Demonstrations in In-context Learning for LLMs with Large Label Space

Zhan Li [1] **Fanghui Liu** [2] **Volkan Cevher** [1] **Grigorios Chrysos** [3]

Abstract

In-context learning (ICL) can solve new tasks on pre-trained Large Language Models (LLMs) given a few demonstrations as input. However, so far there is little understanding of how many demonstrations are required for the real-world applications with large label spaces. In this work, we conduct a meticulous study under various settings with different LLMs and datasets. Our insights suggest that: (i) we might not even need demonstrations, especially when the class names are descriptive and the model is strong-performing (e.g., GPT-4). Nevertheless, (ii) datasets with extremely large label space can benefit with additional human-created demonstrations. Lastly, (iii) automatically generated demonstrations might not yield additional benefits. We believe our study leads to new insights on understanding the inductive bias of ICL.

1. Introduction

Machine learning has witnessed remarkable advances with the advent of Large Language Models (LLMs) such as GPT (Radford et al., 2018; 2019; Brown et al., 2020). These LLMs enable a novel paradigm of in-context learning (ICL), where the LLM can answer a query that is not part of its original objective without changing the LLM parameters. Nevertheless, we can supply few examples as input, which are referred to as *demonstrations* (Brown et al., 2020). ICL offers a significant advantage over training a specialized model, as it can achieve similar results without requiring large amounts of data (Radford et al., 2019). Moreover, ICL allows the user to perform the task by simply and instantly interacting with the LLM. The key question now is *how many demonstrations* to provide.

[1]Lab for Information and Inference Systems, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. [2]Department of Computer Science, University of Warwick, Coventry, UK. [3]Department of Electrical & Computer Engineering, University of Wisconsin-Madison, Madison, WI, United States.. Correspondence to: Zhan Li <zhan.li@epfl.ch>.

Proceedings of the 1st Workshop on Long-Context Foundation Models, Vienna, Austria. 2024.

Zhao et al. (2021); Hashimoto et al. (2023) advocate that few-shot demonstrations outperform zero-shot (i.e., no demonstrations) in ICL under a limited label-space. Min et al. (2022) investigate which parts of the demonstration and the prompt instructions are critical to improve ICL. Despite the progress in understanding the role of demonstrations, the aforementioned works focus on small label spaces, while in real-world applications larger label spaces are important (Minaee et al., 2021; Apté et al., 1994; Kowsari et al., 2017). In other words, the following question remains yet elusive: *How does the number of demonstrations affect the performance under large label spaces?*

In this paper, we explore this question by studying three commonly-used settings across various LLMs. The first setting is retrieval, which is the most successful setting to date for ICL. In retrieval we recover demonstrations from a pool of *similar* queries to our test query. Our first key insight dictates that the number of demonstrations required differs substantially per LLM used. Notably, the highest accuracy is achieved under the retrieval setting with GPT-4, when we use more demonstrations than the recommendations of previous work.

Our second setting alleviates the limitation for a pool of queries. In particular, when demonstrations are not readily available, we can synthesize them using an auxiliary LLM. This setting with self-generated demonstrations, called 2LM, has a high variance depending on the LLM used, while it usually cannot match the performance of the retrieval setting. We examine the potential reasons behind the ineffectiveness of self-generated demonstrations. The third setting, which is called CoT-2LM extends the 2LM setting using the idea of Chain-of-Thought (Wei et al., 2022b), which has been effective in improving the performance of LLMs.

The *key* surprising message though is that demonstrations might not be required in all cases. Notably, we consider a setting without any demonstrations, called "zero-shot", and observe that the model can still perform on par with the retrieval setting in many cases. Given the runtime efficiency of zero-shot setting, we do believe this is a promising avenue even for large label settings. We believe that this zero-shot setting can also further scrutinize the quality of the latent space of LLMs, since we observe high variance in the performance of LLMs on this setting.

2. Methodology

Let us define the problem and the various prompt settings we assess in this work.

2.1. Problem description

We aim to study in-context learning (ICL) in large label-space text classification. We denote $\mathcal{X}$ as the input text space and $\mathcal{Y}$ as the label space. We assume we have N class labels. Let $\mathcal{S} = \cup_{k\in\mathbb{N}}\{(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k, \mathcal{Y}) : \boldsymbol{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$ be the set of k finite-length sequences of $(\boldsymbol{x}, y)$ pairs followed by the label space. We refer to $(\boldsymbol{x}_1, y_1, \ldots, \boldsymbol{x}_k, y_k, \mathcal{Y})$ as context $\mathcal{C}_k$ and the $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^k$ pairs in context as *demonstrations*. The context $\mathcal{C}_k$ and a test query $\boldsymbol{x}_{\text{test}} \in \mathcal{X}$, along with the instruction set ι are fed into a pretrained language model (henceforth referred to as *predictor*). We assume the predictor has fixed parameters. The *predictor* model $\mathcal{M} : \mathcal{S} \times \mathcal{X} \to \mathcal{Y}$ should predict y_{test} as follows:

$$y_{\text{pred}} = \arg\max_{\mathcal{Y}} \mathcal{M}(\cdot \mid \mathcal{C}_k, \boldsymbol{x}_{\text{test}}, \iota) \ .$$

2.2. Prompt configurations

In this paper, we consider a variety of prompt configurations that have emerged in the literature and (partially) depend on the availability of demonstrations.

Zero-shot ICL. We *only* provide the label information and the test query to the predictor. The context $\mathcal{C}_k$ is then reduced to the label space, i.e., $\mathcal{C}_0 = \mathcal{Y}$, while $k = 0$. We call this the "zero-shot" setting throughout this paper. This is the most general and economical setting that does not require any domain-specific demonstrations and uses fewer input tokens. This setting can be applied to any test query without needing any additional "similar" demonstrations as the retrieval setting.

Few-shot ICL with Retrieval. We utilize demonstrations that are "similar" to the test query. We employ a retrieval model to recover demonstrations from a pool of available demonstrations. As a reminder, we have N classes in total, while the retrieval pool contains demonstrations categorized per class. We sample 7 demonstrations per class uniformly at random for our retrieval[1]. Then, the retrieval model compares the embedding vectors of the sampled demonstrations with the embedding vectors of the test query. The top-k samples with the highest cosine similarity are selected as the demonstrations of ICL. This setting, referred to as *retrieval*, is inspired by Milios et al. (2023a). A schematic of the retrieval setting is depicted in Appendix C Figure S4(a).

[1]This number is chosen because most of our experiments have $k \leq 7$. In addition, the experiments in Appendix G exhibit that even larger retrieval pool does not offer any benefits.

Few-shot ICL with 2LM. A core drawback of the retrieval setting is the requirement for a pool of demonstrations that are similar to the test query. In case such human-labeled data are not available, we can synthesize the demonstrations using an additional LM. The first LM, called the *generator*, synthesizes k demonstrations using the label space and the test query as the input. Those demonstrations are then used as the context for the predictor model. A schematic of the 2LM setting is depicted in Appendix C Figure S4(b).

Few-shot ICL with CoT-2LM. We augment the last setting using chain-of-thought (CoT) arguments, which enhance the performance of LMs as reasoners (Wei et al., 2022b). CoT inserts an intermediate reasoning process into the demonstration. We modify the instruction in 2LM setting to require the *generator* to synthesize demonstrations with reasoning steps. Here the input-output pair become input-reason-output tuple $(\boldsymbol{x}_1, \boldsymbol{r}_1, y_1, \ldots, \boldsymbol{x}_k, \boldsymbol{r}_k, y_k)$, where r_i denotes the reasoning process. The components of this framework are shown in Appendix C Figure S9.

