Breaking the Bank with ChatGPT: Few-Shot Text Classification for Finance

Lefteris Loukas[1,2] and **Ilias Stogiannidis**[1,2] and **Prodromos Malakasiotis**[2] and **Stavros Vassos**[1]
[1]Helvia.ai
[2]Department of Informatics, Athens University of Economics and Business, Greece
{lefteris.loukas, ilias.stogiannidis, stavros}@helvia.ai

Abstract

We propose the use of conversational GPT models for easy and quick few-shot text classification in the financial domain using the Banking77 dataset. Our approach involves in-context learning with GPT-3.5 and GPT-4, which minimizes the technical expertise required and eliminates the need for expensive GPU computing while yielding quick and accurate results. Additionally, we fine-tune other pre-trained, masked language models with SetFit, a recent contrastive learning technique, to achieve state-of-the-art results both in full-data and few-shot settings. Our findings show that querying GPT-3.5 and GPT-4 can outperform fine-tuned, non-generative models even with fewer examples. However, subscription fees associated with these solutions may be considered costly for small organizations. Lastly, we find that generative models perform better on the given task when shown representative samples selected by a human expert rather than when shown random ones. We conclude that a) our proposed methods offer a practical solution for few-shot tasks in datasets with limited label availability, and b) our state-of-the-art results can inspire future work in the area.

1 Introduction

Virtual agents have become increasingly popular in recent years, with conversational models like GPT-3.5 (Ouyang et al., 2022) and its successor ChatGPT[1] garnering attention worldwide. While the intent detection task, as seen in the customer assistance domain, has been a well-known problem in academia for many years, it is under-explored in the financial industry due to the limited availability of datasets (Galitsky and Ilvovsky, 2019; Casanueva et al., 2020). This study aims to bridge the gap between the financial industry and the latest developments in academia.

Financial Intent	Label
It declined my transfer.	Declined Transfer
How can I trade currencies with this app?	Exchange Via App
How do your exchange rates factor in?	Exchange Rate
I just topped up, and the app denied it.	Top-up Failed
There has been a red flag on my top up.	Top-up Failed
Tell me how to replace my expired card.	Card About to Expire
...	...
My card is needed soon.	Card Delivery Estimate
What caused my transfer to fail?	Failed Transfer

Table 1: Example financial intents and their labels from the Banking77 dataset. In total, there are 77 different labels in the dataset.

In this paper, we use Banking77 (Casanueva et al., 2020), a real-life dataset of customer service intents and their classification labels. Unlike many datasets in the intent detection literature, Banking77 covers the niche of a single domain, contains a large number of labels (77), and many of the classes have tight overlaps between them, making it perfect for a business use-case scenario. Previous works have focused on fixing labeling errors (Ying and Thomas, 2022) or exploring pre-training intent representations (Li et al., 2022), which require a high level of technical expertise.

First, we demonstrate how well (and quickly) we can solve a few-shot financial text classification task using conversational GPT models. Secondly, we fine-tune other, non-generative, pre-trained models, based on MPNet (Song et al., 2020), with SetFit (Tunstall et al., 2022), a recent contrastive learning technique developed by HuggingFace which minimizes the time and samples needed to fine-tune a pre-trained model.

Our contributions include demonstrating a clever use of in-context learning with GPT-3.5 and GPT-4 to solve a challenging intent classification task. This solution is a) especially handy when rapid and accurate results are needed for few-shot tasks in financial datasets with limited label availability, and b) requires no GPUs and minimizes the need for technical expertise, which is often lacking in the banking industry. We also show that in-context

[1] https://chat.openai.com/

learning can perform better than fine-tuned masked language models (MLMs), even when presented with fewer examples. However, such solutions may be costly for small organizations due to subscription fees and often have limited token capacity, which only allows us to show the model 3 samples, for example. Lastly, we report state-of-the-art results by fine-tuning pre-trained models both when using the whole training dataset (Full-Data setting) and in a few-shot setting where only 10 training instances per class were used (10-shot setting) by employing SetFit and selecting representative samples after hiring a human expert.

2 Related Work

2.1 Studies on Banking77

Previous research papers provide important insights into improving the performance of financial intent classification models on the Banking77 dataset through the correction of label errors, the pre-training of intent representations, and the use of unattended tokens and example-driven training to improve utterance classification models. Initially, Casanueva et al. (2020) established a baseline accuracy of 93.66% by fine-tuning BERT (Devlin et al., 2019) for the Full-Data setting, and an 85.19% for the 10-shot setting by using a Universal Sentence Encoder (Cer et al., 2018) and efficient Transformer representations (Henderson et al., 2020).

