Comparing Specialised Small and General Large Language Models on Text Classification: 100 Labelled Samples to Achieve Break-Even Performance

Branislav Pecher[♠†], **Ivan Srba**[†], **Maria Bielikova**[†]
[♠] Faculty of Information Technology, Brno University of Technology, Brno, Czechia
[†] Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
{branislav.pecher, ivan.srba, maria.bielikova}@kinit.sk

Abstract

When solving NLP tasks with limited labelled data, researchers can either use a general large language model without further update, or use a small number of labelled examples to tune a specialised smaller model. In this work, we address the research gap of how many labelled samples are required for the specialised small models to outperform general large models, while taking the performance variance into consideration. By observing the behaviour of fine-tuning, instruction-tuning, prompting and in-context learning on 7 language models, we identify such performance break-even points across 8 representative text classification tasks of varying characteristics. We show that the specialised models often need only few samples (on average $10 - 1000$) to be on par or better than the general ones. At the same time, the number of required labels strongly depends on the dataset or task characteristics, with this number being significantly lower on multi-class datasets (up to 100) than on binary datasets (up to 5000). When performance variance is taken into consideration, the number of required labels increases on average by $100 - 200\%$ and even up to 1500% in specific cases.

1 Introduction

With the introduction of the GPT-3 model (Brown et al., 2020), large language models have been shown to be an effective generalist models for learning with limited labelled data. They are able to perform well across many NLP tasks with no (using prompting) or only few (using in-context learning) labelled samples and without any parameter update (Qin et al., 2023; Sun et al., 2023; Dong et al., 2022). This performance is achieved by conditioning the model on an appropriate text input (prompt) containing instructions for the given task, a test sample for which to generate output and optionally a set of few in-context examples showcasing the task (Sun et al., 2023; Dong et al., 2022). Many enhancements were proposed to improve the overall few-shot behaviour. For example prompt-tuning and automatic prompt-engineering, where the effective prompts are designed automatically, or instruction-tuning, where language models are tuned to better follow the task instructions (Gao et al., 2021; Logan IV et al., 2022; Dong et al., 2022).

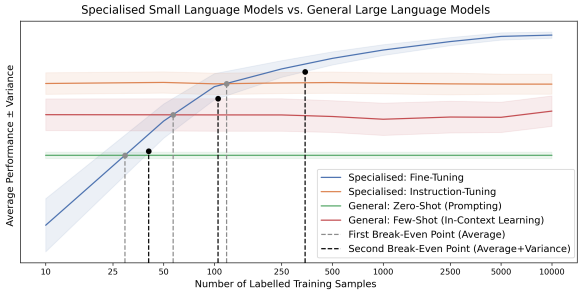

Figure 1: Comparison between the performance of specialised small and general large language models. The break-even points are identified by observing the impact of changing the number of available labelled samples and taking performance variance into consideration. The results are aggregated across the language models relevant for each approach and across 8 datasets (from the individual results presented in Figure 2). Specialised models outperform general ones with only few labelled samples (up to 100), with performance variance showing strong impact on the comparison, increasing the number significantly.

With enough labelled samples, the *specialised* smaller models, obtained through fine-tuning or instruction tuning, can achieve performance on par or better than the *general* large language models used with prompting or in-context learning without further parameter update (Schick and Schütze, 2021; Qin et al., 2023). From practical perspective (e.g., to decide how many labelled samples are needed or to choose the best approach when the number of labels is fixed), it is valuable to know (at least approximately) the number of labelled samples needed to achieve such superior performance,

i.e., to identify performance break-even points.

So far, most of studies comparing the performance of specialised and general language models have been done using different (often non-representative and incomplete) settings and methodologies, leading to divergent findings. The specialised language models are tuned either on the *whole labelled set* (where they achieve higher performance than general language models) (Schick and Schütze, 2021; Qin et al., 2023), or using the same set of *few samples* (e.g., $4 - 32$) used for general models (where they achieve significantly worse performance than general models) (Ma et al., 2023). Only few works study the behaviour of models between the few samples and the whole set, and observe the performance break-even points where specialised language models outperform general ones. Moreover, they often focus only on specific approaches or ignore the performance variance (caused by various sources of randomness, such as random seeds) (Le Scao and Rush, 2021; Hongjin et al., 2022; Hernandez et al., 2023; Gupta et al., 2023).

