Text Classification in the LLM Era - Where do we stand?

Anonymous ACL submission

Abstract

Large Language Models revolutionized NLP and showed dramatic performance improvements across several tasks. In this paper, we investigated the role of such language models in text classification and how they compare with other approaches relying on smaller pre-trained language models. Considering 32 datasets spanning 8 languages, we compared zero-shot classification, few-shot fine-tuning and synthetic data based classifiers with classifiers built using the complete human labeled dataset. Our results show that zero-shot approaches do well for sentiment classification, but are outperformed by other approaches for the rest of the tasks, and synthetic data sourced from multiple LLMs can build better classifiers than zero-shot open LLMs. We also see wide performance disparities across languages in all the classification scenarios. We expect that these findings would guide practitioners working on developing text classification systems across languages.

1 Introduction

Text classification is one of the evergreen problem in NLP and other related areas of research, with widespread applications across different real-world use cases and disciplines of study. Each classification use case is different, and collecting sufficient labeled data for each problem can be challenging. This resulted in the interest in the development of zero-shot text classification systems (Yin et al., 2019). Large Language Models can offer a solution, and their use as zero-shot (English) text classifiers (Gretz et al., 2023) has been explored in recent past. Synthetic data generation with LLMs has also been proposed to address the labeled data scarcity. Different from these two approaches, there is an established body of work on few-shot fine-tuning (e.g., Tunstall et al., 2022; Yehudai and Bandel, 2024), to address situations where we have a very small amount of labeled data, which is not typically sufficient to build a classifier using standard methods.

How does zero-shot text classification compare with few-shot fine-tuning, synthetic data based classification, and building classifiers with full labeled datasets? Although some parts of this broad question received attention in the recent past, a more comprehensive comparison is lacking in NLP research. Further, most of the past research in this direction has focused on English datasets and proprietary LLMs. A detailed comparison across different classification methods spanning more languages and datasets will not only help us understand the state of the art in text classification with LLMs, but also provide guidance to practitioners looking at solving real-world text classification use cases across various tasks and languages. We address these issues in this paper.

Concretely, we study the following research questions in this paper, considering 32 datasets covering 8 languages (4 datasets per language).

1. How well does zero-shot prompting of LLMs (open and proprietary) fare compared to building classifiers with full training data?

2. Does few-shot fine-tuning offer any benefits over zero-shot classification?

3. How well does a synthetic data based classifier fare compared to zero-shot classification with LLMs?

4. Is supervised/instruction fine-tuning of LLMs the way to go for text classification?

Starting with an overview of related work (Section 2), we proceed to the description of our methodology (Section 3) and discuss the details about the results (Section 4). After summarizing the main conclusions (Section 5), we discuss the

limitations (Section 6) and broader impacts (Section 7).

2 Related Work

Text classification, the task of classifying a given text into a pre-defined list of categories, is a well-studied problem. From bag-of-words features to the current state-of-the-art LLMs, numerous approaches have been explored in the past. Access to large amounts of human labeled data has traditionally played a significant role in improving text classifiers, and NLP research in the past two decades addressed this issue by looking at different solutions to learn from little or no labeled data.

Zero-shot Pre-LLM Approaches: Some of the earlier classification approaches relied on using only label names to build "data less" text classifier (Liu et al., 2004; Li and Yang, 2018; Meng et al., 2020; Ye et al., 2020; Gera et al., 2022) and embedding the texts and labels in a shared space (Song and Roth, 2014; Luo et al., 2021; Chu et al., 2021; Sarkar et al., 2022; Gao et al., 2023; Wang et al., 2023). Yin et al. (2019) proposed to formulate zero-shot text classification as a textual entailment problem, although Ma et al. (2021) point to the limitations of this approach in terms of variability across datasets and reliance on spurious lexical patterns. Another practical approach for zero-shot classification is cross-lingual transfer i.e., train a classification model in one or more languages, and use it as a zero-shot classifier on the target language (Wang and Banko, 2021). Except (Wang and Banko, 2021), who studied sentiment and hate speech classification tasks, all the research has focused primarily on English datasets.

Few-shot fine-tuning: Approaches that can learn from a small amount of (< 20 samples per category) labeled examples have also been explored in the recent years (Schick et al., 2020; Dopierre et al., 2021; Ohashi et al., 2021; Zhang et al., 2022). SetFit (Tunstall et al., 2022) introduced an approach based on supervised contrastive learning, transforming a language model into a topic encoder using only a few examples per label, and demonstrated effectiveness with datasets where the number of categories are low (under 5). FastFit (Yehudai and Bandel, 2024) proposed an approach that scales to many classes (50–150) effectively, and showed its usefulness with English datasets. Out of these only SetFit evaluated with a few non-English datasets.

Zero-shot Classification with LLMs: With the arrival of Large Language Models, some recent approaches explored proprietary models like GPT3.5 and GPT4 for zero-shot or few-shot in-context learning for text classification across several datasets (Gretz et al., 2023; Sun et al., 2023; Mozes et al., 2023; Tian and Chen, 2024). Extending this line of work, open LLMs were studied in the context of intent classification (Ruan et al., 2024; Arora et al., 2024) and computational social science (Mu et al., 2024). However, comparing such zero-shot approaches with few-shot and full-data based fine-tuning, (Edwards and Camacho-Collados, 2024) show that smaller, fine-tuned classifiers outperform zero-shot approaches. Whether supervised fine-tuning of LLMs offers any benefit is an unexplored question. Surprisingly, except (Tian and Chen, 2024), all these experiments have been focused only on English datasets so far. We expand this strand of work to 7 other languages, and provide more detailed comparisons across different LLMs.