Ying and Thomas (2022) aimed at reducing label errors in the Banking77 dataset through a confident learning framework (Northcutt et al., 2017, 2021) and a cosine similarity approach. Their classifiers achieved an 88.2% accuracy and 87.8% F1-Score on the original dataset, increasing to 92.4% accuracy and 92.0% F1-Score on the refined dataset.

Li et al. (2022) demonstrated that pre-training intent representations can improve intent classification, achieving an 82.76% accuracy and 87.35% Macro-F1 Score on the Banking77 benchmark. The strategy involved prefix-tuning and only fine-tuning the last layer of an LLM.

Lastly, Mehri and Eric (2021) proposed to enhance text classification models in dialog systems using observer tokens and example-driven training. The combination of these approaches resulted in an 85.95% accuracy in the 10-shot setting and 93.83% in the Full-Data setting.

Banking77 Statistics	**Train**	**Test**
Number of examples	10,003	3,080
Minimum length in characters	13	13
Average length in characters	59.5	54.2
Maximum length in characters	433	368
Minimum word count	2	2
Average word count	11.9	10.9
Maximum word count	79	69

Table 2: Dataset statistics for the Banking77 dataset. The dataset contains 10,003 examples for training and 3,080 examples for testing, with 77 different intents. Text length statistics are also provided.

2.2 Few-Shot Text Classification

Learning from just a few training instances is crucial when data collection is difficult. Interestingly, the predominant training paradigm of fine-tuning LMs exhibits poor performance in few-shot scenarios (Dodge et al., 2020), while the growing size of LMs often makes their use in this paradigm prohibitive. An alternative is to use in-context learning (Brown et al., 2020), where a generative LLM is prompted with a context and is asked to solve NLP tasks without any fine-tuning. The context typically contains a short description of the task, a few demonstrations (the context), and the instance to be classified. The intuition behind in-context learning is that the LLM has already learned several tasks during its pre-training and the prompt tries to locate the appropriate one (Reynolds and McDonell, 2021). Selecting the appropriate prompt is not trivial, though; LLMs are unable to understand the meaning of the prompt (Webson and Pavlick, 2022). This phenomenon was somewhat alleviated by fine-tuning LLMs to follow human instructions (Ouyang et al., 2022; OpenAI, 2023). Nonetheless, in-context learning is still correlated with term frequencies encountered during pre-training (Razeghi et al., 2022), while instruct-based LLMs like GPT-3.5 and GPT-4 carry the biases of the human annotators that provided the training instructions. To further deal with the difficulties of in-context learning, prompt-tuning has emerged as a promising research direction (Lester et al., 2021; Zhou et al., 2021; Jia et al., 2022).

3 Task and Dataset

Intent detection is a special case of text classification, and it has a crucial role in task-oriented conversational systems in various domains. It reflects the complexity of real-world financial and commercial systems which can be attributed to the partially overlapping intent categories, the need for fine-grained decisions, and the usual lack of data in finance (Casanueva et al., 2020; Loukas et al., 2021, 2022; Zavitsanos et al., 2022).

However, publicly available intent detection datasets are limited, and existing datasets oversimplify the task and do not reflect the complexity of real-world industrial systems (Braun et al., 2017; Coucke et al., 2018). Following the recent trends towards building robust datasets for industry-ready systems (Larson et al., 2019; Liu et al., 2019a, 2021), Banking77 (Casanueva et al., 2020) was created by PolyAI[2] as part of their study on a new intent classifier using pre-trained dual sentence encoders based on fixed Universal Sentence Encoders (Cer et al., 2018) and ConveRT (Henderson et al., 2020). In contrast to other multi-domain and broad-intent datasets, which may not capture the full complexity of each domain, Banking77 is a single-domain dataset that contains a large number (77) of fine-grained intents related to banking. Casanueva et al. believe that the dataset's single-domain focus and the large number of intents make the intent detection task more realistic and challenging. However, some intent categories partially overlap with others, requiring fine-grained decisions that cannot rely solely on the semantics of individual words, indicating the tasks's difficulty.