3. Experiments

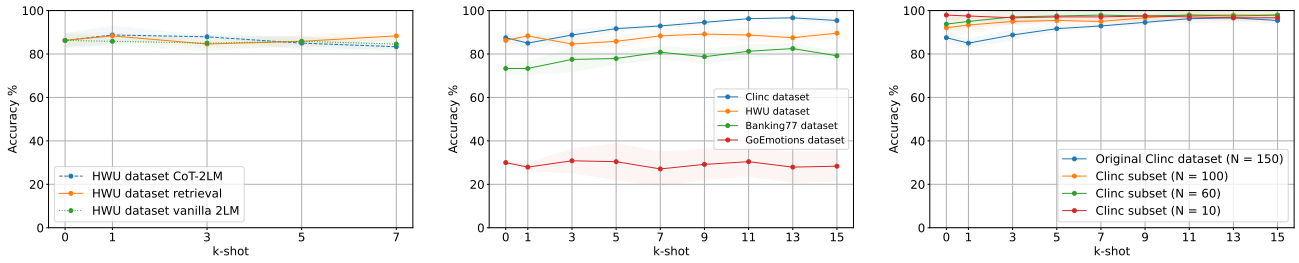
In this section, we conduct experiments on GPT-4, GPT-3.5[2], Mixtral (Jiang et al., 2024), and LLaMA-2 (Touvron et al., 2023) under three different configurations, with four datasets: Banking77 with 77 unique labels (Casanueva et al., 2020), Clinc150 with 150 unique labels (Larson et al., 2019), HWU64 with 59 unique labels (Liu et al., 2019) and GoEmotions with 27 unique labels (Demszky et al., 2020). Our detailed experimental setup can be found in Appendix B.

3.1. GPT-4 on different prompt configurations

Our first experiment assesses the performance of the configurations of Section 2.2 on GPT-4, which is currently the strongest performing model. The results *on HWU* in Figure 1(a) exhibit that the *zero-shot* is on par with the k-shot configurations. Notice that the performance remains stable across different number of demonstrations.

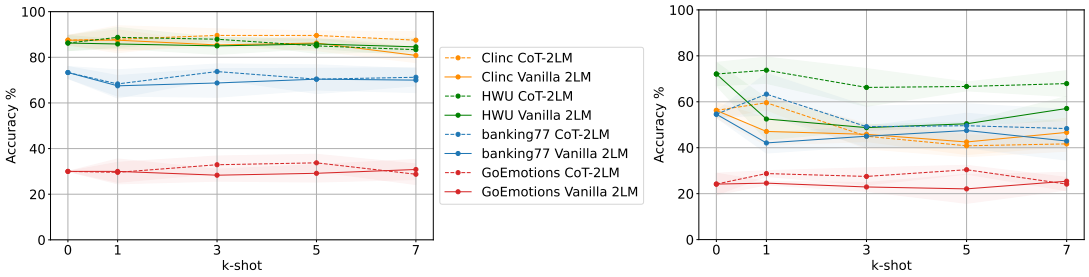
To further evaluate the effect of the number of demonstrations k on the retrieval setting, we generalize our validation to all datasets. We observe in Figure 1(b) that the retrieval performance improves with increasing k in 2 out of the 4 datasets, namely Banking77 and Clinc, with Clinc showing the largest improvement. A natural question is whether this improvement is related to the larger number of classes of Clinc and Banking77 datasets over the HWU and GoEmotions. We measure the impact of the number of classes on the number of demonstrations by randomly sampling a subset of classes from the Clinc dataset. The results in

[2]https://platform.openai.com/docs/models



(a) GPT-4 on HWU across three settings (b) retrieval with GPT-4 on four datasets (c) retrieval with GPT-4 on Clinc

Figure 1. (a) Accuracy on HWU over three settings on GPT-4. Notice that the few-shot settings do not result in significant performance gains compared to the zero-shot on GPT-4. (b) Accuracy over retrieval setting on GPT-4 under the retrieval pool: $7N$. Retrieved demonstrations result in a gradual increase on intent classification datasets. The Clinc dataset has the largest increase, but only when additional demonstrations are utilized. (c) Accuracy under a decreasing size of label space on Clinc dataset. The larger the label space, the higher the number of retrieved demonstrations required. Highlighted regions denote the standard deviation in this paper.



(a) GPT-4 (b) GPT-3.5

Figure 2. (a) Accuracy of vanilla 2LM and CoT-2LM settings (depicted with solid and dashed lines respectively) on GPT-4. CoT results in a minor improvement from vanilla 2LM, especially in the regime of $\{3, 5\}$ demonstrations. (b) Accuracy of vanilla 2LM and CoT-2LM settings on GPT-3.5. Vanilla-2LM brings dramatic accuracy decrease, while CoT greatly improves the performance from vanilla 2LM. When k = 1, CoT-2LM outperforms zero-shot, but decreases as k gradually increases.

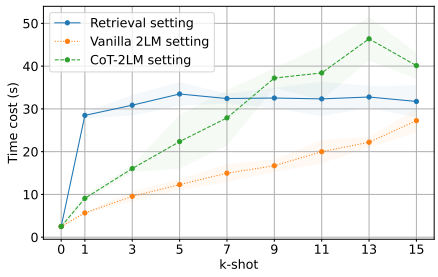


Figure 3. Time (in seconds) required to perform a full-flow GPT-4 ICL task on Clinc dataset. Once retrieval is involved, the cost of time increases steeply.

Figure 1(c) exhibit that *the number of classes significantly affects the performance with respect to the number of demonstrations*. Specifically, when the number of classes is reduced to 100 or less, adding more than 5 demonstrations does not lead to significant improvement, which is much lower than the optimal $k = 13$ in the full dataset. Moreover, when the number of classes is only 10, the zero-shot setting outperforms all the k-shot retrieval settings.

We extend our experiments on both the vanilla 2LM and the CoT-2LM settings. The results in Figure 2(a) indicate that neither the vanilla 2LM nor the CoT-2LM settings have a consistent improvement across datasets and different number of demonstrations. However, notice that the CoT-2LM setting results in a slight increase, particularly in the regime of 3 or 5 demonstrations. This is more evident in the challenging dataset of GoEmotions, where the performance of CoT-2LM consistently surpasses retrieval setting for all values of k up to and including 7.

Beyond the accuracy, the computational complexity of each configuration differs. To evaluate that, we sample randomly five test queries in the Clinc dataset and repeat each experiment three times. As visualized in Figure 3 the retrieval setting is more computationally demanding for up to 9 demonstrations with the zero-shot setting being the fastest.

3.2. Experiments on different models

We extend our validation to the popular models of GPT-3.5, Mixtral 8x7B and LLaMA-2. As we elaborate below, the three models differ significantly from GPT-4 with respect

to their ICL performance. Noticeably, in all cases the performance is not better than the corresponding accuracy on GPT-4. Initially, we focus on the HWU dataset as in GPT-4 and then extend our experimentation to all datasets in the retrieval setting.

GPT-3.5 and Mixtral 8x7B: As illustrated in Figure 2(b) and Figure S6, the zero-shot setting performs favorably to the vanilla 2LM setting, which is consistent with our previous observations. However, there are *two* key differences from GPT-4: (i) the CoT-2LM setting improves substantially the vanilla 2LM setting, (ii) the gap between the retrieval and the vanilla 2LM setting is large. The results with CoT-2LM on GPT-3.5 across all datasets (see Figure 2(b)) exhibit the peak performance at $k = 1$. That differs from the observations on GPT-4, in which $k > 1$ performs favorably in CoT-2LM.

We extend our experiments on the retrieval setting to further compare the models. The results in Figure S5 depict *two* key differences from GPT-4: (i) the retrieval setting is more important on GPT-3.5 and Mixtral when compared to GPT-4, (ii) the performance saturates after 5 demonstrations. Interestingly, in Clinc dataset, Mixtral performs worse than GPT-3.5, while in GoEmotions the demonstrations hurt the performance in Mixtral.

LLaMA-2: As indicated in Figure S6(c), LLaMA-2 differs from all the aforementioned models in: (i) its zero-shot performance is much lower than the previous models, (ii) CoT-2LM does not have consistent benefits across different number of demonstrations. Nevertheless, in the retrieval setting with more demonstrations LLaMA-2 performs on par with GPT-3.5 and Mixtral. Notably, in GoEmotions dataset the retrieval setting of LLaMA-2 outperforms the equivalent results in Mixtral as visualized in Figure S5(c).

