Leveraging positional bias of LLM in-context learning with Class-few-shot and Maj-Min alternating ordering *

Aleksander Szczęsny[0009−0003−6808−2321], Maciej Markiewicz[0009−0004−2882−6741], Łukasz Radliński[0000−0002−7366−3847], and Przemysław Kazienko[0000−0001−5868−356X]

Wrocław University of Science and Technology, Department of Artificial Intelligence, Wybrzeże Stanisława Wyspiańskiego 27 50-370 Wrocław, Poland

Abstract. Selecting appropriate examples for in-context learning significantly impacts the performance of Large Language Models. In this paper, we show that leveraging LLMs' positional biases and incorporating knowledge of class distribution can improve classification outcomes, especially for underrepresented classes. We introduce *Class-few-shot*, a method that balances class representation among few-shot examples. To investigate this, we conduct almost 10,000 experiments on four datasets and three models, cross-checking how different biases affect models' performance and how they interact. We show that presenting classes from the most to least numerous using an alternating pattern leads to better results than standard few-shot prompting with the same number of examples. Additionally, we compare the general few-shot and *Class-few-shot* results, outlining the strengths of both approaches. All of our raw experiment results, prompts and codes are publicly available on GitHub[1].

Keywords: Few-shot · Class-few-shot · In-context learning · Large Language Models · Positional bias · Alternating ordering · Maj-Min

1 Introduction

Few-shot in-context learning [5] is one of the key approaches to increasing the performance of Large Language Models (LLMs) [10], and one that can be applied with almost no additional effort. Over the years, the identification of factors that contribute to the effectiveness of few-shot has become an area of research. This paper aims to verify how positional bias and class distribution influence the

* Work financed by: (1) the National Science Centre, Poland, project no. 2021/41/B/ST6/04471; (2) CLARIN ERIC funded by the Polish Minister of Science, no. 2024/WK/01; (3) the Polish Ministry of Education and Science within the programme "International Projects Co-Funded"; (4) the European Union under the Horizon Europe, grant no. 101086321 (OMINO). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor European Research Executive Agency can be held responsible for them. (5) "The Digital Research Infrastructure for the Arts and Humanities DARIAH-PL" supported by the development plan under the investment A2.4.1 Investment in research capacity building of the National Recovery and Resilience Plan.

[1] https://github.com/olorules/Class-few-shot

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7

results and proposes **Class-few-shot** – a method designed to increase performance, especially the recall for minority classes, by balancing the class distribution. The key contributions of the paper include: (1) A Class-few-shot *Maj-Min* alternating order method for selecting in-context examples that improves performance. (2) Analysis of class balancing influence on the most numerous class *Maj* and the least numerous class *Min* results. (3) Nearly 10,000 publicly available experiments for further study.

2 Related work

In-context learning [5] is an important alternative to fine-tuning [19] and can be used in a wide range of applications, from simple classification [10] to ad hoc personalization [18]. Initial research on in-context learning quickly identified some factors that influenced performance, such as recency bias and majority label bias [21], or the number and quality of examples [13]. These discoveries led to further exploration, particularly in two areas central for this work: the impact of information conveyed in the prompt, and the impact of order and example distribution on model performance.

2.1 Example order, distribution and demonstration

One of the key biases demonstrated in [21] was models' tendency to prioritize the majority label and examples closer to the end of the prompt, which was present even in many-shot in-context learning [1]. Multiple strategies have been suggested to overcome the order of examples issue [12,16,20]. There has not been much research comparing the ordering bias with other biases [21]. Most works related to class distribution focus on developing methods, for example, selection [2, 11, 17], or dataset balancing [9] and expansion [8]. However, recent research confirms the significant impact of dataset imbalance on in-context learning [6].

3 *Class-few-shot* and positional bias

In the standard few-shot approach, the examples presented in a prompt are usually chosen randomly or manually, and the number of those examples is fixed. While sampling makes the class distribution of few-shot examples resemble the class distribution of the dataset, it may result in some classes being missing or overrepresented. Additionally, random few-shot sampling does not fully take advantage of the positional bias, as it cannot guarantee consistent class presence.