The dataset comprises 13,083 annotated customer service queries labeled with 77 intents and is split into two subsets: train (10,003 examples) and test (3,080 samples) (Table 2). The label distribution is heavily imbalanced in the training subset (Figure 1), demonstrating the challenge in developing classifiers in the Full-Data setting.

4 Methodology

4.1 In-Context Learning

For in-context learning, we use **GPT-3.5** (Ouyang et al., 2022) and **GPT-4** (OpenAI, 2023), which are based on the Generative Pre-trained Transformer (GPT) (Radford et al., 2018, 2019) and further trained with Reinforcement Learning from Human Preferences (RLHF) (Christiano et al., 2017) to follow instructions. GPT-3.5 is a 175B-parameter model able to consume a context o 4,096 tokens, while GPT-4 is a multi-modal model able to consume 32,768 tokens.

4.2 Fine-tuning MLMs

MPNet (Song et al., 2020) is a family of models based on the transformer architecture (Vaswani et al., 2017; Devlin et al., 2019), which adopts a novel pre-training objective that leverages the dependency among predicted tokens through permuted language modeling and takes auxiliary position information as input. MPNet is pre-trained on 160GB text corpora and outperforms other models like BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), and RoBERTa (Liu et al., 2019b) on various downstream tasks. We use a variation of MPNet, establishing it as a prominent method for our task. We use two variants of MPNet, dubbed **S-MPNet-v2**[3] and **P-MPNet-v2**.[4] Both variants were trained to identify similarities between pairs of texts which we believe allows the model to learn representations that encapsulate the more salient semantic details of the texts. Also, P-MPNet-v2 was trained with a more strict objective than S-MPNet-v2, which required both texts in the pair to have the exact same meaning.

4.3 Few-Shot Contrastive Learning

SetFit (Tunstall et al., 2022) is a few-shot learning methodology that fine-tunes a pre-trained Sentence Transformer (like S-MPNet-v2) on a small number of text pairs with contrastive learning (Chen et al., 2020). Tunstall et al. showed that using SetFit and 8 training examples has comparable performance to training models on the complete dataset.

4.4 Human Expert Annotation

Casanueva et al. (2020) identified class overlaps during the creation of Banking77. To address these challenges, we curated a subset of Banking77 for few-shot text classification with the help of a human expert who reviewed a sample of 10 examples per class and selected the top 3 examples based on their relevance to the intent they represent. This approach provided a light curation that helped avoid overlaps and ensured that each example was highly relevant to its intended intent. We expect these training instances to lead to better performance than randomly selecting training instances per class in the few-shot setting.

[2] https://github.com/PolyAI-LDN/task-specific-datasets

[3] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[4] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

5 Experimental Setup

Fine-tuning: For all of our methods, we use TensorFlow (Abadi et al., 2015) and HuggingFace (Wolf et al., 2020). For the Few-shot Experiments, we use SetFit following the developers' recommended practises.[5]

Prompt Engineering: We experimented with different prompt settings, as found in Appendix B.

In-context Learning: We utilize the OpenAI API when employing GPT-3.5.[6] Due to maximum token limitations, we use the 1-Shot setting for GPT-3.5 and the 3-shot setting for GPT-4. The prompt we use can be broken down into three parts. The first contains the description of the task and the available classes, the second provides a few examples, and the third presents the text to be classified. The prompt can be found in the Appendix A.

Note that although models like GPT-3.5 or GPT-4 can provide a quick solution without the need for technical expertise, they come at a cost as they are only accessed behind a paywall. Our experiments cost around 60$ when using GPT-3.5 ($0.002 per 1K tokens) and 1,480$ when using GPT-4 ($0.03 per 1K tokens for the 8K context model).[7]

6 Results

To understand the model's performance, we report micro-F1 (μ-F_1) and macro-F1 (m-F_1). Table 3 shows that S-MPNet-v2 achieves competitive results across all few-shot settings using SetFit. When trained on only 3 samples, it achieves scores of 76.3 μ-F_1 and 75.6 m-F_1. As we increase the number of samples, the performance improves, reaching a 91.2 micro-F1 and 91.3 macro-F1 score with 20 samples. This is only 3 percentage points (pp) lower than fine-tuning the model with all the data. Lastly, S-MPNet-v2 outperforms the previous state-of-the-art (Mehri and Eric, 2021), both in the 10-shot setting (by 2.2 pp) and in the Full-Data setting (by 0.2 pp). P-MPNet-v2 has a similar but slightly worse behavior than S-MPNet-v2.