3.3. Key takeaway messages on configurations

Zero-shot setting: Both GPT models and Mixtral offer a strong-performing zero-shot setting. We hypothesize that as long as the class labels are descriptive, zero-shot is a valid option that offers little to no overhead as a side benefit.

Retrieval setting: All four datasets include demonstrations, which makes them amenable to the retrieval setting. However, having available pool of demonstrations on arbitrary ICL tasks might be harder, especially under the more realistic scenario of large label spaces. Yet, the retrieval setting can be beneficial, particularly when the model is not GPT-4. Concretely, for GPT-3.5, Mixtral and LLaMA-2 a few demonstrations can bring significant performance increase. Additionally, our ablation experiment in Appendix G exhibits that randomly selecting demonstrations instead of optimizing their selection may hurt the performance.

2LM settings: CoT leads to an improved performance in most cases. CoT 2LM setting is particularly helpful in the GoEmotion dataset, where a step-by-step process is more beneficial than human-labelled emotions, which might not be ideal for ICL.

When should we avoid the zero-shot setting? The datasets used so far include descriptive class labels, which can partly explain the strong performance of the zero-shot setting. On the contrary, when the labels are not descriptive, demonstrations are critical for ICL. We provide evidence with the following experiment: we manually modify the class labels of HWU by randomly inserting letters and digits. For instance, the altered version of the label 'lists_createoradd' is transformed into 'lisPts_7creyatelora1dd'. The zero-shot performance of GPT-3.5 in this modified label space is significantly reduced as reported in Figure S7. If we use as few as 1 demonstration, the performance increases significantly.

Why do self-generated demonstrations degrade the performance on GPT-3.5 but not on GPT-4? As mentioned above, the vanilla 2LM setting degrades the performance on GPT-3.5, while on GPT-4 the performance remains mostly the same. Can this be attributed to the *generator* or to the *predictor* model? We conduct the following ablation to assess which of the two models has a larger impact on the final result. We utilize GPT-3.5 and GPT-4 for this task as two representative instances. As illustrated in Figure S8(a) when GPT-4 serves as the *generator* and GPT-3.5 as the *predictor* LM, there is a significant improvement in performance with respect to the vanilla 2LM setting of GPT-3.5. Noticeably, in the 1-shot setting, this combination of GPT-4 and GPT-3.5 surpasses even the CoT-2LM setting. This is indicative of higher quality demonstrations synthesized by GPT-4. Let us now assess the symmetric setting, where GPT-3.5 synthesizes the demonstrations and GPT-4 acts as the *predictor* model. The results in Figure S8(b) exhibit that there is no significant difference from the vanilla 2LM setting with GPT-4. We hypothesize that GPT-4 is capable of weighting the quality of the demonstrations and thus the performance does not suffer significantly.

4. Conclusion

In this paper, we explore the critical question of how many demonstrations are required for ICL under different settings and across datasets. We compare the performance of the retrieval setting, 2LM setting, and CoT-2LM setting on the most prevailing LLMs for ICL text classification in a large label space. We show that state-of-the-art LLMs deliver promising zero-shot performance, but retrieved demonstrations can help LLMs improve accuracy.

Impact Statement

In this paper, we investigate the phenomenon of in-context learning (ICL), where large language models (LLMs) can perform new tasks without any parameter updates, just by conditioning on natural language prompts that include task demonstrations. We explore the factors that influence the effectiveness of ICL, such as the dataset size, the model, and the number of demonstrations. We acknowledge that our work has potential implications for the security and ethics of LLMs, especially as they become more accessible and powerful. On one hand, our work can help developers and users of LLMs to leverage ICL for various applications, such as natural language understanding, generation, and translation, with minimal computational and data resources. On the other hand, our work can also expose the vulnerabilities and biases of LLMs, which can be exploited or manipulated by malicious actors. Nevertheless, we strive in our work to focus on datasets that have been ethically approved in the past, and also encourage the community to further increase the safety and guardrails of these models.

Acknowledgements

This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043). The research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-24-1-0048. This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_205011. GC acknowledges travel support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951847.

References

An, S., Lin, Z., Fu, Q., Chen, B., Zheng, N., Lou, J.-G., and Zhang, D. How do in-context examples affect compositional generalization? *arXiv preprint arXiv:2305.04835*, 2023.

Anonymous. On task description of in-context learning: A study from information perspective. In *Submitted to The Twelfth International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=TFR0GrzERG. under review.

Apté, C., Damerau, F., and Weiss, S. M. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*, 12(3):233–251, 1994.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901, 2020.

Casanueva, I., Temcinas, T., Gerz, D., Henderson, M., and Vulic, I. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020*, mar 2020. URL https://arxiv.org/abs/2003.04807. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.

Chen, J., Chen, L., Zhu, C., and Zhou, T. How many demonstrations do you need for in-context learning? In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://openreview.net/forum?id=JPUx2nVgWa.

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., and Ravi, S. GoEmotions: A Dataset of Fine-Grained Emotions. In *Association for Computational Linguistics (ACL)*, 2020.

Hashimoto, K., Raman, K., and Bendersky, M. Take one step at a time to know incremental utility of demonstration: An analysis on reranking for few-shot in-context learning. *arXiv preprint arXiv:2311.09619*, 2023.

Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T. L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. Mixtral of experts, 2024.

Kim, H. J., Cho, H., Kim, J., Kim, T., Yoo, K. M., and Lee, S.-g. Self-generated in-context learning: Leveraging auto-regressive language models as a demonstration generator. *arXiv preprint arXiv:2206.08082*, 2022.

Kossen, J., Rainforth, T., and Gal, Y. In-context learning in large language models learns label relationships but is not conventional learning. *arXiv preprint arXiv:2307.12375*, 2023.

Kowsari, K., Brown, D. E., Heidarysafa, M., Jafari Meimandi, K., , Gerber, M. S., and Barnes, L. E. Hdltex: Hierarchical deep learning for text classification. In *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE, 2017.

Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., Kummerfeld, J. K., Leach, K., Laurenzano, M. A., Tang, L., and Mars, J. An evaluation dataset for intent classification and out-of-scope prediction. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1311–1316, Hong Kong, China,

November 2019. Association for Computational Linguistics (ACL). doi: 10.18653/v1/D19-1131. URL https://aclanthology.org/D19-1131.

Li, J., Zhang, Z., and Zhao, H. Self-prompting large language models for open-domain qa. *arXiv preprint arXiv:2212.08635*, 2022.

Li, R., Wang, G., and Li, J. Are human-generated demonstrations necessary for in-context learning? *arXiv preprint arXiv:2309.14681*, 2023.

Li, X. and Qiu, X. Finding supporting examples for in-context learning. *arXiv preprint arXiv:2302.13539*, 2023.

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. What makes good in-context examples for GPT-3? In Agirre, E., Apidianaki, M., and Vulić, I. (eds.), *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.deelio-1.10. URL https://aclanthology.org/2022.deelio-1.10.

Liu, X., Eshghi, A., Swietojanski, P., and Rieser, V. Benchmarking natural language understanding services for building conversational agents. *arXiv preprint arXiv:1903.05566*, 2019.

Lu, Y., Bartolo, M., Moore, A., Riedel, S., and Stenetorp, P. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Association for Computational Linguistics (ACL)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics (ACL). doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Milios, A., Reddy, S., and Bahdanau, D. In-context learning for text classification with many labels. In Hupkes, D., Dankers, V., Batsuren, K., Sinha, K., Kazemnejad, A., Christodoulopoulos, C., Cotterell, R., and Bruni, E. (eds.), *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pp. 173–184, Singapore, December 2023a. Association for Computational Linguistics (ACL). URL https://aclanthology.org/2023.genbench-1.14.

Milios, A., Reddy, S., and Bahdanau, D. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pp. 173–184, 2023b.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? In *Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., and Gao, J. Deep learning–based text classification: A comprehensive review. *ACM Comput. Surv.*, 54(3), apr 2021. ISSN 0360-0300. doi: 10.1145/3439726. URL https://doi.org/10.1145/3439726.

Nguyen, T. and Wong, E. In-context example selection with influences. *arXiv preprint arXiv:2302.11042*, 2023.