Synthetic Data: One approach to address the labeled data problem is to augment existing data by creating new data by applying text transformations such as replacing synonyms, paraphrasing, back translation etc. Bayer et al. (2022) presents a detailed survey of such data augmentation techniques for text classification. An extension of this idea is to directly synthesize the labeled data using generative language models (Yu et al., 2023; Yue et al., 2023; Kurakin et al., 2023; Choi et al., 2024). In the recent past, Large Language Model based synthetic data generation is increasingly observed across different NLP tasks (Tan et al., 2024). GPT4 has been used for English (Li et al., 2023; Yamagishi and Nakamura, 2024; Peng et al., 2024) and code-mixed (Zeng, 2024) synthetic data generation for text classification with mixed results. We extend this line of work by covering more languages and exploring multiple LLMs as sources for synthetic data instead of relying on one, and extending to handle datasets with a larger label set.

Overall, we address several gaps in existing research by comparing zero-shot classification, few-shot fine-tuning, synthetic data based classification, and classification with full data together, and also study how the comparison works out once we go beyond English. In this process, we also present a comparison between different open and closed

recent LLMs.

3 Approach

We experimented with zero-shot classification, few-shot fine-tuning, and synthetic data based classification, and compared them with classifiers trained on full amount of labeled data. Our methods are described below, followed by a description of the datasets used.

3.1 Zero-shot Prompting

We compared three open LLMs - Qwen2.5-7B (Team, 2024), Aya23-8B (Aryabumi et al., 2024) and Aya-Expanse-8B (Dang et al., 2024), which is a more recent, instruction tuned version of Aya23, and one proprietary LLM - GPT4 (Achiam et al., 2023) (*gpt-4-0613*) in a zero-shot prompting setup across all languages and classification tasks. Initial experiments showed a tendency to generate a lot of explanation for the prediction despite specifying not to in the prompt. So, we controlled for the output structure using Instructor.[1] Further details on Instructor setup are mentioned in the Appendix (Figure 6). All LLMs still generated explanations beyond labeling, (as high as 10% for some open LLMs) which were treated as classification errors. All prompts were in English, as changing the language to the target language of the dataset resulted in poorer results in early experiments, which was also observed in some recent studies on other problems/datasets (Dey et al., 2024; Jin et al., 2024). We did not attempt few-shot prompting, considering the large label set with some of the datasets, but looked into few-shot fine-tuning, instead, as described below.

3.2 Few-shot fine-tuning

We performed few-shot fine-tuning with FastFit (Yehudai and Bandel, 2024) which integrates batch contrastive learning with a token similarity score to learn few-shot task specific representations for text classification. We used 10 examples per label in all cases, as that had the best result in the original FastFit paper.[2] We experimented with another few-shot fine-tuning approach SetFit (Tunstall et al., 2022) but it quickly became intractable to train for some of the datasets with >10 categories. Hence, we reported results with only FastFit in this paper. Comparisons with SetFit for the datasets with under 10 categories can be seen in the Appendix (Section B).

3.3 Synthetic Data Generation

We generated equal amounts of synthetic data from three sources - GPT4, Qwen2.5-7B and Aya-Expanse-8B, for all the classification tasks, across all languages, to ensure diversity in the generated text. Initial experiments showed that generating data from multiple LLMs was beneficial than relying on a single source, which is corroborated by recent research on other tasks (Maheshwari et al., 2024). This is also useful for controlling the costs, as the two open LLMs can be run locally on a laptop and do not incur any inference costs (and consumed less power). We used the same prompt across all LLMs, changing the task/language as needed. Details about the prompting strategy can be seen in Appendix A.

3.4 Classification with Synthetic and Real Data

We compared three approaches for text classification with real or synthetic training data, listed below:

1. A logistic regression classifier with the embedding representations from a state-of-the-art transformer model as the feature vector generator, without any further fine-tuning. We used *gte-multilingual-base* (Zhang et al., 2024), a 305 million parameter multilingual model, as our feature extractor.

2. A fine-tuned version of BERT (Devlin et al., 2019) with multilingual BERT as the base,[3] trained for 5 epochs, across all languages and datasets.

3. Instruction fine-tuning of Qwen-2.5-7B-Instruct (Yang et al., 2024; Team, 2024) for 3 epochs on the training data (10 epochs for TAXI1500, the smallest dataset) for all languages.[4]

All the classifiers were trained in two setups: first only with real data, and then only with synthetic data. More details on the experimental setup such as parameters, time taken to train, GPU requirements etc are described in the Appendix (Section A).

[1] https://python.useinstructor.com/

[2] Base model: paraphrase-multilingual-mpnet-base-v2

[3] https://huggingface.co/google-bert/bert-base-multilingual-cased

[4] Early experiments showed superior performance with Qwen compared to Aya-Expanse

3.5 Datasets and Evaluation

We experimented with four publicly available datasets and each dataset has eight language subsets for Arabic, English, French, German, Hindi, Italian, Portuguese and Spanish (i.e, 32 datasets in total) with official train-validation-test splits. Arabic and Hindi datasets are in their native scripts and all the other languages are in Roman script. Our choice of datasets primarily depended on finding all languages represented across all datasets. The datasets cover sentiment and topic classification, and are described below:

1. Multilingual twitter sentiment (Barbieri et al., 2022) which we will refer to as SENTIMENT is a dataset of tweets manually labeled with positive/negative/neutral sentiment.

2. Taxi 1500 (Ma et al., 2023) is a topic classification dataset, manually labelled with 6 categories - recommendation, faith, description, sin, grace and violence that describe sentences from the bible. The dataset covers 1500 languages in total, with a mapping between parallel sentences across bible versions that is used to build a labeled dataset from English labeled data. Some languages have multiple bibles, and we took the alphabetically first bible for that language to build our dataset.

3. Amazon Massive (FitzGerald et al., 2023) is a one million sample dataset covering 51 languages consisting of parallel virtual assistant commands classified into 60 intents spread across 18 domains ("scenario" field in the dataset). We modeled intent and scenario classification as two separate tasks, which we refer to as INTENT and SCENARIO datasets respectively.

Note that all the datasets contain short texts of different genres (tweets, bible sentences and commands to voice assistants). Table 1 shows a summary of the datasets used.