3.1 *Class-few-shot*: a new approach to balancing few-shot class distribution

We present a new method called *Class-few-shot*, Figure 1. Its purpose is to ensure that when sampling few-shot examples, each class is represented n times. This means that the final *shot* number depends on the class number of the specific task. The precise class presentation is controlled by parameters described below:

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7

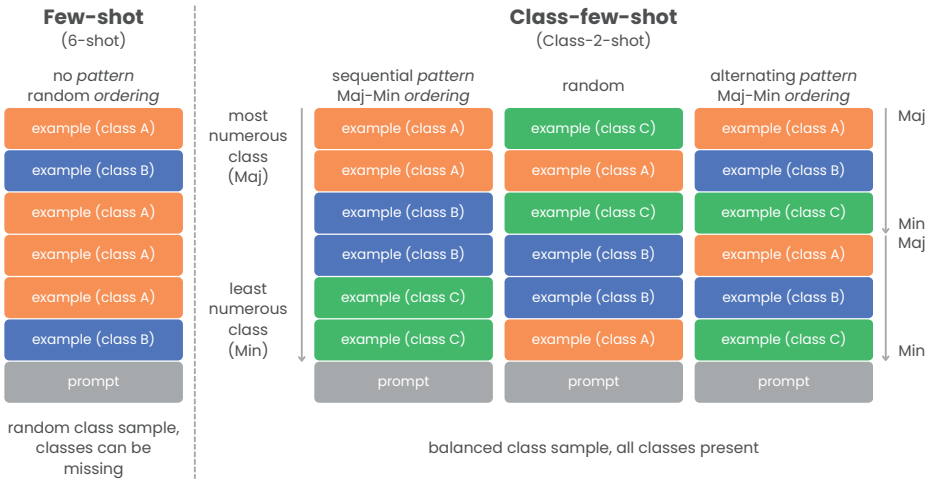

Fig. 1. *Class-few-shot* and few-shot in three classes scenario

- **Class order** (*Ordering*) determines the order of examples representing each class in prompts, based on the number of examples belonging to the class:
  - ***Maj-Min*** – Sequence from the most to the least numerous class (first majority class *Maj*, last minority class *Min*),
  - ***Min-Maj*** – Sequence from the least to the most numerous class.
- **Class-few-shot pattern** (*Pattern*) defines a pattern of the *class order* in a prompt. It is only relevant for $n > 1$ shots per class:
  - **Sequential** (*Seq*) – all examples from a class are presented together, then all examples from the next class, etc., following given *Ordering*,
  - **Alternating** (*Alt*) – examples are given one at a time, presenting different classes sequentially.
- **Random** – a special case of Class-few-shot with a fully random example order. Class balance is maintained with n examples per class, but no specific *ordering* or *pattern* is applied. This serves as a baseline where only class balance influences results, isolating the effect from positional bias.

3.2 Positional bias

The suggested approach aims to exploit the positional bias of LLMs. Prior research indicated bias towards information near the end of the prompt [21] and suggested that models are prone to favoring a specific position of the label [1, 12, 16, 20]. We test various orderings and patterns to identify which one can be leveraged with *Class-few-shot* to provide better example selection for unbalanced datasets, as the bias effects vary by task and model.

We test various orderings and patterns to identify which one can be leveraged with *Class-few-shot* to provide better example selection for unbalanced datasets, as the bias is proven to work differently depending on the task and model.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7

4 Experimental setup

We would like to compare the impact of *Class-few-shot* to all other factors influencing the performance, as it is difficult to isolate the impact of one specific factor. We decided to choose four different datasets, described in Section 4.1, as well as three representative models of different sizes: **Llama-3.1-8B-Instruct** (small), **Mixtral-8x7B-v0.1-Instruct** (medium, MoE), and **Llama-3.3-70B-Instruct** (large). To perform the experiments, we used the *lm-evaluation-harness* [7] framework with custom samplers implementing Class-few-shot. Models were hosted using the *vLLM* library.

4.1 Datasets

We chose 4 single-label classification datasets of different type, balance, class number, and difficulty (based on [10]). Dataset statistics are presented in Table 1. For the SNLI [4] dataset, we used the train set for few-shot samples and have sampled a 10% (5k) stratified random subset from the set to increase performance. The test set has not been altered. For Sarcasmania [15] we built the train and test sets by splitting the whole dataset with a train/test ratio of 9:1, with stratification.