Peng, K., Ding, L., Yuan, Y., Liu, X., Zhang, M., Ouyang, Y., and Tao, D. Revisiting demonstration selection strategies in in-context learning. *arXiv preprint arXiv:2401.12087*, 2024.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. *Technical Report*, 2018.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics (ACL). doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Rubin, O., Herzig, J., and Berant, J. Learning to retrieve prompts for in-context learning. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (ACL): Human Language Technologies*, pp. 2655–2671, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.191. URL https://aclanthology.org/2022.naacl-main.191.

Shin, S., Lee, S.-W., Ahn, H., Kim, S., Kim, H., Kim, B., Cho, K., Lee, G., Park, W., Ha, J.-W., and Sung, N. On the effect of pretraining corpora on in-context learning by a large-scale language model. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Association for Computational Linguistics (ACL)*, pp. 5168–5186, Seattle, United States, July 2022. Association for Computational Linguistics (ACL). doi: 10.18653/v1/2022.naacl-main.380. URL https://aclanthology.org/2022.naacl-main.380.

Sorensen, T., Robinson, J., Rytting, C. M., Shaw, A. G., Rogers, K. J., Delorey, A. P., Khalil, M., Fulda, N., and Wingate, D. An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*, 2022.

SU, H., Kasai, J., Wu, C. H., Shi, W., Wang, T., Xin, J., Zhang, R., Ostendorf, M., Zettlemoyer, L., Smith, N. A., and Yu, T. Selective annotation makes language models better few-shot learners. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=qY1hlv7gwg.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24824–24837, 2022b.

Yang, J., Ma, S., and Wei, F. Auto-icl: In-context learning without human supervision. *arXiv preprint arXiv:2311.09263*, 2023.

Yoo, K. M., Kim, J., Kim, H. J., Cho, H., Jo, H., Lee, S.-W., Lee, S.-g., and Kim, T. Ground-truth labels matter: A deeper look into input-label demonstrations. *arXiv preprint arXiv:2205.12685*, 2022.

Zhang, Y., Feng, S., and Tan, C. Active example selection for in-context learning. *arXiv preprint arXiv:2211.04486*, 2022.

Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning (ICML)*, pp. 12697–12706. PMLR, 2021.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. V., and Chi, E. H. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WZH7099tgfM.

Contents of the Appendix

The Appendix is organized as follows:

- Appendix A discusses the related work.
- Appendix B states our experimental setup.
- In Appendix C, we provide additional information to the main paper.
- In Appendix D, we explore why self-generated demonstrations are useless.
- Appendix G contains the ablation experiments conducted on randomly selected demonstrations, as well as experiments on varying retrieval pool sizes.
- Appendix H includes the ablation experiments on varying testset sizes.
- In a supplementary experiment detailed in Appendix E, we test the "confidence" of GPT-4 towards each label.
- We describe in Appendix F whether the predictions of the different models are constrained by the label space.

A. Related work

In-context learning (ICL) is considered a property of pre-trained LLMs (Brown et al., 2020). Despite the fact that ICL does not change the parameters of the pre-trained model, ICL has demonstrated impressive outcomes. This has resulted in a flurry of papers attempting to explain the underlying causes behing the emergence of ICL and its properties. Shin et al. (2022); Wei et al. (2022a) investigate ICL from the perspective of pre-training stage. For inference stage, the importance of demonstrations for ICL was first indicated by Liu et al. (2022). Anonymous (2023) demonstrate that in addition to demonstrations, the information provided by the task description in the prompt also largely influences ICL performance. Many prompting strategies are proposed to extend the reasoning performance of LLMs such as Chain-of-thought (CoT) (Wei et al., 2022b) and Least-to-Most Prompting (Zhou et al., 2023).

Inspired by the aforementioned observations, new research focuses on the topic of demonstrations and how to effectively select them (SU et al., 2023; Li & Qiu, 2023; Nguyen & Wong, 2023; Peng et al., 2024). Liu et al. (2022) propose a demonstration retrieval method that uses K-NN for demonstration selection based on semantic distance and KATE to be the sentence encoder. Rubin et al. (2022) take another approach and train a dense retrieval which leverages an additional LM to score the retrieval. Zhang et al. (2022) leverage reinforcement learning to selection examples. However, none of the previous studies focuses on the more realistic case of large label space.

An interesting work that is concurrent and close to our is that of Milios et al. (2023b). They use the retrieval setting with extreme number of demonstrations (over 20) to obtain the best performance on ICL with large label space. This is complementary to our observations, since we focus on more realistic cases, since collecting such a large dataset requires additional cost.

In addition, in our work, we conduct comprehensive experiments on a larger breadth of LLMs including the state-of-the-art GPT family and Mixtral in this work. Our comparisons in Section 3 validate that the key insights with recent models, such as GPT-4 or Mixtral, differ when compared to prior works. We hypothesize this can be attributed to the stronger capabilities of recent models.

Beyond the retrieval setting, self-generation of demonstrations or prompts has also emerged (Kim et al., 2022; Sorensen et al., 2022; Yang et al., 2023). Li et al. (2022) developed a self-prompting method where a language model creates a pseudo QA dataset for use as retrieval pool. Li et al. (2023) leverage self-generated demonstrations combined with CoT strategy on multi-task language understanding, math reasoning task and code generating task. Contrary to the aforementioned works, we have two core differences: (i) we focus on large label space and (ii) we conduct experiments on more diverse settings.

Lastly, our work is related to works exploring how models learn the demonstrations or how demonstrations affect the final performance (Min et al., 2022; Yoo et al., 2022; Kossen et al., 2023; An et al., 2023; Hashimoto et al., 2023). Our work differs from previous efforts in that we encompass the exploration of self-generated demonstrations, rather than demonstrations annotated by humans.

B. Experimental Setup

Let us describe the concrete models and datasets used.

Models. We conduct a series of ICL classification experiments on GPT-3.5 (i.e., `gpt-3.5-turbo-0613`) and GPT-4 (i.e., `gpt-4-turbo-1106`).[3] Beyond GPT-based models, we assess the following popular open-source models: LLaMA-2 (i.e., `LLaMA-2-70b-chat`) (Touvron et al., 2023) and Mixtral 8x7B (i.e.,`Mixtral-8x7B-Instruct-v0.1`) (Jiang et al., 2024) [4]. We employ the `all-mpnet-base-v2`

[3] https://platform.openai.com/docs/models

[4] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

model in Sentence-BERT family (Reimers & Gurevych, 2019) as our retrieval model.

Datasets. We conduct experiments on four datasets: Banking77 with 77 unique labels (Casanueva et al., 2020), Clinc150 with 150 unique labels (Larson et al., 2019), HWU64 with 59 unique labels (through the concatenation of "scenario" and "intent") (Liu et al., 2019) and GoEmotions with 27 unique emotional labels plus "Neutral" label (Demszky et al., 2020). The first three are intent analysis datasets and the last one is a fine-grained sentiment analysis task dataset. All datasets have a larger label space than popular classification datasets used in previous studies (Min et al., 2022; Chen et al., 2023; Yoo et al., 2022).

For the GoEmotions dataset we excluded the "Neutral" class. This decision was based on the fact that when annotators were not certain about the emotion, they were asked to select "Neutral". We believe that excluding this ambiguous label will facilitate a more accurate and straightforward evaluation process. For HWU datasets, we use the "answer_normalised" as task input text, which standardizes the formatting of times, percentages, and values, uniformly replaces the names of persons from the original text.

We randomly sample 7 data points from each class to form a $7N$ sized retrieval pool for model retrieval. For each value of k, we randomly selected 80 test queries[5] in the test set. We repeat each experiment 3 times and calculate the mean and standard deviation. For imbalanced test sets, we perform an undersampling operation on them, reducing all the samples per class to have a uniform number of samples.

C. Supplementaries to the main text

C.1. Prompt configurations

We give a visual overview of the retrieval setting and 2LM setting in Figure S4.

C.2. Prompt visualization

In this subsection we visualize our prompt for better understanding. Different components of our prompts are illustrated in Figure S9, taking COT-2LM as an example.