Dataset	# categories	# train	# test
SENTIMENT	3	1839	870
TAXI1500	6	860	111
SCENARIO	18	12000	2974
INTENT	60	12000	2974

Table 1: Dataset statistics per language

For synthetic data generation, we aimed to generate datasets comparable to the size of the training data for all dataset-language combinations except in the case of TAXI1500 where we generated a training set that is double the size of the original human labeled training set owing to its small size compared to others. Table 2 shows the sizes of the generated datasets and the split between different LLMs (GPT4, Aya-Expanse-8B, Qwen2.5-7B).

Dataset	# train	description
SENTIMENT	1800	200 per category, per LLM
TAXI1500	1800	100 per category, per LLM
INTENT	13500	75 per category, per LLM
SCENARIO	13500	from intent dataset

Table 2: Synthetic training data (per language)

Evaluation: We report classification accuracy as the evaluation measure in this paper. Since two of the datasets are imbalanced across categories (TAXI1500 and SCENARIO), we considered reporting macro-F1 additionally. But considering the fact that there is not much difference between the measures and the order is always preserved (i.e., if approach A gets a higher accuracy than approach B, it always has a higher macro-F1 as well), we decided to report only accuracy.

4 Results

We report results addressing the four research questions and also discuss the variation across languages and tasks in this section. Detailed per language/per task/per method results can be seen in Appendix B.

4.1 Zero-shot Classification

Figure 1 shows the zero-shot performance of various LLMs, compared to a logistic regression classifier trained with full data and embeddings based feature representation, averaged across all languages per task.

These results reveal some interesting insights: For sentiment classification, GPT4 and Aya-Expanse are better than a classifier trained on full training data. But, for the other three tasks, GPT4 is clearly much better among the zero-shot methods, although we observe that a custom classifier is much better, especially as the number of labels in the dataset increases. The difference in performance trends between sentiment classification versus other tasks we studied here may indicate

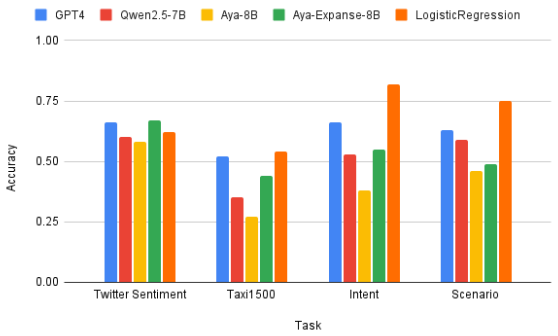


Figure 1: Zero-shot LLMs versus a logistic regression classifier trained with full data

a more subjective versus topical task difference which would warrant further scrutiny. Interestingly, all the models performed poorly on French sentiment classification compared to other languages, while Arabic was the language where most models performed the worst for other tasks (see Tables 6– 9 in Appendix B for details).

4.2 Few-shot Classification

Figure 2 shows how few-shot fine-tuning with FastFit compares with zero-shot classification and training with full data, taking GPT4 as the representative zero-shot classifier, as that was the best among the zero-shot options we explored.

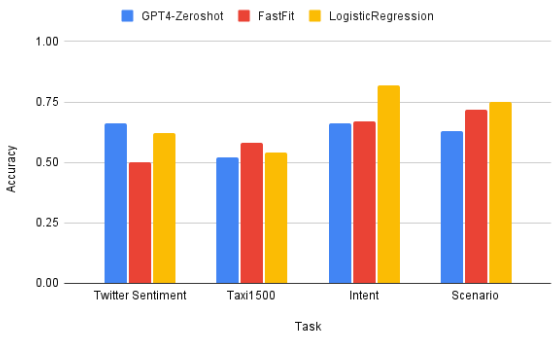


Figure 2: Few-shot Fine-tuning

The results show that while few-shot fine-tuning is not useful for sentiment classification, there is a >5% improvement over zeroshot GPT4 for two tasks (TAXI1500 and SCENARIO) and similar performance for INTENT datasets. For TAXI1500, it even results in a small improvement over training with full labeled dataset, presumably due to the contrastive learning objective used for learning the representations for fine-tuning. While there are performance disparities across languages (See Figure 7 in Appendix B for details), they are much larger for the datasets with a smaller number of categories (SENTIMENT and TAXI1500) compared to the datasets with larger number of categories (SCENARIO and INTENT).Some of this can be attributed to the fact that the datasets with more categories see a larger sample of data during fine-tuning (as we take 10 samples per category), which is probably helping the model learn the task better across languages, reducing disparities among them in terms of overall accuracy. While we would need further experiments with other datasets with many classes (covering multiple languages), we can conclude based on these results that few-shot fine-tuning can be a viable alternative to zero-shot classifiers if at least a small amount of labeled data is available.

4.3 Synthetic Data and Text Classification

We now turn to the question of the usefulness of synthetic data for text classification. Figure 3 shows a comparison between all the zero-shot LLMs and a logistic regression model trained entirely with synthetic data, averaged across languages and grouped by task.

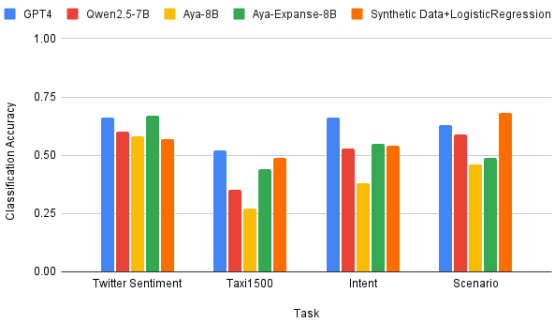


Figure 3: Zero-shot versus synthetic data based Classification

In all tasks except sentiment classification, we notice that the synthetic data based classifier either performs comparably or out-performs all the open LLMs and outperforms GPT4 too in one task (SCENARIO). We can infer that synthetic data can be considered a viable option over zero-shot classification, from these results. In practical terms, that can mean a one time cost (for building the synthetic dataset) rather than an ongoing cost of prompting a proprietary LLM as a zero-shot classifier instead.

Figure 4 compares among zero-shot, synthetic data based, and real data based classifiers, taking GPT4 as the zero-shot classifier.

