LOOC	**0.668**	**0.674**
Bias Score (↓)		
Original LM	0.098	0.122
CC	0.081	0.128
DC*	0.060	0.109
LOOC	0.043	0.061
ICC	0.043	–
SDC	0.041	0.055
SDC + LOOC	**0.033**	**0.051**
RSD (↓)		
Original LM	0.562	0.506
CC	0.496	0.509
DC*	0.385	0.426
LOOC	0.275	0.259
ICC	0.271	–
SDC	0.257	0.234
SDC + LOOC	**0.227**	**0.228**

Table 2: The averaged results of SDC and the comparisons across 276 tasks from SUPER-NATURALINSTRUCTIONS. The best results are highlighted in **bold**, and the second best are underlined. SDC achieves the highest performance in improving task outcomes and mitigating label bias on both models. Additionally, combining SDC with LOOC further enhances task performance and reduces label bias. * indicates the method require the assess of in-domain data.

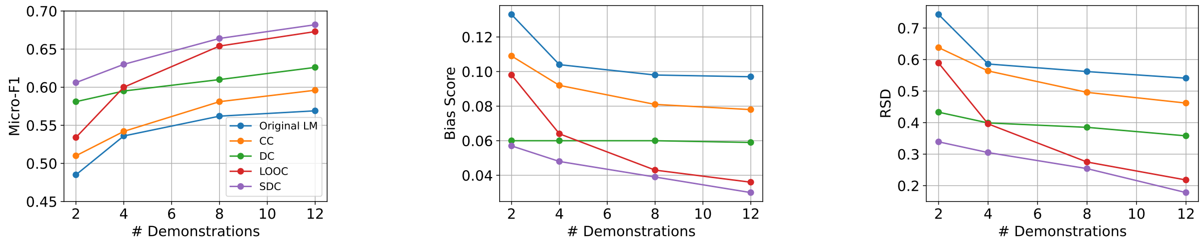
4 Experimental Settings

4.1 Datasets

We follow Reif and Schwartz (2024) to conduct experiments on 276 classification and multiple-choice tasks from the SUPER- NATURALINSTRUCTIONS benchmark (Wang et al., 2022). In this benchmark, there are 1,000 evaluation instances and an additional set of 32 held-out instances for estimating

---

[1] We try multiple strategies to construct the prompt, and this one performs the best. Results and discussion can be seen in Appendix A.2



(a) Micro-F1 (b) Bias Score (c) RSD

Figure 1: Results of SDC across various numbers of demonstrations.

the Bias Score. The possible labels for all tasks are predefined, such as “`Positive/Negative`” or “`Yes/No`”.

4.2 Implementation Details

We use Llama3-7b (AI@Meta, 2024) and Qwen2-7b (Yang et al., 2024) as the base models. For each task, we randomly sampled 8 instances as demonstrations for both generating synthetic in-domain inputs and evaluating models on each task. We apply Nucleus Sampling (Holtzman et al., 2020) with a threshold of p=0.85 to sample diverse synthetic in-domain inputs. For each task, 40 synthetic in-domain unlabeled instances are generated to estimate the model’s prior. We use greedy search when evaluating the model. Regarding DC, we also sample 40 random texts of the average input length, keeping the same number as the synthetic instances. We conduct all experiments 3 times and report the averaged results. Running the experiment once requires nearly one GPU hour on an RTX A6000.

5 Results and Analysis

5.1 Main Results

The results of SDC and baselines applied to two LLMs are shown in Table 2. All methods reduce label bias in the original models, as seen in higher Micro-F1 scores and lower Bias Scores and RSD. Notably, DC, which uses in-domain data for calibration, reduces Bias Score by an average of 24.7% across the two models compared to the original LMs. In contrast, SDC, which does not use in-domain data, significantly reduces Bias Score by an average of 57.5% across the two models. This *demonstrates the effectiveness of using synthetic in-domain data in mitigating label bias.*

Moreover, we combine SDC with LOOC by averaging their estimated priors. The results indicate that this combination further improves task performance and reduces label bias, with an average 17.6% increase in Micro-F1 and reductions of 62.3% and 57.5% in Bias Score and RSD, respectively. This *highlights the adaptability of* SDC, *which is further enhanced when integrated with other methods.*

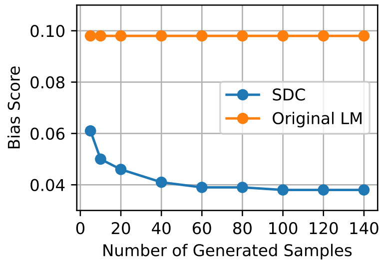


Figure 2: Bias Score of SDC with various number of synthetic in-domain samples.

5.2 Analysis

Generalizability on Number of Demonstrations: The number of demonstrations is a crucial parameter that influences both synthetic data generation and model predictions. We conducted additional experiments on Llama3-7b using 2, 4, 8, and 12 demonstrations, with the results shown in Fig. 1. Notably, under this setting, SDC uses the same number of demonstrations for both synthetic data generation and model predictions. The figure shows that SDC effectively mitigates bias across all tested demonstration sizes and consistently outperforms alternatives in every comparison. This highlights its *strong generalizability to different numbers of demonstrations*.