C.3. Complete experimental results

In this subsection we present supplementary results which are discussed in Section 3. Accuracy of retrieval setting over all datasets on GPT-3.5, Mixtral and LLaMA-2 can be found in Figure S5. Accuracy on HWU over three settings on GPT-3.5, Mixtral, and LLaMA-2 are presented in Figure S6. Figure S7 shows the results on GPT-3.5 after modifying the descriptiveness of label space.

[5]We conduct ablation studies in Appendix H to show that experiments on test sets of different sizes present consistent results.

D. Why self-generated demonstrations are not helpful?

The experimental results in Section 3 indicate that the 2LM setting does not enhance the performance. Can this be attributed to the bias of self-generated demonstrations? We investigate this question on GPT-4, since this is the strongest-performing model. We focus on two topics: (i) the distribution over classes on the self-generated demonstrations, (ii) the match between demonstration vs test query class.

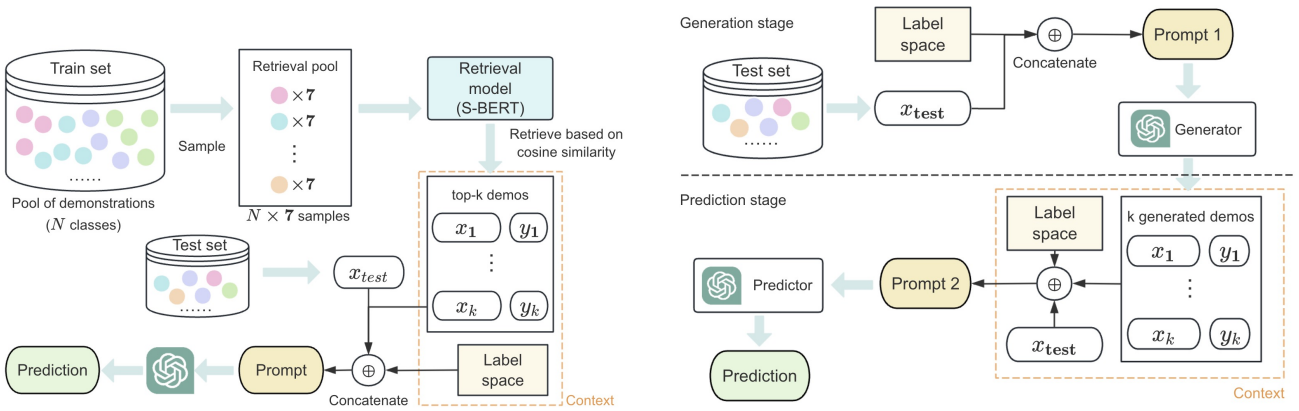
D.1. Distribution over classes

We utilize the GoEmotions dataset, since this contains a more manageable number of classes. We request the *generator* model to synthesize text-label pairs, given only the label space. The label space for this particular experiment is ordered alphabetically with the name of classes, while we repeat the experiment 40 times. The results in Figure S10 indicate that the self-generated demonstrations do not distribute evenly across the classes. Notably, the model exhibited a strong preference for certain class "joy". Despite expecting 27 unique classes based on the dataset, our 40 repetitions yield only 7 distinct classes. A similar trend is identified in LLaMA-2. In 40 repetitions of the experiment, only 13 non-repeated classes appeared in all demonstrations (we expect 27), and most of these demonstrations belonged to positive emotions, as reported in Figure S11.

Could the generation improve if we "inform" the generator model about the test query? Yes, more diverse samples are synthesized as visualized in Figure S12 when we concatenate the label space and the test query. Nevertheless, notice that the distribution is still not uniform.

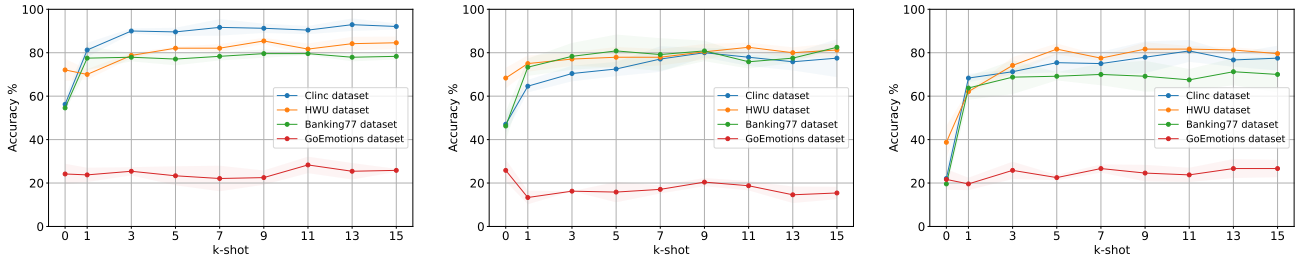
We conduct an ablation study on the ordering of the classes and assess whether we can obtain more diverse demonstration labels. Specifically, we only prompt the label space to *generator* and ask *generator* to generate one informative demonstration. Random shuffle the order of classes in the label space set each time. Figure S13 verifies that we cannot synthesize demonstrations evenly across classes. Furthermore, the *generator* has a preference for the labels that appear first in the label space when it is provided with only a randomized order of label space, but this preference becomes less apparent when the test query is provided.

Just as previous work (Lu et al., 2022) has found that the order of the demonstrations affects model performance similarly, we find that the position order in which the labels appear in the space set also affects the results generated by *generator* GPT-4 (shown in Figure S14). Given a label spaces in random order, labels of output demonstration



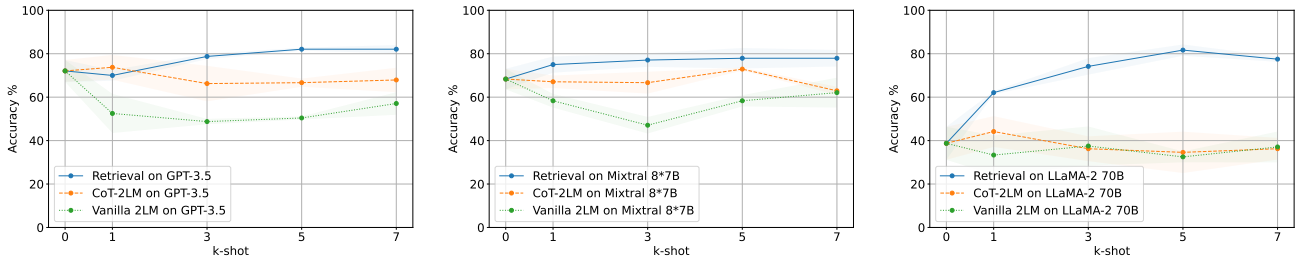
(a) Retrieval (b) 2LM

Figure S4. Overview of two prompt configurations. (a) depicts the pipeline of the *retrieval* setting under a retrieval pool of 7 demonstrations. (b) depicts the vanilla 2LM setting, which consists of two stages: a generation stage and a prediction stage. Notice that the 2LM setting does not require any 'pool of demonstrations'.



(a) GPT-3.5 (b) Mixtral 8x7B (c) LLaMA-2

Figure S5. Accuracy of retrieval setting over all datasets on (a) GPT-3.5, (b) Mixtral, (c) LLaMA-2. GPT-3.5 and Mixtral have similar performance in the HWU and the Banking77 datasets. However, in Clinc dataset, Mixtral does not benefit as much from the demonstrations when compared to GPT-3.5. Contrary to GPT-4, the number of demonstrations to reach the peak performance is less in these three models. In most cases, the first demonstration makes the largest difference, especially in the case of LLaMA-2.



(a) GPT-3.5 (b) Mixtral 8x7B (c) LLaMA-2

Figure S6. Accuracy on HWU over three settings on (a) GPT-3.5, (b) Mixtral, (c) LLaMA. Notice that GPT-3.5 and Mixtral offer similar patterns (e.g., significant improvement when CoT is used), while the performance of LLaMA-2 differs. A common pattern is that self-generated demonstrations in vanilla 2LM setting do not perform as well as the zero-shot setting.

generated by *generator* becomes more diverse compared to alphabetical order case, although still very skewed. Further, when test query is also given, the position skew is not evident, as shown in Figure S15. GPT-4 can almost overcome the sensitivity to order when self-generating context.