4.2 Experiment plan

Our experiments considered testing all permutations of parameters against each other, for all models and datasets. *Class-few-shot* parameters included *Class Order* and *Class-few-shot Pattern*. Performance in standard *few-shot* was also evaluated as a reference. Each experiment has been repeated 10 times, using a fixed seed for each run: *1234*, *1235*, *1236*, etc.

To ensure reproducibility, we separated examples sampling from order and pattern. Thanks to that, all experiments were carried out with identical sets of

Dataset	Class	Count	Percentage	Total Samples	Difficulty	Balance
SNLI [4]	Entailment	3368	34.3%	9824	Easy	Well-balanced
	Contradiction	3237	32.9%			
	Neutral	3219	32.8%			
TweetEval [3]	Anger	558	39.3%	1421	Medium	Imbalanced
	Sadness	382	26.9%			
	Joy	358	25.2%			
	Optimism	123	8.7%			
Sarcasmania [15]	No-Sarcasm	2129	53.5%	3978	Hard	Well-balanced
	Sarcasm	1849	46.5%			
Word Context [14]	Yes (Same Meaning)	319	50.0%	638	Medium	Perfect
	No (Other Meaning)	319	50.0%			

Table 1. Class distribution of datasets used in the study (test/validation set). Percentages are calculated relative to the total number of samples in each dataset.

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7

examples, which were constant between configurations. For example, a *Min-Maj Alt* run on seed *1234* had the same few-shot examples as *Maj-Min Alt*, *Min-Maj Seq*, or *random* runs on the same seed. Class list in the prompts was presented in the same way across all experiments.

Experiments with *class-n-shot* were carried out in three scenarios in which the parameter $n \in \{1, 2, 3\}$, and the model was given n examples representing each class. The number of examples used in reference *few-shot* experiments was the same as in corresponding *class-n-shot* configurations. E.g., for the TweetEval dataset (4 classes), the model was given either $n = 4$, $n = 8$, or $n = 12$ examples. This enables a direct comparison of the results obtained by the two approaches.

4.3 *Class-few-shot* as an experimental methodology

To test the positional and class bias, we have to make sure that the test environment provides as much stability and reproducibility as possible. This is why we use the *Class-few-shot* method to perform all of the experiments. In standard few-shot, class distribution of examples will reflect the class distribution of the dataset. In the case of a strongly imbalanced dataset, there might be some runs with no possibility of creating a consistent *pattern*, or *class order*, as few-shot could only consist of one class. This could potentially influence the results. Therefore, we use *Class-few-shot* to assess biases, and perform standard few-shot experiments only as a reference.

5 Results

The following section presents a selection of experiments and observed trends, along with a thorough analysis.

5.1 Positional bias in *Class-few-shot*

The results presented in Table 2 illustrate the impact of *ordering* and *pattern* on the performance of LLMs in classification tasks across selected datasets. While the performance differences between various ordering strategies are not striking, the consistently better results of the *Maj-Min alternating ordering* across all tested models suggest that structured variation in example presentation plays a role in optimizing model performance. Additionally, the random scenario tends to provide stable, above average results.

Furthermore, it is worth highlighting the variability in the impact of these parameters on specific models. For both Llama models, despite the gap between sizes, the differences observed across various parameter configurations were significantly bigger than those seen for Mixtral-8x7B. This suggests that model architecture may play a critical role in leveraging positional bias.

Table 3 shows the mean F1 score for the least numerous class, *optimism* (9%), in the most imbalanced dataset, TweetEval. The results indicate that any *Class-n-shot* setting significantly outperforms standard few-shot for the minority class,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7