We can observe from this figure that for at least

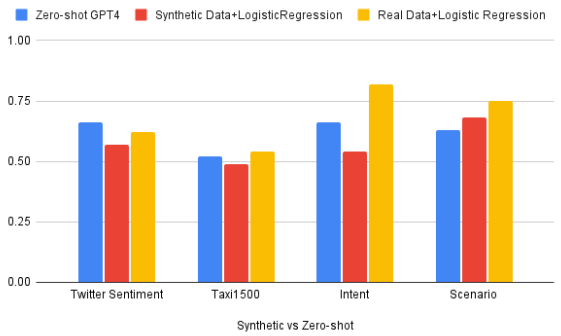


Figure 4: zero-shot GPT4, synthetic data based, and real data based classifiers

one task TAXI1500, synthetic data is performing at the same level as the other two approaches, whereas for the SCENARIO task, it show slightly better results than zero-shot GPT4. Considering Figures 3 and 4 together, we can conclude that synthetic data generated from multiple LLMs can be useful and sometimes, better than zero-shot classifiers, while being competitive compared to real-data in some cases.

One versus Many Synthetic Data Sources Since we used three sources for synthetic data generation, a natural next question to look into is what is a good source of data. To understand this, we compared different sources of synthetic data by using only one source to build a logistic regression based classifier each time. Table 3 shows the summary of these results. Note that these single-source datasets form only 1/3rd of the full dataset which uses all three sources together. Hence, we don't compare with classifiers trained on full data here, and compare only one LLM versus another as a source of synthetic data.

Task	GPT4	Aya-Expanse	Qwen2.5
TWITTER	0.5	0.47	**0.51**
TAXI1500	**0.43**	0.4	**0.43**
INTENT	**0.53**	0.48	0.40
SCENARIO	**0.65**	0.64	0.59

Table 3: Average accuracy across languages of synthetic data based classification

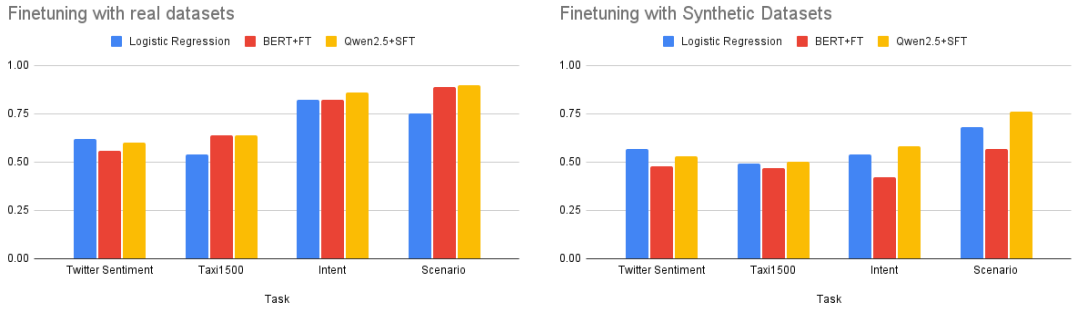
We can see that Qwen2.5 gave better results for the two datasets with smaller number of categories, but started to perform the worse of the three LLMs once we moved to datasets with larger number of categories. GPT4 consistently seems to be a reasonable source of synthetic data across all datasets. Aya-Expanse does better with datasets with larger number of categories than smaller ones. Training on the data from all sources consistently gives better results despite these differences in individual sources (Figure 3). Thus, we can conclude that multi-source generation also potentially results in more diverse data and using open LLMs as the sources of synthetic data along with GPT4 can be a cost effective way of synthetic data generation. Note that the open LLMs are both very small (7B/8B) compared to GPT4 and can be used locally on a laptop.

Classification with Real versus Synthetic Data We compared three classification approaches: a logistic regression classifier, which we used in all the above described experiments to compare against zero-shot and few-shot approaches, a classifier fine-tuned on the multilingual BERT model, and an instruction tuned classifier built from the Qwen2.5-7B-Instruct model. Figure 5 presents the performance of these classifiers using both real and synthetic datasets. With real datasets, we can see the plain logistic regression model give the best average performance for sentiment classification over all languages. It falls behind other approaches (although not dramatically) with other tasks. The Qwen2.5 and BERT fine-tuned models achieved similar accuracy across most tasks, except for the INTENT dataset, where Qwen fine-tuned model outperformed BERT by 5%. On the synthetic datasets, BERT fine-tuned model consistently had lower accuracy across tasks. The Qwen finetuned model showed the best performance for datasets with a large number of labels (INTENT and SCENARIO). In summary, our results indicate that instruction tuning is perhaps more effective for synthetic datasets and tasks with more categories, whereas logistic regression remains a strong choice for simpler tasks with fewer categories.

4.4 Performance Differences Across Languages

Across all tasks and methods, we noticed large differences in performance across languages in these experiments. Table 4 shows the performance difference between the best and worst performing languages for all methods averaged across the four tasks. Exact language specific and task specific details can be seen in the Appendix (Section B).

Zero-shot prompting, followed by synthetic data based classification had relatively less variation



(a) real data

(b) synthetic data

Figure 5: Synthetic versus Real-Data

Method	Difference
Zero-shot GPT4	13%
Zero-shot Qwen2.5-7B	15%
Zero-shot Aya-8B	13%
Zero-shot Aya-Expanse-8B	11%
FastFit	22%
Logistic Regression-Real	29%
BERT-FT-Real	23%
SFT-Real	17%
Logistic Regression-Synthetic	17%
BERT-FT-Synthetic	12%
SFT-Synthetic	17%

Table 4: Average (Absolute) Performance Difference Across Languages for the different methods

across languages, although they are still in the 10-20% range. When we had some labeled data for all languages, we still had $> 20\%$ performance difference for few-shot finetuning and BERT fine-tuning, whereas a logistic regression classifier has an almost 30% average difference between best and worst performing languages. In specific cases, such as few-shot finetuning on TAXI1500, the difference between the best performing (French-69%) and least performing (Arabic-30%) was nearly 40%. Even when there was full access to training data, BERT fine-tuning had $> 40\%$ performance difference between Italian (35%) and English (77%) for TAXI1500.