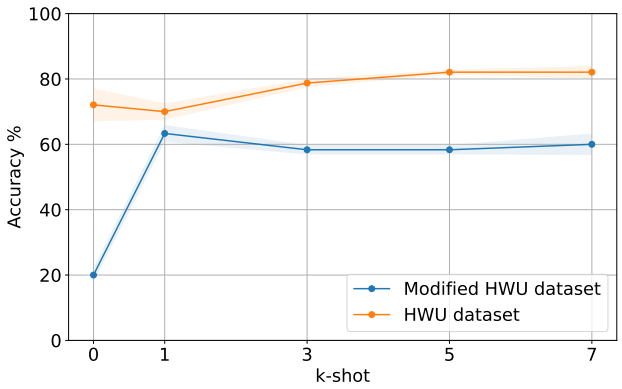


Figure S7. Prediction accuracy (mean±std %) of GPT-3.5 retrieval setting after randomly inserting alpha\digit into labels on HWU, e.g. `lists_createoradd` → `lisPts_7creyatelora1dd`

Table S1. Domain analysis of 1-shot setting for 20 correct and 20 incorrect final predictions in the Banking77 dataset. The number before the slash represents the count of 1-shot demonstrations that feature under-specified/same/different labels compared to the test query, out of the 20 correct (incorrect) samples.

Prediction	Under-specified class	In-class	Out-of-class
Incorrect	4/20	4/20	12/20
Correct	0/20	13/20	7/20

D.2. Demonstration vs test query labels

One reasonable question is whether the 2LM setting fails because the demonstrations synthesized are not matched with the test query class. In this experiment we select the Banking77 dataset, since the 2LM setting results in the largest degradation in the performance. We select 20 samples in the 1-shot setting and define three patterns (which as visualized in Figure S16) for the self-generated demonstrations. Those patterns are: (i) **In-class:** The demonstration belong to the same class as the test query, (ii) **Out-of-class:** The demonstration belongs to a different than the test query and (iii) **Under-specified class:** The demonstration is difficult to categorize.

Table S1 reports that only 20% of the incorrect prediction prompts have in-class demonstrations. However this number for correct prediction prompts is 65%. Notice that the *generator* synthesizes under-specified demonstrations approximately 20% of the time, which will confuse the predictor model. Furthermore, this weakness may be amplified in tasks involving large label space classification, where labels are more detailed and fine-grained.

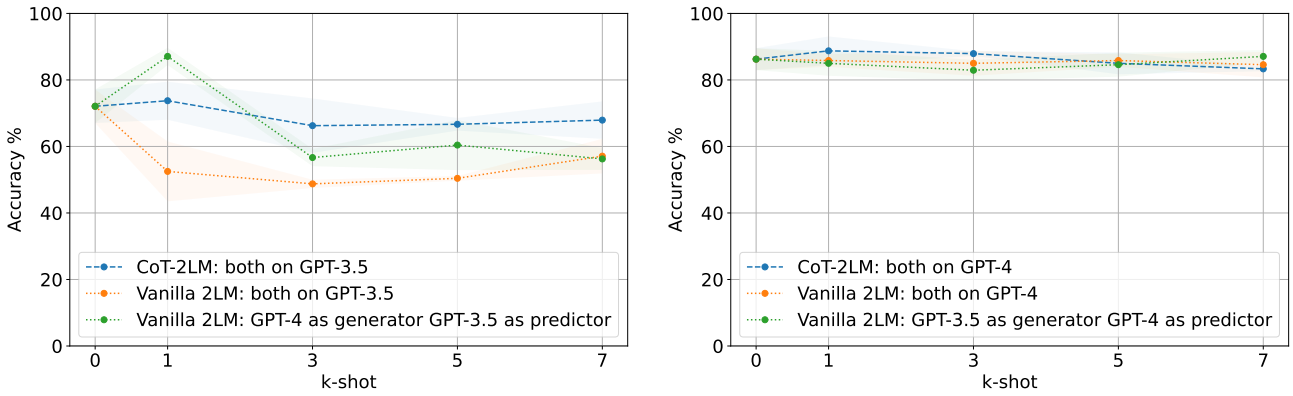
E. How confident is the model in its prediction?

One interesting question is how confident the *predictor* LM is on its predictions. We evaluate this asking GPT-4 to directly output the confidence on the retrieval setting. In our experiment, one experimental group is denoted as in-class group, in which the labels of the demonstrations in prompts are the same as the label of the test query (see the visualization examples in Figure S16), while the other group is all different labels than the test query (out-of-class). When prompting, we ask the model to sort the label space according to the probability of each label to be the right class of test query. The probability value given by the model after normalizing is considered by us as the model's confidence in that label. In the samples that were correctly predicted in both sets of experiments (where the TOP-1 label given by the model matches the ground truth), we visualize the confidence distribution of the model as Figure S17. We find that while in-class can give the model more extreme confidence (0.99), out-of-class demonstrations do not confuse the model to any great extent and the model remains very confident (with most confidence greater than 0.85).

F. Is the class prediction constrained to the label space?

Evaluating whether the prediction belongs in the label space is a challenging problem, which is particularly important in our large label space setup. In order to create a fair comparison, we take a number of measures to ensure that the output is a class from the label space. Firstly, we make sure the context includes the label set. Secondly, we ask the predictor model the following in the instruction: "Please be brief and concise, and do not say any other words except the output label, no pleasantries, no explanation". Nevertheless, Mixtral 7x8B and LLaMA-2 still give some non-compliant output, while GPT-4 and GPT-3.5 adhere to the requirements. Specifically, 37.5% of responses of LLaMA-2 prefix the label with something like "The output of the last input is:", while 9% of responses include pleasantries such as "Sure, I'd love to help you". 17% of the responses from Mixtral 8x7B provide a note in parentheses to explain the answer.

When evaluating the results, previous works determine whether the prediction is correct or not by first retrieving the label in the label space that is most similar to the model output based on semantic distance and then comparing this label with ground-truth label (Milios et al., 2023a). However, this approach can lead to a higher final accuracy than the actual accuracy because the model may generate answers that are outside the label space. In our work, we strictly compare whether the model output label matches



(a) GPT-3.5 as the predictor

(b) GPT-4 as the predictor

Figure S8. Ablation study on HWU when different model are used as the generator and the predictor. (a) GPT-4 is the generator LM and GPT-3.5 the predictor LM. When GPT-4 generates the demonstrations the results are significantly improved, especially for less than 7 demonstrations. (b) GPT-3.5 is the generator and GPT-4 the predictor. Interestingly, GPT-4 performs similarly regardless of whether the demonstrations are synthesized from GPT-3.5 or GPT-4.

the ground-truth label.

G. Ablation studies on the configurations

Baseline ICL: What happens if instead of retrieved demonstrations we use random samples from the pool of available demonstrations?

To assess this setting, we sample k data points uniformly at random and use those as the demonstrations. The result on HWU is reported in Figure S18(a) for GPT-4 and in Figure S18(b) for GPT-3.5. In GPT-4, the performance remains stable even under randomly sampled demonstrations, which is consistent with our prior observations. On the contrary, on GPT-3.5 the performance is significantly decreased when comparing with the other configurations on GPT-3.5. Notably, when there is only 1 demonstration, the performance deteriorates drastically.

Larger retrieval pool: In our experiments, the size of the retrieval pool is $7N$, which means that we have 7 demonstrations for each of the N classes. The number was selected to minimize the annotation cost without altering our key takeaways. The following question arises though: Does a larger retrieval pool necessarily lead to better performance?

We conduct an ablation experiment on HWU and Clinc to scrutinize this question. We set the size of the retrieval pool to $k \times N$, i.e., the size of retrieval pool will increase as k increases. As reported in Figure S19, the performance is almost indistinguishable from that of the fixed retrieval pool size $7N$, indicating that a larger retrieval pool is not necessarily better. Importantly, our key findings hold with the $7N$ retrieval pool.