Dataset	Model	*Maj-Min*		*Min-Maj*		Random
		Seq	*Alt*	*Seq*	*Alt*	
SNLI	Llama-3.1-8B	0.750	**0.770**	0.740	0.760	0.764
	Mixtral-8x7B	**0.815**	0.802	0.778	0.791	0.799
	Llama-3.3-70B	**0.821**	**0.820**	0.794	0.805	0.811
TweetEval	Llama-3.1-8B	0.773	0.778	**0.781**	0.774	0.777
	Mixtral-8x7B	0.728	**0.740**	**0.738**	0.731	0.734
	Llama-3.3-70B	0.767	**0.775**	0.772	0.771	**0.774**
Sarcasmania	Llama-3.1-8B	0.721	**0.778**	0.709	0.706	0.732
	Mixtral-8x7B	**0.945**	**0.945**	0.912	0.910	0.928
	Llama-3.3-70B	0.822	**0.847**	**0.848**	0.840	0.839
Word Context	Llama-3.1-8B	0.613	0.638	**0.651**	0.621	0.631
	Mixtral-8x7B	0.684	0.679	0.684	**0.692**	0.685
	Llama-3.3-70B	0.725	**0.728**	0.718	0.722	0.721
Average	Llama-3.1-8B	0.714	**0.741**	0.720	0.715	0.726
	Mixtral-8x7B	**0.793**	**0.791**	0.778	0.781	0.786
	Llama-3.3-70B	0.784	**0.793**	0.783	0.784	0.786

Table 2. Comparison of different *Class-few-shot* scenarios. The results presented above are averaged from Class-2-shot and Class-3-shot (Class-1-shot does not enable *pattern* usage).

suggesting that *Class-few-shot* helps balance output distribution and increases sensitivity to minority classes. In contrast to the outcomes detailed in Table 2, here the *Seq* pattern performs best, placing rare class instances closer to the prompt, which is consistent with prior findings [21].

5.2 *Class-n-shot* vs. few-shot comparison

Figure 2 shows the average performance differences between standard few-shot and *Class-few-shot* across all experiments, highlighting the potential gains of using *Class-few-shot*.

For Llama-8B, both *class-2-shot* and *class-3-shot* outperform in the *random* and *Maj-Min Alternating* scenario, with the latter improving the results by 2 percentage points compared to standard few-shot. In the case of Mixtral 8×7B, *Class-few-shot* consistently outperforms standard few-shot across all tested configurations. Llama-70B exhibits the least gain from employing *Class-few-shot*. Performance generally improves with more examples, and while the *Maj-Min Alternating* strategy yields the best results, it is worth noticing that the *random*

Model	Few-Shot	Random	*Min-Maj*		*Maj-Min*	
			Seq	*Alt*	*Seq*	*Alt*
Llama3.1_8B	0.568	0.618	0.597	0.604	**0.633**	0.621
Mixtral_8x7B	0.439	0.497	0.496	0.502	**0.505**	0.481
Llama3.3_70B	0.619	0.628	0.616	0.628	**0.637**	0.632

Table 3. F1 score for the minority class of the TweetEval dataset (*optimism*).

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7

strategy results in better performance, suggesting that class balancing alone can boost in-context learning. Note that results for Sarcasmania were excluded for Llama-70B due to extremely poor 0-shot performance, with the model often failing to produce coherent responses.