Although this dataset has the least amount of training data, we observe similar trend in datasets where large amounts of training data is available. For example, the logistic regression classifier on SCENARIO dataset has 58% accuracy with Arabic, but 90% with English. All these may point to the insufficiency of the base multilingual embedding models in representing data across languages. It is also possible that the difference in the script and the writing direction is a contributing factor in some cases (e.g., Arabic).

While the variance across languages is lesser with synthetic data, as we noticed earlier, multilingual synthetic data generation comes with other challenges. For example, generation is much slower for languages with a different script (Arabic and Hindi in our data). The average number of tokens was also higher (in some cases, 3 times higher) for non-Roman script based languages. This directly influences the costs involved in synthetic data generation, especially with proprietary LLMs, and may impose limitations on their use as synthetic data generators for problems involving data from non roman script based languages.

5 Discussion

We compared different ways of performing text classification (zero-shot classification, few-shot fine-tuning, using full labeled data, and using synthetic data) across 32 datasets (8 languages, 4 datasets per language). Returning to our original research questions, our findings are summarized below:

RQ1: *How well does zero-shot prompting of LLMs (open and proprietary) fare compared to building classifiers with full training data?*

Zero-shot classifiers perform well in terms of accuracy, but primarily for datasets with fewer categories, especially SENTIMENT, where GPT4 and an open LLM Aya-Expanse-8B achieve comparable results which are better than training with full human labeled data. In all other cases, while GPT4 has the best performance among the zero-shot LLMs, it trails behind classifiers built with

labeled datasets, especially as the number of categories increases. With high amounts of labeled data, even a logistic regression classifier with text embedding representations performs much better, and is of course less resource and cost intensive.

RQ2: *Does few-shot fine-tuning offer any benefits over zero-shot classification?*

Few-shot fine-tuning generally offers higher accuracy (upto 10% gain) than zero-shot classifiers on 3 out of 4 tasks. However, for the SENTIMENT task, which has fewer categories and can be perceived as a more subjective compared to topic classification, GPT-4 outperforms few-shot finetuning. Overall, considering the fast training and inference times, without any added costs, few-shot finetuning can be considered a reliable option as the number of categories increases, in the absence of sufficient labeled data.

RQ3: *How well does a synthetic data based classifier fare compared to zero-shot classification with LLMs?*

In all tasks except sentiment classification, the synthetic data-based classifier either performs on par (INTENT) or achieves a 5-10% improvement compared to zero-shot classification using open LLMs. It outperforms GPT-4 too in one task (SCENARIO). These results indicate that synthetic data can be a viable alternative to zero-shot classification with open LLMs when number of categories are more, as it involves a one-time cost for creating the dataset. In terms of what is a better source of synthetic data, in most cases, GPT4 and at least one open LLM achieve comparable performance as the single source of synthetic data. Over all, the best results are achieved by combining all data sources, which can also be a cost-effective solution.

RQ4: *Is supervised fine-tuning of LLMs the way to go for text classification?*

Excluding sentiment classification, we see that the best performance is with either BERT fine-tuned or an instruction fine-tuned classifier with real data, and supervised fine-tuning does better than other approaches with synthetic data. However, it has to be noted that supervised fine-tuning is compute intensive both for training and inference across all languages and tasks (See Table 5 in the appendix for details on the time taken). On the other hand, we also notice a strong performance with a simple logistic regression based classifier too in some cases, although as the number of categories increases, the advantage seems to wane away.

Based on these results, we can summarize the following guidelines for practitioners:

1. For sentiment classification, zero-shot classification with LLMs is a better option than task specific fine-tuning.

2. In all other cases, few-shot fine-tuning achieves better performance than zero-shot classification. So, collecting a handful of labeled data is useful.

3. Synthetic data based classifiers perform better than zero-shot classification with open LLMs, but are not always better than GPT4. However, sourcing data from multiple LLMs is useful.

4. Despite all the recent advances, training classifiers using high-quality labeled data still gives the best performance, and SFT gives the best performance, especially when dealing with a large number of categories, and even a logistic regression classifier can give a strong performance in some cases when such datasets are available.

While the datasets and languages covered are by no means exhaustive, these results provide some guidance on what methods work for what kinds of data, what to expect in terms of language disparities, and how to work with synthetic data from multiple sources, across multiple languages. Future directions can include increasing the coverage of languages, and expanding to multi-label datasets, to draw more comprehensive conclusions about LLMs and the task of text classification.

6 Limitations

While the study allows us to draw a few concrete conclusions based on our experiments, it is not free from limitations. Firstly, we have limited ourselves to one prompt for zero-shot models and one prompt for synthetic data generation. While the use of instructor for structured output generation in the case of zero-shot classification automatically adjusts the prompt during retries, essentially taking care of prompt engineering, we cannot steer the internal prompt creation process ourselves. Few-shot prompting (instead of few-shot fine-tuning) was also not explored, as the benefits are unclear with the increasing number of labels in some datasets. Few-shot fine-tuning can be sample sensitive, and

while we did not notice any variations across different runs, we did not systematically explore that aspect. We also did not explore multilingual classifiers and all our classifiers are monolingual, built on top of pre-trained multilingual models. No language-specific choices were made (e.g., using a English embedding model may give better performance for English). We also did not do any qualitative analysis of the results or of the generated synthetic data.

Since our goal is to compare broadly across different approaches, an extensive evaluation of fine-tuning options or parameter variations was not performed, nor did we repeat the experiments with different initializations, to keep the number of experiments (and costs involved) in check. Additionally, all the datasets deal with only short texts (tweets, sentences, voice assistant commands) and hence, the results of this study might not extend to long texts. Finally, the question of potential data contamination is inevitable while discussing the zero-shot performance. While we don't know the specifics of the training data for various LLMs, considering that the performance across all tasks (except sentiment classification) is lower than models relying on full training data, perhaps it is not a serious concern for these experiments. In terms of the languages covered, while there is some typological diversity, 7/8 languages belong to the Indo-European language family (covering three sub-families: Germanic, Italic and Indo-Aryan). Thus, the extendability of these conclusions to other languages and language families is not guaranteed. Finally, the open LLMs we explored are much smaller in size (7B-8B parameters) compared to GPT4, and the results should not be seen as a verdict against the use of open LLMs for text classification.