H. Ablation studies on varying testset sizes

In order to extend our experimentation to larger demonstrations, while maintaining a reasonable number of experiments to avoid increasing our energy footprint, our main experiments are conducted on 80 randomly sampled data points. Would the results differ significantly if we used the whole testset?

We assess the impact of the testset in this ablation study using the strongest-performing model, i.e., GPT-4. We conduct experiments on HWU and Banking77 with varying size of testset. Each experiment is repeated 3 times. The results in Figure S20 indicate that the core insights remain similar even for larger testset sizes across different number of demonstrations. Therefore, in the rest of the paper, we use 80 data points for our experimentation.

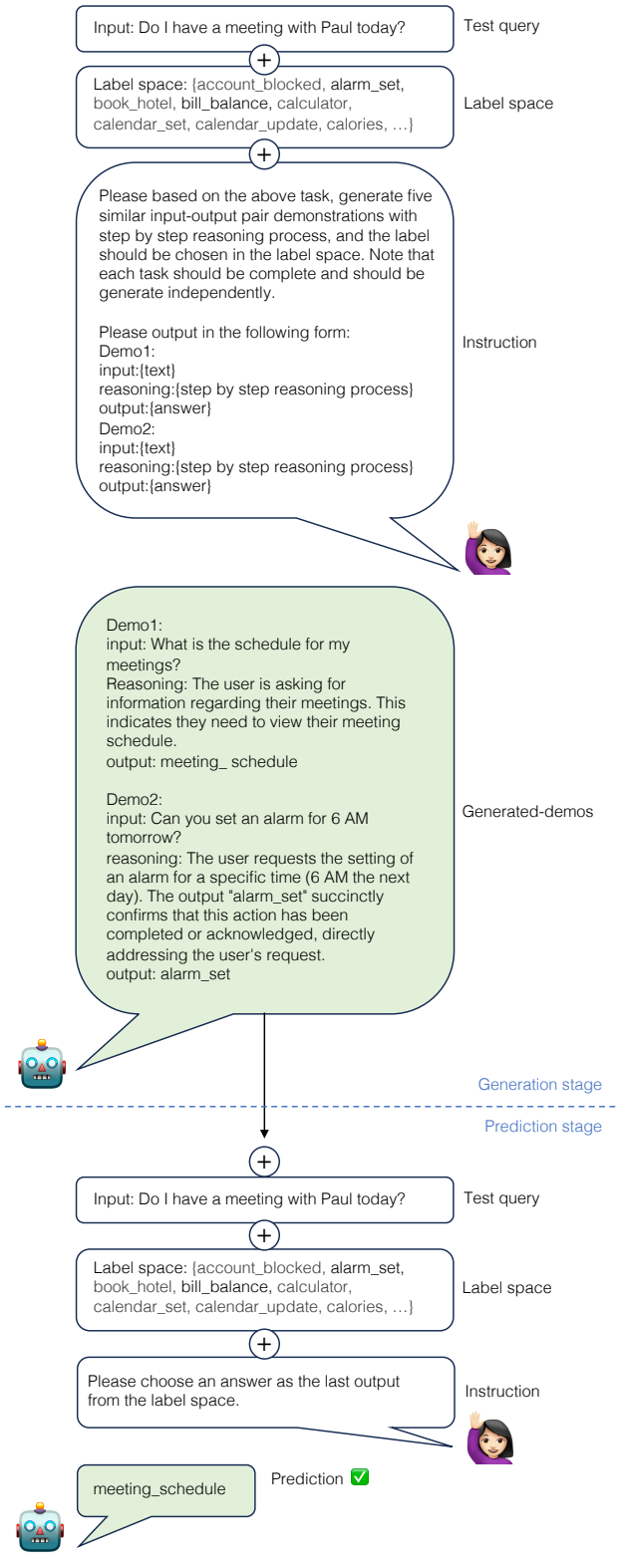


Figure S9. Different components of prompts in COT-2LM setting.

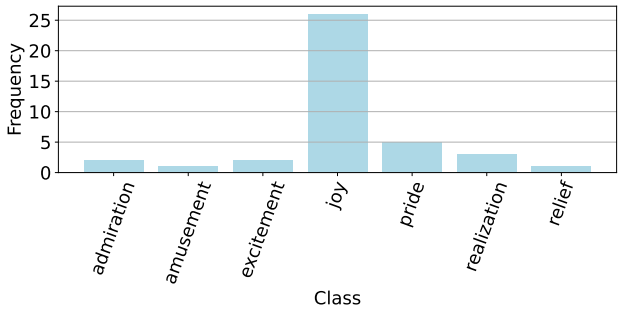


Figure S10. Distribution of the classes of demonstrations generated by *generator* given only alphabetical order label space, repeating 40 times. The model has a preference on "joy" class.

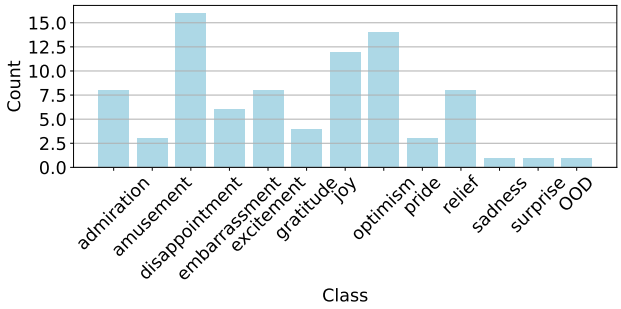


Figure S11. LLaMA-2 prefers to generate demonstrations with positive emotion labels. OOD means out-of label space.

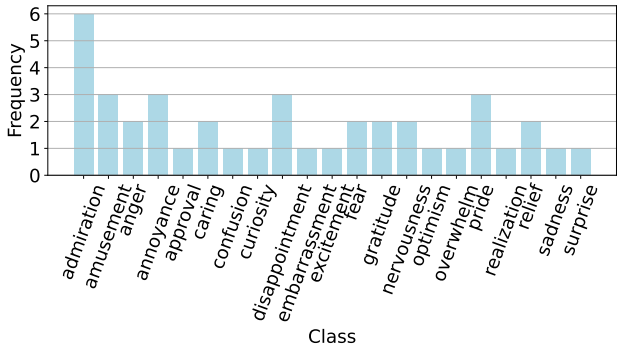


Figure S12. Distribution of the classes of demonstrations generated by *generator* given alphabetical order label space w/ test query.

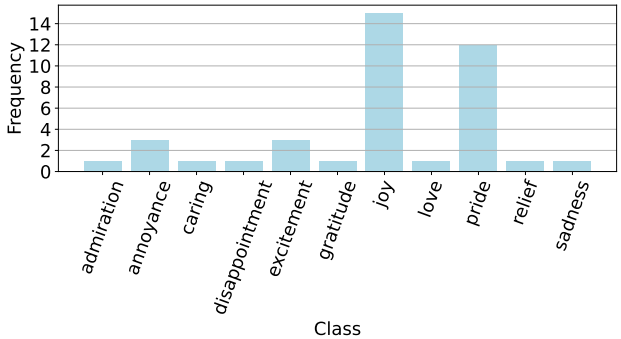


Figure S13. Distribution of the classes of demonstrations generated by GPT-4 given random order label space w/o query each time.

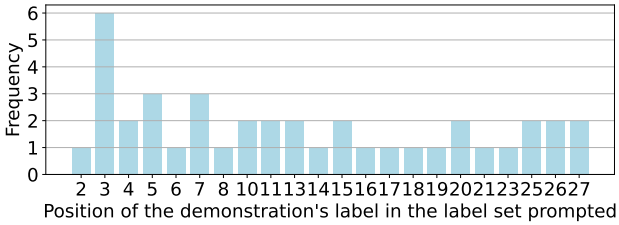


Figure S14. Distribution of label positions for GPT4-generated demonstrations when the model is given a randomly ordered label space w/o querying each time. The x-axis represents the position of the demonstration's label in the label set of the prompt.