Impact of Synthetic Data Quantity: The amount of synthetic in-domain data is crucial for SDC, as the model’s prior estimation relies on averaging the model’s prediction distribution over this data. Increasing the amount reduces randomness in the estimated prior. To assess the impact of data quantity, we conducted experiments on Llama3-7b with SDC using synthetic instances ranging from 5 to 140. The results, shown in Fig. 2, demonstrate that SDC consistently mitigates label bias regardless of the data quantity. As the amount of synthetic data increases, SDC achieves a lower Bias Score, indicating stronger bias mitigation. Notably, SDC performs effectively with even only 5 synthetic instances, matching the Bias Score of DC, which uses real in-domain data. These findings suggest that SDC *is effective even with a small number of synthetic examples*, providing a flexible and efficient approach to reducing label bias without the need for real in-domain data.

6 Conclusion

This work introduces SDC (Synthetic Data Calibration) to mitigate label bias in LLMs without requiring real in-domain data. By leveraging LLMs to generate synthetic calibration data, SDC significantly reduces label bias, achieving a 57.5% Bias Score reduction across 279 tasks. Moreover, combining SDC with LOOC further enhances performance, demonstrating its effectiveness and adaptability. These results highlight SDC 's potential in improving LLM reliability across diverse tasks.

Limitations

While our proposed Synthetic Data Calibration (SDC) method demonstrates promising improvements in mitigating label bias across a variety of classification and multiple-choice tasks, several limitations warrant discussion. First, the quality and representativeness of the synthetic in-domain data depend heavily on the underlying generative capabilities of the LLM. In domains with highly specialized or nuanced language, the generated examples may not fully capture the true distribution of real inputs, potentially limiting calibration effectiveness. Second, SDC 's performance is sensitive to the prompt design and the choice of demonstration examples. Small variations in these factors can affect the diversity and accuracy of the synthetic data, suggesting a need for further investigation into robust prompt engineering strategies.

References

AI@Meta. 2024. Llama 3 model card.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? A study on judgement biases. *CoRR*, abs/2402.10669.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen R. McKeown, and He He. 2023. On the relation between sensitivity and accuracy in in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*. Association for Computational Linguistics.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Joonwon Jang, Sanghwan Jang, Wonbin Kweon, Minjin Jeon, and Hwanjo Yu. 2024. Rectifying demonstration shortcut in in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4294–4321. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8086–8098. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics.

Yuval Reif and Roy Schwartz. 2024. Beyond performance: Quantifying and mitigating label bias in llms. *CoRR*.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan

Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*. Association for Computational Linguistics.

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, Jianxin Yang, 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, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *CoRR*.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, Proceedings of Machine Learning Research. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Appendix

A.1 Bias Evaluation Metrics

We follow (Reif and Schwartz, 2024) to use **Bias Score** and **Relative Standard Deviation of class-wise accuracy (RSD)** to assess the label bias in the model's predictions. The Bias Score directly measures the model's tendency toward each class by holding out a set of instances $\mathcal{I}_{BS} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{|\mathcal{I}_{BS}|}, y_{|\mathcal{I}_{BS}|})\}$ from the test set and calculating the average predicted probabilities for each class:

$$p_{BS}(y^i) = \frac{1}{|\mathcal{I}_{BS}^{y^i}|} \sum_{(x,y) \in \mathcal{I}_{BS}^{y^i}} p(y|x, \mathcal{C})$$

Method	Baseline	Label First	Input First	No Label
Bias Score ↓	0.098	0.045	**0.041**	0.073
RSD ↓	0.562	0.287	**0.257**	0.384

Table 3: The bias evaluation results for various prompting strategies.

where $\mathcal{I}_{BS}^{y^i} = \{(x, y) \in \mathcal{I}_{BS} | y = y^i\}$, y^i denotes the answer of the i-th class. Given the average predicted probabilities for each class, the Bias Score is computed as the L1 distance between the model's prediction distribution and the uniform distribution.

$$BiasScore = \frac{1}{2} \sum_{y^i \in Y} \left| p_{BS}(y^i) - \frac{1}{|Y|} \right|.$$

Additionally, RSD assesses the variance in the model's prediction accuracy across classes, defined as:

$$RSD = \frac{\sqrt{\frac{1}{|Y|} \sum_{i=1}^{|Y|} (\text{acc}_i - \text{acc})^2}}{\text{acc}},$$

where acc_i denotes the accuracy of the model's prediction for the i-th class. Note that a lower Bias Score or RSD indicates the model has less tendency toward certain answers, representing lower label bias.

A.2 Prompt Design

We explore three ways to construct the prompt of synthetic data generation:

- **Label First Prompting**, where the demonstration sequence is $\left(y_1, x_1, y_2, x_2, \ldots, y_{|\mathcal{C}|}\right)$ and the LLM is asked to generate the next input $x_{|\mathcal{C}|+1}$ for a (randomly selected) label $y_{|\mathcal{C}|+1}$.
- **Input First Prompting**, where the demonstration sequence is $\left(x_1, y_1, x_2, y_2, \ldots, x_{|\mathcal{C}|}, y_{|\mathcal{C}|}\right)$ and the LLM is asked to *only* generate new input x, without conditioning on a specific label.
- **No Label Prompting**, where the demonstration contains *only* input examples, e.g. $\left(x_1, x_2, \ldots, x_{|\mathcal{C}|}\right)$. This format prompts the model to continue with a new input example $x_{|\mathcal{C}|+1}$, but makes no mention of any label.

We apply these three strategies to Llama3-7b and report their results on SUPER-NATURALINSTRUCTIONS in Table 3. From the table, we see that Input First Prompting achieves the best performance. We suspect this is because it does not require the model to learn explicit input-label correspondences, thus simplifying free-form generation of synthetic in-domain data. At the same time, including the label in the demonstration provides a helpful hint about the overall task.