Our goal is to remedy these shortcomings and answer the following question: *"How many labelled samples do the **specialised smaller models** need to outperform more **general larger models**?"* To achieve this, we investigate and observe how the performance of different approaches changes when increasing the number of labelled training samples. We identify the break-even points between performance of specialised models and more general models, while taking the performance variance into consideration (aggregated results are presented in Figure 1). Our main contributions and findings are[1]:

- We perform a comprehensive investigation of the impact of increasing the number of labelled training samples on performance and its variance of fine-tuning, prompting, in-context learning and instruction-tuning approaches across 7 language models and 8 representative text classification datasets of various characteristics, such as number of classes, text length or task type.
- By identifying break-even points between the specialised and general models, we find that the smaller specialised models (obtained through fine-tuning or instruction-tuning), require only small number of labelled training examples ($10 - 1000$) to achieve performance on par or even better than general large language models used in zero/few-shot settings. The required number of labelled samples is dependent on the dataset characteristics, with binary datasets and datasets that require better language understanding (e.g., question answering) requiring significantly more labelled samples for the specialised models to outperform the general ones. In addition, we find significant impact of performance variance, especially originating from in-context learning or fine-tuning on few samples, increasing the number of required labelled samples by up to 1500%, with average increase of $100\% - 200\%$.
- Based on the observed findings, we provide practical recommendations on how to choose the best approach, model or to decide how many additional data to label considering the available annotation and computation budget.

[1]To support replicability and extension of our results, we openly publish the source code of our experiments at https://anonymous.4open.science/r/investigation-with-interactions-F6D4

2 Related Work

Extensive focus is dedicated to the evaluation and comparison between different data efficient approaches utilising large language models, such as prompting, in-context learning, fine-tuning, prompt-tuning or prompt-based/instruction tuning (Dong et al., 2022; Liu et al., 2023). The comparisons usually focus on a specific setting and methodologies, such as different model sizes, used approaches or number of labelled samples, and often lead to divergent findings. In some comparisons, the performance of specialised small models is compared with significantly larger general models (Schick and Schütze, 2021; Ma et al., 2023; Liu et al., 2022; Hernandez et al., 2023; Hongjin et al., 2022), while in others the setting is kept as similar as possible (i.e., comparing specialised and general models of the same sizes) (Mosbach et al., 2023; Qin et al., 2023; Logan IV et al., 2022; Gao et al., 2021; Le Scao and Rush, 2021; Gupta et al., 2023).

Majority of the comparisons focus on comparing fine-tuned models with in-context learning (Ma et al., 2023; Hongjin et al., 2022; Hernandez et al., 2023; Qin et al., 2023), or on comparing prompt-based/instruction-tuned models on the specific tasks with their general counterparts (Liu

et al., 2022; Mosbach et al., 2023; Schick and Schütze, 2021). Only a limited number of papers compare fine-tuning with instruction-tuning (Gupta et al., 2023; Le Scao and Rush, 2021) or multiple approaches at the same time (i.e., fine-tuning, instruction-tuning, prompting and in-context learning at the same time) (Logan IV et al., 2022; Gao et al., 2021).

If enough labelled samples are used for the tuning, the smaller models, specialised either through fine-tuning or instruction-tuning, can achieve performance on par with, or in many cases even better than the performance of the larger general models (using prompting or in-context learning) (Schick and Schütze, 2021; Qin et al., 2023; Hernandez et al., 2023). At the same time, in the extremely limited settings, where the models are fine-tuned using the same low number of samples as it is used for in-context learning, the general large language models excel over the small specialised models (Ma et al., 2023; Hongjin et al., 2022). Other papers study the impact of varying the training data sizes on the performance of specialised models and their comparison with general models. However, they often either perform a comprehensive comparison across multiple approaches, but focus only on few samples and vary the data sizes only to up 32 samples (Logan IV et al., 2022; Gao et al., 2021), or focus on small subset of approaches while using a larger part or the whole dataset, but often in specific domains (Le Scao and Rush, 2021; Gupta et al., 2023; Hongjin et al., 2022; Hernandez et al., 2023; Ma et al., 2023). Only in specific cases the sensitivity of the compared approaches to the sources of randomness (e.g., choice of samples or their order for in-context learning) or to more systematic choices (e.g., choice of prompt format) and the performance variance these factors introduce is taken into consideration (Ma et al., 2023).