GPT-3.5 achieves competitive results despite that it is presented with only 1 sample per class (either representative or random). It outperforms S-MPNet-v2 and P-MPNet-v2 by a large margin (over 17 pp) in the 1-shot setting, while being comparable in the 3-shot setting. As expected, using our human-curated representative samples leads to better in-context learning results. GPT-4 also shows potential for few-shot classification, outperforming all other models on the 3-shot setting by more than 6 pp. Similarly to GPT-3.5, its performance drops substantially (approximately 9 pp) when trained on random samples as opposed to when trained on the human-curated representative ones.

Methods	Setting	μ-F_1	m-F_1
Mehri and Eric (2021)	Full-Data	93.8	NA
Mehri and Eric (2021)	10-shot	85.8	NA
Ying and Thomas (2022)	Full-Data	NA	92.0
S-MPNet-v2 (ours)	Full-Data	**94.0**	**93.9**
P-MPNet-v2 (ours)	Full-Data	93.0	93.0
S-MPNet-v2	1-shot	57.4	55.9
P-MPNet-v2	1-shot	50.6	48.7
GPT-3.5 (representative samples)	1-shot	**75.2**	**74.3**
GPT-3.5 (random samples)	1-shot	74.0	72.3
S-MPNet-v2	3-shot	76.3	75.6
P-MPNet-v2	3-shot	71.4	70.9
GPT-4 (representative samples)	3-shot	**83.1**	**82.7**
GPT-4 (random samples)	3-shot	74.2	73.7
S-MPNet-v2	5-shot	83.5	83.3
S-MPNet-v2	10-shot	88.0	87.9
S-MPNet-v2	15-shot	90.6	90.5
S-MPNet-v2	20-shot	91.2	91.3
P-MPNet-v2	5-shot	79.2	79.1
P-MPNet-v2	10-shot	85.7	85.8
P-MPNet-v2	15-shot	88.4	88.4
P-MPNet-v2	20-shot	90.1	90.0

Table 3: Classification results for all models on the test data, with N-Shot indicating the number of samples used during training. All MPNet variants are fine-tuned without the `SetFit` method on the Full-Data setting.

7 Conclusion

We presented a few-shot text classification study on the financial domain. Experimenting with Banking77, a financial intent classification dataset, we showed that in-context learning with conversational LLMs can be a straightforward solution when one needs fast and accurate results in few-shot settings. In addition, we demonstrated that generative LLMs, like GPT-3.5 and GPT-4, can perform better than MLM models, even with fewer examples. While LLMs minimize the technical expertise needed or omit GPU training times, they can be considered costly for small organizations, given that LLMs can be only accessed behind a paywall (approximately 1,600$ for GPT-3.5 and GPT-4). On the other side, by fine-tuning S-MPNet-v2 with SetFit, we surpassed the previous state-of-the-art in the 10-shot setting by 2 pp. The same model also achieved state-of-the-art results in the Full-Data setting with standard fine-tuning.

[5] https://github.com/huggingface/setfit

[6] We use the `gpt-3.5-turbo` variant.

[7] https://openai.com/pricing

8 Acknowledgements

This work has received funding from European Union's Horizon 2020 research and innovation programme under grant agreement No 101021714 ("LAW GAME").

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Daniel Braun, Adrian Hernandez Mendez, Florian Matthes, and Manfred Langen. 2017. Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185, Saarbrücken, Germany. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *ArXiv*, abs/1805.10190.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Boris Galitsky and Dmitry Ilvovsky. 2019. On a chatbot conducting a virtual dialogue in financial domain. In *Proceedings of the First Workshop on Financial Technology and Natural Language Processing*, pages 99–101, Macao, China.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. 2022. Visual prompt tuning. In *Computer Vision – ECCV 2022*, pages 709–727, Cham. Springer Nature Switzerland.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference* *on Natural Language Processing (EMNLP-IJCNLP)*, pages 1311–1316, Hong Kong, China. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xianzhi Li, Will Aitken, Xiaodan Zhu, and Stephen W. Thomas. 2022. Learning better intent representations for financial open intent classification. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)*, pages 68–77, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2019a. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction - 10th International Workshop on Spoken Dialogue Systems, IWSDS 2019, Syracuse, Sicily, Italy, 24-26 April 2019*, volume 714 of *Lecture Notes in Electrical Engineering*, pages 165–183. Springer.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. 2021. *Benchmarking Natural Language Understanding Services for Building Conversational Agents*, pages 165–183. Springer Singapore, Singapore.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Lefteris Loukas, Manos Fergadiotis, Ion Androutsopoulos, and Prodromos Malakasiotis. 2021. EDGAR-CORPUS: Billions of tokens make the world go round. In *Proceedings of the Third Workshop on Economics and Natural Language Processing*, pages 13–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. FiNER: Financial numeric entity recognition for XBRL tagging. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.