7 Ethics and Broader Impact

We have used publicly available datasets and did not do any experiments involving human participants. We have also used small, locally run LLMs for several experiments, and most of the experiments are run locally on a laptop (details in the appendix), thus, consuming less power and perhaps with less carbon footprint than fine-tuned models or larger LLMs that require GPUs for training and/or inference. We do not foresee any harms due to the approaches described in this paper and it is only helpful for those working on text classification in the real-world to give a more realistic perspective about working with LLMs, which can potentially save time/money in a short/long run in terms of choosing the solution space to explore. We also share all the code and generated synthetic datasets as supplementary material to support reproducible research.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Gaurav Arora, Shreya Jain, and Srujana Merugu. 2024. Intent detection in the age of LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1559–1570, Miami, Florida, US. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, et al. 2024. Aya 23: Open weight releases to further multilingual progress. *arXiv preprint arXiv:2405.15032*.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Juhwan Choi, Yeonghwa Kim, Seunguk Yu, JungMin Yun, and YoungBin Kim. 2024. UniGen: Universal domain generalization for sentiment classification via zero-shot dataset generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1–14, Miami, Florida, USA. Association for Computational Linguistics.

Zewei Chu, Karl Stratos, and Kevin Gimpel. 2021. Unsupervised label refinement improves dataless text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4165–4178.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

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.

Krishno Dey, Prerona Tarannum, Md Arid Hasan, Imran Razzak, and Usman Naseem. 2024. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings. *arXiv preprint arXiv:2410.13153*.

Thomas Dopierre, Christophe Gravier, and Wilfried Logerais. 2021. A neural few-shot text classification reality check. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 935–943, Online. Association for Computational Linguistics.

Aleksandra Edwards and Jose Camacho-Collados. 2024. Language models for text classification: Is in-context learning enough? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10058–10072, Torino, Italia. ELRA and ICCL.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Lingyu Gao, Debanjan Ghosh, and Kevin Gimpel. 2023. The benefits of label-description training for zero-shot text classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13823–13844, Singapore. Association for Computational Linguistics.

Ariel Gera, Alon Halfon, Eyal Shnarch, Yotam Perlitz, Liat Ein-Dor, and Noam Slonim. 2022. Zero-shot text classification with self-training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1119, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shai Gretz, Alon Halfon, Ilya Shnayderman, Orith Toledo-Ronen, Artem Spector, Lena Dankin, Yannis Katsis, Ofir Arviv, Yoav Katz, Noam Slonim, et al. 2023. Zero-shot topical text classification with llms-an experimental study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9647–9676.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM on Web Conference 2024*, pages 2627–2638.

Alexey Kurakin, Natalia Ponomareva, Umar Syed, Liam MacDermed, and Andreas Terzis. 2023. Harnessing large-language models to generate private synthetic text. *arXiv preprint arXiv:2306.01684*.

Ximing Li and Bo Yang. 2018. A pseudo label based dataless naive bayes algorithm for text classification with seed words. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1908–1917.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461.

Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S Yu. 2004. Text classification by labeling words. In *Aaai*, volume 4, pages 425–430.

Qiaoyang Luo, Lingqiao Liu, Yuhao Lin, and Wei Zhang. 2021. Don't miss the labels: Label-semantic augmented meta-learner for few-shot text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2773–2782.

Chunlan Ma, Ayyoob ImaniGooghari, Haotian Ye, Renhao Pei, Ehsaneddin Asgari, and Hinrich Schütze. 2023. Taxi1500: A multilingual dataset for text classification in 1500 languages. *arXiv preprint arXiv:2305.08487*.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796.

Gaurav Maheshwari, Dmitry Ivanov, and Kevin El Haddad. 2024. Efficacy of synthetic data as a benchmark. *arXiv preprint arXiv:2409.11968*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017.

Maximilian Mozes, Jessica Hoffmann, Katrin Tomanek, Muhamed Kouate, Nithum Thain, Ann Yuan, Tolga Bolukbasi, and Lucas Dixon. 2023. Towards agile text classifiers for everyone. In *Findings of the Association for Computational Linguistics: EMNLP 2023*,

pages 400–414, Singapore. Association for Computational Linguistics.

Yida Mu, Ben P Wu, William Thorne, Ambrose Robinson, Nikolaos Aletras, Carolina Scarton, Kalina Bontcheva, and Xingyi Song. 2024. Navigating prompt complexity for zero-shot classification: A study of large language models in computational social science. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12074–12086.

Sora Ohashi, Junya Takayama, Tomoyuki Kajiwara, and Yuki Arase. 2021. Distinct label representations for few-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 831–836.

Letian Peng, Zilong Wang, and Jingbo Shang. 2024. Incubating text classifiers following user instruction with nothing but LLM. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3753–3766, Miami, Florida, USA. Association for Computational Linguistics.

Qian Ruan, Ilia Kuznetsov, and Iryna Gurevych. 2024. Are large language models good classifiers? a study on edit intent classification in scientific document revisions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15049–15067, Miami, Florida, USA. Association for Computational Linguistics.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2022. Exploring universal sentence encoders for zero-shot text classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 135–147, Online only. Association for Computational Linguistics.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 28.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ke Tian and Hua Chen. 2024. ESG-GPT:GPT4-based few-shot prompt learning for multi-lingual ESG news text classification. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing*, pages 279–282, Torino, Italia. Association for Computational Linguistics.

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 preprint arXiv:2209.11055*.

Cindy Wang and Michele Banko. 2021. Practical transformer-based multilingual text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 121–129, Online. Association for Computational Linguistics.