Compared to these works, we focus on more comprehensive comparisons across: 1) full training dataset, increasing the size from 10 samples to full dataset in exponential fashion; 2) multiple approaches for obtaining specialised models (fine-tuning, prompt-based/instruction-tuning) and for using the general models (prompting, in-context learning); and 3) multiple runs to carefully take into consideration the sensitivity of different approaches and the performance variance this sensitivity introduces.

3 Investigating the Impact of Dataset Size

Approaches. We investigate the impact of training dataset size across a set of currently popular approaches for dealing with limited labelled data in NLP: 1) *fine-tuning*; 2) *prompting*; 3) *in-context learning*; and 4) *instruction-tuning*. For fine-tuning, we use the **BERT** (Devlin et al., 2019) and **RoBERTa** (Liu et al., 2019) base models. For prompting and in-context learning the **Flan-T5** (Chung et al., 2024) base, **ChatGPT** (3.5-turbo-0613 version) (Brown et al., 2020; Ouyang et al., 2022), **LLaMA-2** (Touvron et al., 2023) 13B chat optimised model, **Mistral-7B** (Jiang et al., 2023) and **Zephyr-7B** (Tunstall et al., 2023) models. For instruction-tuning we use the **Flan-T5** (full instruction-tuning), **Mistral-7B** and **Zephyr-7B** models (both instruction-tuned using LoRA (Hu et al., 2021)).

Datasets. The investigation covers 8 classification datasets composed of tasks with different number of classes and characteristics. We focus on 4 binary datasets from GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) benchmarks: *SST2* (Socher et al., 2013) for sentiment classification, *MRPC* (Dolan and Brockett, 2005) for determining semantic equivalence relationship between two sentences, *CoLA* (Warstadt et al., 2019) for determining the grammatical acceptability of a sentence, and *BoolQ* (Clark et al., 2019) for question answering. In addition, we use 4 multi-class text datasets: *AG News* (Zhang et al., 2015) for news classification (4 classes), *TREC* (Voorhees and Tice, 2000) for question classification (6 classes), *SNIPS* (Coucke et al., 2018) for intent classification (7 classes) and *DB Pedia* (Lehmann et al., 2015) for topic classification (14 classes).

Experimental setup. For each dataset, we split all the available labelled samples into training, validation and testing set using 60:20:20 split. The investigation is done across a scale of available labelled samples, starting with 10 and increasing exponentially up to the full dataset (or a dataset fraction where fine-tuning outperforms all other approaches). Each experiment is repeated 100 times for fine-tuning and 20 times for the remaining approaches (due to the significant costs of inference or training of the larger language models) in order to observe and reduce the performance variance (as recommended by Gundersen et al. (Gundersen et al., 2023) and Pecher et al. (Pecher et al.,

2023)), since particularly in-context learning and fine-tuning with few samples were identified to produce results with significant variance due to effects of randomness (Lu et al., 2022; Zhao et al., 2021; Zhang et al., 2022; Mosbach et al., 2021; Dodge et al., 2020). Each repeat uses the same fixed random seed for all models to guarantee deterministic and replicable results. The random seed covers all sources of randomness (Gundersen et al., 2022) – the split of data, selection of labelled samples, initialisation of models, order of data in training, non-deterministic operations (e.g., dropout), and selection of samples for in-context learning. At each point, we report the mean F1 macro and standard deviation. For fine-tuning, we follow the typical setup and recommendations from related work: adding dropout layer with rate of 0.3 followed by a classification layers to the pre-trained transformers, using learning rate of 1e-5 with AdamW optimiser with warmup and training the model until convergence using a maximum of 10 epochs and early stopping, with variable batch size across different sizes (starting at 4 and ending at 32 for the full dataset). The prompt format used for prompting, in-context learning and instruction tuning is included in Table 1. It is a result of a simple prompt engineering based on prompt formats used in related work. In addition, we set all the general models to be deterministic (e.g., no sampling and no beam search) and set maximum number of tokens for generation to 10. For in-context learning, we randomly select 2 samples for each class. Instruction tuning is done using learning rate of 1e-5 for 5 epochs using AdamW optimiser with warmup, batch size of 4 and early stopping, with a maximum of 250 steps per epoch for the Flan-T5 model (as we observed severe overfitting of Flan-T5 without this constraint on larger dataset sizes).