Shikib Mehri and Mihail Eric. 2021. Example-driven intent prediction with observers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2979–2992, Online. Association for Computational Linguistics.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research (JAIR)*, 70:1373–1411.

Curtis G. Northcutt, Tailin Wu, and Isaac L. Chuang. 2017. Learning with confident examples: Rank pruning for robust classification with noisy labels. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'17. AUAI Press.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Accessed: 06 May 2023.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pretraining for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *ArXiv*, abs/2209.11055.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. pages 5998–6008.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. Curran Associates Inc., Red Hook, NY, USA.

Cecilia Ying and Stephen Thomas. 2022. Label errors in BANKING77. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 139–143, Dublin, Ireland. Association for Computational Linguistics.

Elias Zavitsanos, Dimitris Mavroeidis, Konstantinos Bougiatiotis, Eirini Spyropoulou, Lefteris Loukas, and Georgios Paliouras. 2022. Financial misstatement detection: A realistic evaluation. In *Proceedings of the Second ACM International Conference on AI in Finance*, ICAIF '21, New York, NY, USA. Association for Computing Machinery.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130:2337 – 2348.

A In-context Learning Prompt

You are an expert assistant in the field of customer service. Your task is to help workers in the customer service department of a company. Your task is to classify the customer's question in order to help the customer service worker to answer the question. In order to help the worker, you MUST respond with the number and the name of one of the following classes you know. If you cannot answer the question, respond: "-1 Unknown". In case you reply with something else, you will be penalized.

The classes are:
0 activate_my_card
1 age_limit
.. ..
75 wrong_amount_of_cash_received
76 wrong_exchange_rate_for_cash_withdrawal

Here are some examples of questions and their classes:
How do I top-up while traveling? automatic_top_up
How do I set up auto top-up? automatic_top_up
... ...
It declined my transfer. declined_transfer

How do I locate my card?

B Prompt Engineering

We experiment with two different prompt settings using GPT-4 in a 3-shot setting on a held-out validation subset.[8] In the first setting, we present the few-shot examples as the previous chat history. In the second setting, the few-shot examples are presented as a message from the `system`, which is one of the roles in the conversational setting of OpenAI. The second setting yielded the best results (Table 4), and we proceed to use it for the rest of our experiments. As seen in Table 4, by presenting the few-shot examples to the OpenAI API via previous chat history, we score a 77.5 μ-F_1 and a 74.4 m-F_1 score. However, presenting the examples as a `system` message hyperparameter to the API, which sets the assistant behavior, we achieve an improved μ-F_1 of 77.7 and a m-F_1 of 77.0.

Thus, we present the few-shot examples as `system` in the OpenAI later in our prompt-tuning methods (GPT-3.5 and GPT-4).

Few shot examples given as	μ-F_1	m-F_1
Previous chat history	75.5	74.4
System context	**77.7**	**77.0**

Table 4: Validation Micro-F1 and Macro-F1 scores for our two prompt settings with GPT-4 in the 3-Shot scenario.

C Class Distribution

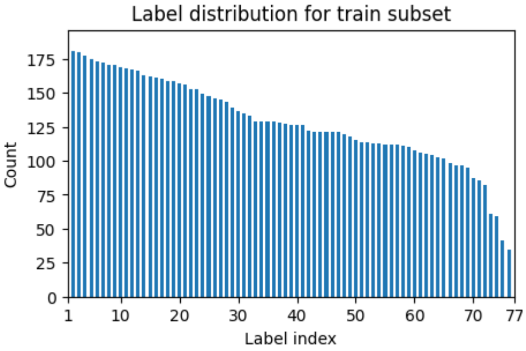

Figure 1: Class distribution of the 77 intents used over the training subset. Intent indices are shown instead of tag names for brevity.

[8] We used 5% of the training data.