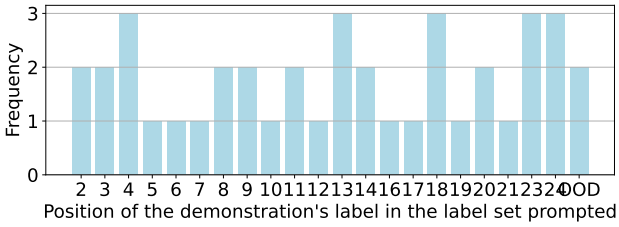


Figure S15. Distribution of label positions for GPT4-generated demonstrations when the model is given a randomly ordered label space w/ querying each time. After being provided with a test query, GPT-4 is significantly less sensitive to labeling order.

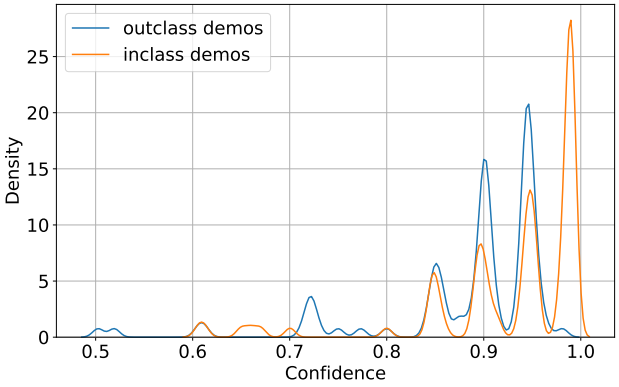


Figure S17. The x-axis represents the model's confidence in each label as the correct answer, and the y-axis represents the smoothed frequency density of these confidence.

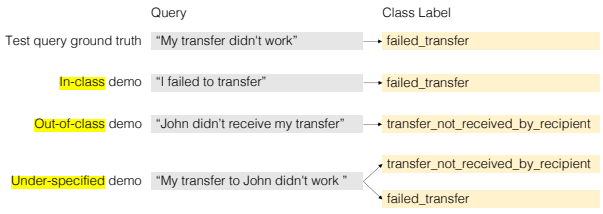
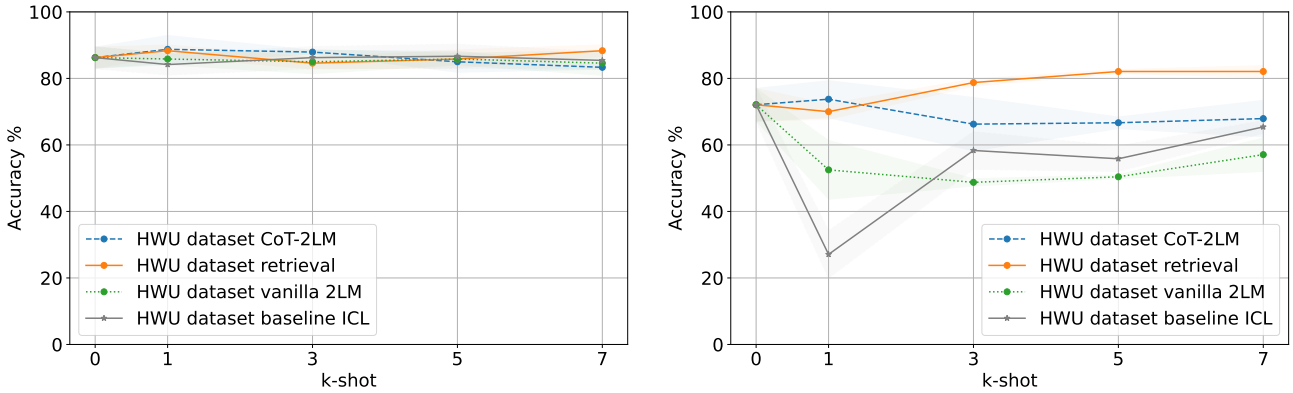


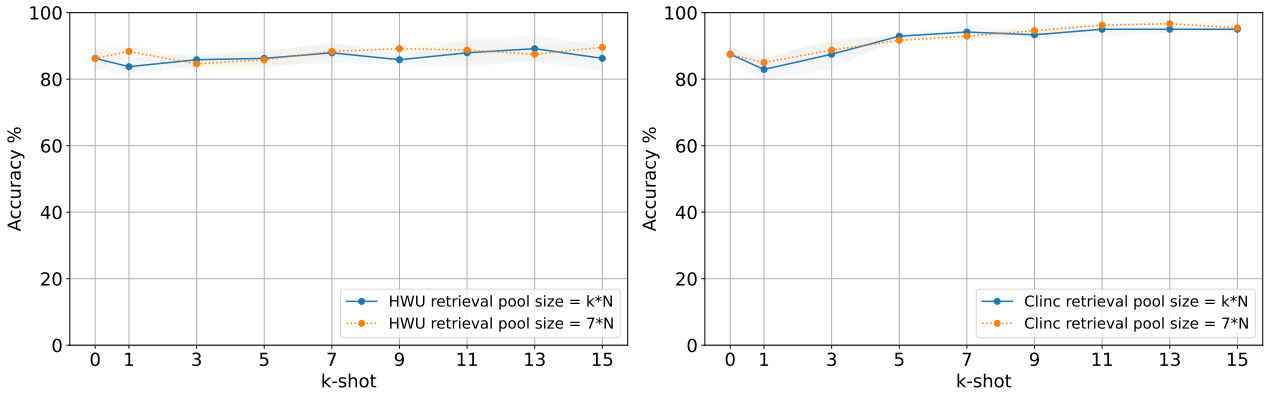
Figure S16. In-class demo, out-of-class demo and under-specified demo. The latter two patterns are more likely to lead to eventual prediction failure. We provide precise experiment detail utilizing these patterns in Table S1 and Appendix E.



(a) GPT-4 with random demonstrations

(b) GPT-3.5 with random demonstrations

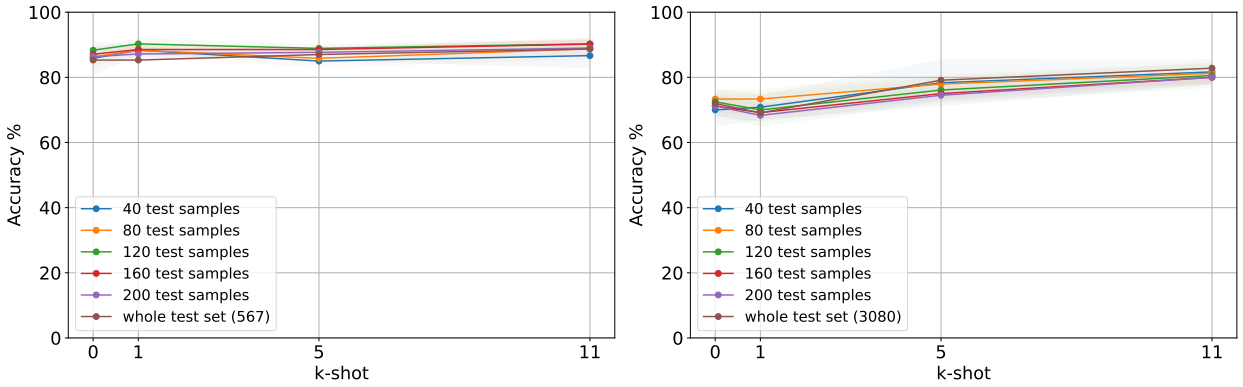
Figure S18. Ablation study on HWU. We compare the performance of the main settings with the baseline ICL, which considers using randomly selected demonstrations (gray line). On GPT-4 the performance remains similar, while on GPT-3.5 selecting randomly the demonstrations decreases the accuracy.



(a) HWU

(b) Clinc

Figure S19. Ablation study on the size of the retrieval pool size with GPT-4 on (a) HWU (b) Clinc datasets. The results indicate that the size of the retrieval pool does not have a large influence on the performance.



(a) HWU

(b) Banking77

Figure S20. Accuracy of retrieval setting with GPT-4 on testset of different sizes over (a) HWU (b) Banking77 dataset. The test set of HWU is acquired by undersampling by classes and then splitting by a 60-40 ratio. Notice the consistent trends over the different sizes of the testset.