Table 1: Prompt formats and verbalisers used for prompting, in-context learning and instruction-tuning across different datasets. The *[Class 1-N]* and the *[Output]* are replaced with the names of the classes defined by the verbaliser. The *[Input]* is replaced by the sentence of the samples. The *[Input]* and *[Output]* are repeated for each in-context sample, while the final *[Output]* is used to determine the predicted class.

Prompt Format
Determine { sentiment | grammatical acceptability | semantic equivalence | whether the passage contains answer | topic | intent } of the { sentence | sentence pair | question } using following options: 1) *[Class 1]* 2) *[Class 2]* ... N) *[Class N]*. *[Input]* *[Output]*

Dataset	**Verbaliser**
SST2	{Negative, Positive}
MRPC	{No, Yes}
CoLA	{No, Yes}
BoolQ	{No, Yes}
AG News	{World, Sports, Business, Science and Technology}
TREC	{Expression, Entity, Description, Human, Location, Number}
SNIPS	{Playlist, Weather, Event, Musing, Creative Work, Rate Book, Book Restaurant}
DB Pedia	{Company, Educational Institution, Artist, Athlete, Office Holder, Transportation, Building, Natural Place, Village, Animal, Plant, Album, Film, Written Work}

3.1 Comparison of Specialised and General Models

In this section, our goal is to answer the following main research question: ***RQ1:*** *How do the specialised models compare to general models on average as the number of labelled samples increases?* To answer this research question, we identify the first break-even points (average performance) in the performance of specialised (fine-tuning, instruction-tuning) models in comparison to general ones (prompting, in-context learning), but also with each other. These break-points specify the point after which the performance of fine-tuning is better on average, but may still be lower due to performance variance, such as when the randomness is not sufficiently addressed (low number of runs is used or runs are cherry-picked). The aggregated results are presented in Figure 1, with the results for individual models and datasets presented in Figure 2.

Specialised models can outperform the general ones using only small number of labelled examples. To outperform zero-shot setting (prompting), fine-tuning approaches require on average between $10 - 1000$ labelled samples. In case of few-shot setting (in-context learning), the number of required labelled samples is higher, on average between $100 - 5000$ samples.