Yau-Shian Wang, Ta-Chung Chi, Ruohong Zhang, and Yiming Yang. 2023. PESCO: Prompt-enhanced self contrastive learning for zero-shot text classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14897–14911, Toronto, Canada. Association for Computational Linguistics.

Yosuke Yamagishi and Yuta Nakamura. 2024. UTRad-NLP at #SMM4H 2024: Why LLM-generated texts fail to improve text classification models. In *Proceedings of The 9th Social Media Mining for Health Research and Applications (SMM4H 2024) Workshop and Shared Tasks*, pages 42–47, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhiquan Ye, Yuxia Geng, Jiaoyan Chen, Jingmin Chen, Xiaoxiao Xu, SuHang Zheng, Feng Wang, Jun Zhang, and Huajun Chen. 2020. Zero-shot text classification via reinforced self-training. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 3014–3024.

Asaf Yehudai and Elron Bandel. 2024. Fastfit: Fast and effective few-shot text classification with a multitude of classes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 174–184.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. 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 3914–3923.

Yue Yu, Yuchen Zhuang, Rongzhi Zhang, Yu Meng, Jiaming Shen, and Chao Zhang. 2023. ReGen: Zero-shot text classification via training data generation with progressive dense retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11782–11805, Toronto, Canada. Association for Computational Linguistics.

Xiang Yue, Huseyin Inan, Xuechen Li, Girish Kumar, Julia McAnallen, Hoda Shajari, Huan Sun, David Levitan, and Robert Sim. 2023. Synthetic text generation with differential privacy: A simple and practical recipe. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1321–1342, Toronto, Canada. Association for Computational Linguistics.

Linda Zeng. 2024. Leveraging large language models for code-mixed data augmentation in sentiment analysis. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 85–101, Miami, Florida, USA. Association for Computational Linguistics.

Haoxing Zhang, Xiaofeng Zhang, Haibo Huang, and Lei Yu. 2022. Prompt-based meta-learning for few-shot text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1342–1357, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

A Details about the Experimental Settings

Compute Infrastructure: All the zero-shot prompting, few-shot fine-tuning, and logistic classifier training, and synthetic data generation experiments were performed on an Apple M1 Pro laptop with 32GB memory. Open LLMs were downloaded and used locally using Ollama (`https://ollama.com/`), and the OpenAI model was called with their API. BERT-Finetuning for the datasets with smaller number of categories (SENTIMENT and TAXI1500) was done locally and for the other two datasets, it was done on a Google Colab T4 GPU. Instruction fine-tuning was performed on a V100 GPU. Transformers[5] and Unsloth[6] were used for BERT-finetuning and instruction fine-tuning of Qwen2.5-7B respectively. All the implementation code is provided in the supplementary material.

Zero-shot Prompting: Zero-shot prompting controlling for the output structure was performed using Instructor (`https://python.useinstructor.com/`), which utilizes Pydantic (`https://docs.pydantic.dev/`) for efficient data validation. The code snippet to prompt an LLM with instructor is shown in Figure 6 below. The variable catnames contains the list of labels. More details on how structured output generation works can be seen in the code submitted as supplementary material and in the documentation for the library Instructor.

```python
class ClassificationResponse(BaseModel):
    label: Literal[tuple(catnames)] = Field(
        ...,
        description="The predicted class label.",
    )

def classify(data: str) -> ClassificationResponse:
    """Perform single-label classification on the input text."""
    return client.chat.completions.create(
        model=myconfig['ollama_model'], #qwen2.5:7b or aya:8b
        response_model=ClassificationResponse,
        messages=[
            {
                "role": "user",
                "content": f"Classify the following text: <text>{data}</text>",
            },
        ],
    )
```

Figure 6: Prompt for Zero-shot Classification

Synthetic Data Generation: The prompt used for synthetic data generation is specified as follows, where *text_lang*, *text_genre*, *task_desc* and list_of_cats[i] come from config files and are task-dataset specific. For example, for generating sentiment classification data for Arabic, text_lang = arabic, text_genre = tweets, task_desc = sentiment classification, list_of_cats = positive or negative or netural.

[5] `https://github.com/huggingface/transformers`
[6] `https://github.com/unslothai/unsloth`

```
prompt = f"""generate a {text_lang} language
text that looks like {text_genre}
that can be categorized as {list_of_cats[i]}
in the context of {task_desc}.
The generated text should be under 50 words,
and ensure some diversity of vocabulary
in the generated texts.
            """
```

Instruction fine-tuning: The following instruction format was used for fine-tuning the Qwen2.5-7B model:

```
''' <s>[INST] Consider the text:
    "{input_text}" Please select the
    most relevant category for the
    given text from following OPTIONS:
    {all_categories}.
    CHOICE: {response} </s>
'''
```

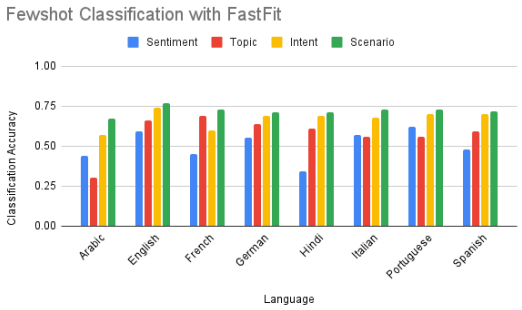
Parameter settings: We used evaluation loss as the metric for selecting the best model. Some training parameters are presented in Table 5. Instruction tuning was done for 3 epochs each for SENTIMENT, INTENT and SCENARIO datasets, and 10 epochs for TAXI1500 dataset which has a smaller amount of training data compared to the rest and did not converge in 3 epochs.