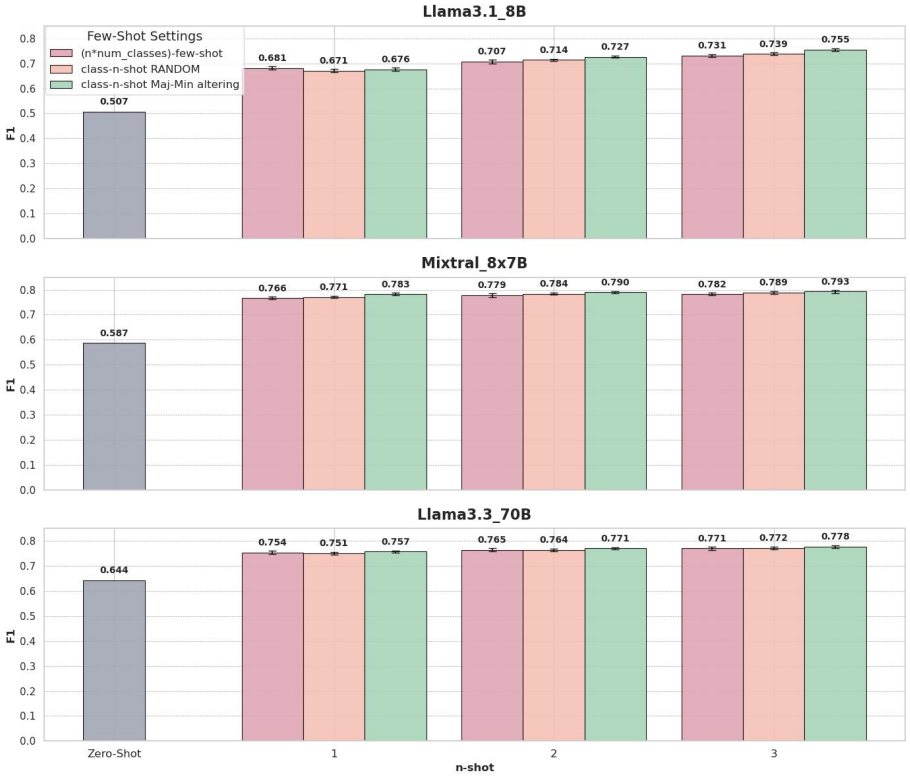

Fig. 2. F1 comparison between 0-shot, few shot and *Class-few-shot*, averaged across all the experiments with $n \in \{1, 2, 3\}$.

6 Conclusions and future work

Our findings reveal that *Maj-Min Alternating Class-few-shot* improves the performance of LLMs over the standard few-shot approach. Furthermore, the results confirm that *Class-few-shot* can improve the F1 score on minority classes by approximately 5 p.p.

Potential future work could explore how *Class-few-shot* affects the class distribution, as minority class results have improved significantly. Other directions include analysing positional bias with *Class-few-shot* in a many-shot [1] context,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7

as well as expanding the number of datasets and studying the role of class order in prompts beyond few-shot examples.

References

1. Agarwal, R., et al.: Many-shot in-context learning (2024)
2. Baldassini, F.B., Shukor, M., Cord, M., Soulier, L., Piwowarski, B.: What makes multimodal in-context learning work? (2024)
3. Barbieri, F., Camacho-Collados, J., Neves, L., Espinosa-Anke, L.: Tweeteval: Unified benchmark and comparative evaluation for tweet classification (2020)
4. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference (2015)
5. Brown, T.B., et al.: Language models are few-shot learners (2020)
6. Gao, H., Zhang, F., Zeng, H., Meng, D., Jing, B., Wei, H.: Exploring imbalanced annotations for effective in-context learning (2025)
7. Gao, L., et. al.: A framework for few-shot language model evaluation (12 2023). https://doi.org/10.5281/zenodo.10256836
8. Kaszyca, O., Kazienko, P., Kocoń, J., Cichecki, I., Kochanek, M., Szydło, D.: Is it possible for chatgpt to mimic human annotator? Authorea Preprints (2023)
9. Kochanek, M., et al.: Improving training dataset balance with chatgpt prompt engineering. Electronics **13**(12), 2255 (2024)
10. Kocoń, J., et al.: Chatgpt: Jack of all trades, master of none. Information Fusion p. 101861 (2023)
11. Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., Chen, W.: What makes good in-context examples for gpt-3? (2021)
12. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity (2021)
13. Min, S., et al.: Rethinking the role of demonstrations: What makes in-context learning work? (2022)
14. Pilehvar, M.T., Camacho-Collados, J.: Wic: the word-in-context dataset for evaluating context-sensitive meaning representations (2019)
15. Siddiqui, R.: Sarcasmania dataset (2019)
16. Sorensen, T., et al.: An information-theoretic approach to prompt engineering without ground truth labels (2022)
17. Wang, S., Yang, C.H.H., Wu, J., Zhang, C.: Bayesian example selection improves in-context learning for speech, text, and visual modalities (2024)
18. Woźniak, S., Duszenko, J., Kocoń, J., Kazienko, P.: Improving llm-based recommender systems with user-controllable profiles. In: The 1st Workshop on Human-Centered Recommender Systems at TheWebConf 2025 - The ACM Web Conference 2025. ACM (2025)
19. Woźniak, S., Koptyra, B., Janz, A., Kazienko, P., Kocoń, J.: Personalized large language models (2024)
20. Wu, Z., Wang, Y., Ye, J., Kong, L.: Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering (12 2022)
21. Zhao, T.Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models (2021)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97635-3_7