Model	epochs	learning rate	weight decay	GPU/CPU hours
FastFit	10	5e-5	0.01	5-15 min for training; fast inference
BERT	5	5e-5	0.01	0.5 hr for training; fast inference
Qwen2.5-7B	3	1e-5	0.001	6hr for training; 3hr for inference

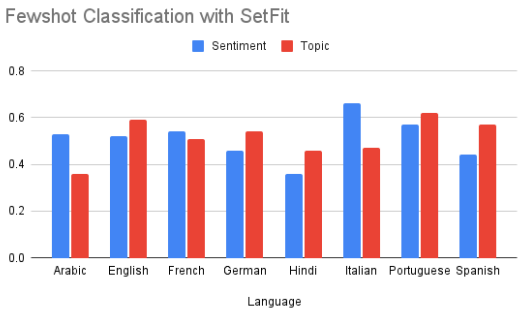
Table 5: Parameter setting for fine tuning

More details can be seen in the code provided as supplementary material.

B Additional Results



(a) FASTFIT



(b) SETFIT

Figure 7: Few-shot Fine-tuning Across Tasks and Languages

	Zero-shot				Few-shot	Real Data			Synthetic Data		
Language	GPT4	Aya:8B	Qwen2.5:7B	Aya-Expanse:8B	FastFit	Logistic Reg.	BERT-FT	SFT	Logistic Reg.	BERT-FT	SFT
Arabic	**0.68**	0.59	0.58	**0.68**	0.44	0.61	0.45	0.65	0.51	0.5	0.53
English	**0.71**	0.65	0.68	**0.71**	0.59	0.65	0.65	0.58	0.59	0.52	0.51
French	0.54	0.52	0.47	0.58	0.45	0.65	**0.68**	0.57	0.61	0.49	0.45
German	0.67	0.62	0.65	0.7	0.55	0.7	0.68	**0.73**	0.57	0.47	0.6
Hindi	**0.62**	0.53	0.56	0.61	0.34	0.41	0.33	0.36	0.47	0.38	0.46
Italian	**0.72**	0.58	0.6	0.71	0.57	0.59	0.59	0.64	0.61	0.49	0.59
Portuguese	0.69	0.6	0.61	0.69	0.62	**0.7**	0.54	0.68	0.58	0.49	0.59
Spanish	0.68	0.59	0.62	**0.69**	0.48	0.65	0.56	0.61	0.61	0.49	0.52

Table 6: All Results for TWITTER sentiment classification task

	Zero-shot				Few-shot	Real Data			Synthetic Data		
Language	GPT4	Aya:8B	Qwen2.5:7B	Aya-Expanse:8B	FastFit	Logistic Reg.	BERT-FT	SFT	Logistic Reg.	BERT-FT	SFT
Arabic	0.58	0.22	0.23	0.44	0.3	0.36	0.5	**0.69**	0.32	0.48	0.61
English	0.53	0.35	0.4	0.43	0.66	0.74	**0.77**	0.69	0.54	0.51	0.33
French	0.54	0.23	0.34	0.41	0.69	0.57	**0.73**	0.67	0.55	0.5	0.62
German	0.44	0.32	0.33	0.41	0.64	0.54	**0.72**	**0.72**	0.44	0.43	0.57
Hindi	0.52	0.27	0.35	0.44	0.61	0.5	**0.64**	0.62	0.46	0.41	0.57
Italian	0.49	0.27	0.32	0.38	0.56	0.56	0.35	**0.6**	0.48	0.49	0.43
Portuguese	0.47	0.3	0.41	0.45	0.56	0.55	**0.69**	0.56	0.58	0.48	0.47
Spanish	0.56	0.23	0.4	0.52	0.59	0.47	**0.72**	0.56	0.52	0.48	0.37

Table 7: All Results for TAXI1500 Bible topic classification task

	Zero-shot				Few-shot	Real Data			Synthetic Data		
Language	GPT4	Aya:8B	Qwen2.5:7B	Aya-Expanse:8B	FastFit	Logistic Reg.	BERT-FT	SFT	Logistic Reg.	BERT-FT	SFT
Arabic	0.6	0.31	0.46	0.49	0.57	0.74	0.79	**0.81**	0.44	0.3	0.51
English	0.72	0.44	0.6	0.6	0.74	0.87	0.88	**0.89**	0.58	0.43	0.62
French	0.67	0.4	0.54	0.56	0.6	0.83	0.86	**0.87**	0.58	0.46	0.56
German	0.66	0.37	0.54	0.56	0.69	0.8	0.84	**0.86**	0.51	0.43	0.58
Hindi	0.66	0.39	0.5	0.52	0.69	0.83	0.82	**0.86**	0.57	0.45	0.59
Italian	0.68	0.41	0.55	0.56	0.68	0.83	0.85	**0.86**	0.57	0.43	0.57
Portuguese	0.66	0.38	0.54	0.55	0.7	0.83	0.86	**0.88**	0.56	0.44	0.59
Spanish	0.66	0.37	0.54	0.54	0.7	0.84	**0.86**	0.85	0.54	0.44	0.61

Table 8: All Results for INTENT classification task

	Zero-shot				Few-shot	Real Data			Synthetic Data		
Language	GPT4	Aya:8B	Qwen2.5:7B	Aya-Expanse:8B	FastFit	Logistic Reg.	BERT-FT	SFT	Logistic Reg.	BERT-FT	SFT
Arabic	0.59	0.43	0.52	0.45	0.672	0.58	**0.86**	**0.86**	0.58	0.54	0.696
English	0.67	0.5	0.66	0.52	0.77	0.9	0.9	**0.94**	0.74	0.54	0.77
French	0.64	0.47	0.59	0.49	0.73	0.81	**0.9**	**0.9**	0.7	0.59	0.72
German	0.64	0.46	0.6	0.5	0.71	0.79	**0.9**	**0.91**	0.68	0.59	0.82
Hindi	0.62	0.45	0.54	0.48	0.71	0.55	**0.87**	**0.87**	0.7	0.57	0.79
Italian	0.63	0.46	0.59	0.49	0.73	0.8	**0.89**	**0.89**	0.68	0.61	0.75
Portuguese	0.63	0.46	0.61	0.5	0.73	0.81	0.9	**0.93**	0.67	0.57	0.76
Spanish	0.63	0.46	0.6	0.48	0.72	0.79	0.89	**0.9**	0.68	0.56	0.74

Table 9: All Results for SCENARIO classification task