The instruction-tuned models (small or

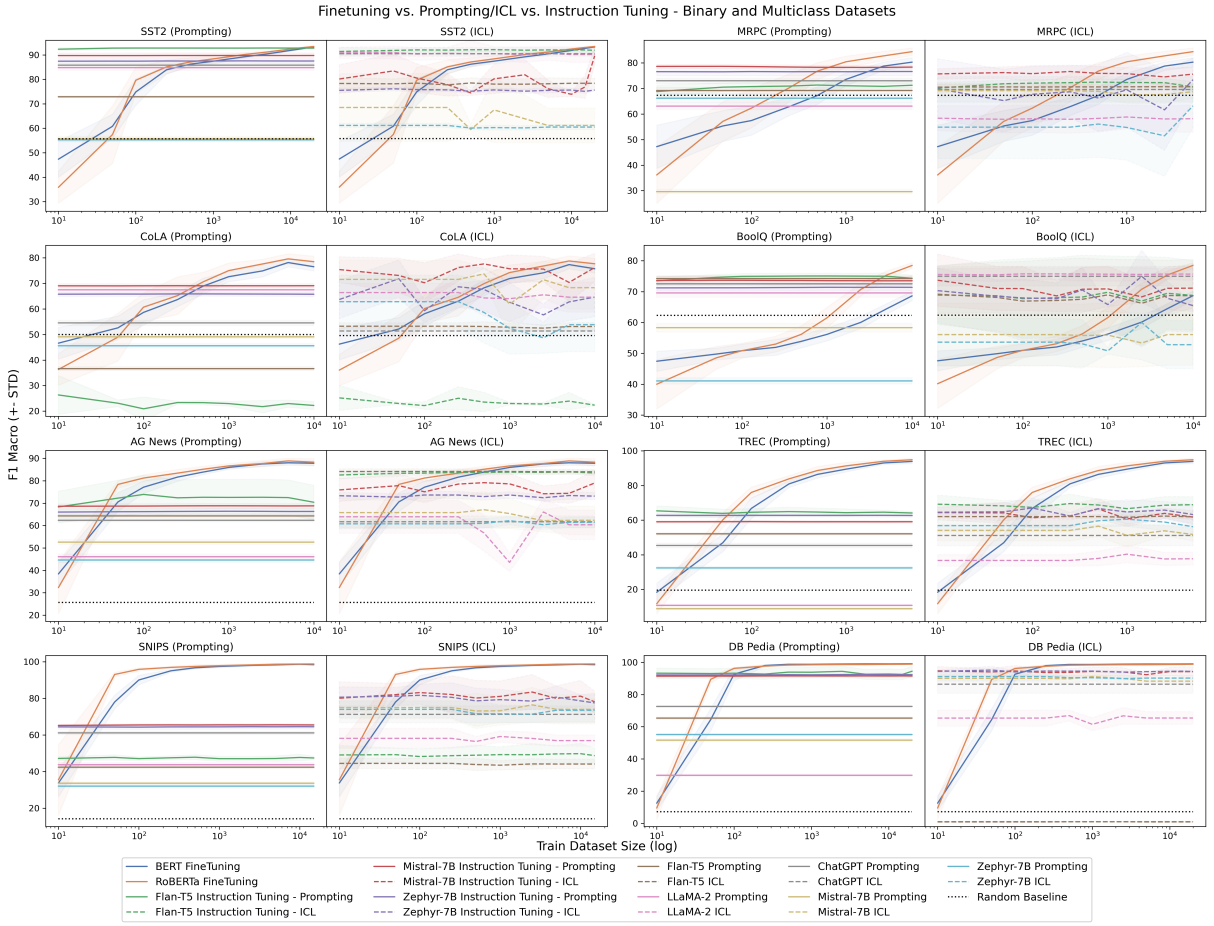

Figure 2: The impact of varying size of available labelled training samples (in logarithmic scale) on the performance of fine-tuning, prompting, in-context learning and instruction-tuning approaches across the binary (SST2, MRPC, CoLA, BoolQ) and multi-class datasets (AG News, TREC, SNIPS, DB Pedia), reported using F1 macro and its standard deviation. Although the impact is dataset dependent (with binary datasets requiring more labelled samples), we can observe that **fine-tuning can outperform large language models used in zero/few-shot setting when trained only small number of labelled samples** ($10 - 1000$). On the other hand, **instruction-tuning** of small (Flan-T5) or medium (Mistral/Zephyr) language models **with only 10 labelled samples** and using them in zero/few-shot setting often **leads to performance on par, or better**, than the one in much larger language models (LLaMA-2/ChatGPT) used in zero/few-shot setting without additional tuning.

medium sized) **represent a good balance between the generality of the models and the samples required for specialisation.** To outperform these models, fine-tuning approaches often require a majority of the labelled dataset. In addition, we observe a consistent performance of instruction-tuned models, regardless of how many labelled samples are used. The performance achieved by instruction-tuned models with 10 labelled samples is close to the one achieved with the full labelled dataset. Furthermore, they consistently outperform all the general larger counterparts. As such, instruction-tuned models achieve the almost-best performance with only a fraction of labelled samples required by fine-tuning, making them an ideal specialised models in many cases.

However, **the number of required labelled samples and the comparison between models is heavily dependent on the dataset and task characteristics**, such as how many classes are used (binary vs. multi-class setting), length of sentences (e.g., SST2 vs. CoLA), whether the task requires working with a single sentence or a paired/multiple inputs (e.g., SST2/CoLA vs. MRPC/BoolQ), or the overall type of the task (e.g., simple sentiment classification vs. question answering) that is defined by these characteristics. **The largest difference can be observed between binary and multi-class datasets**. On multi-class datasets, fine-tuning requires only up to 100 labelled samples to outperform any of the models (even instruction-tuned ones). On binary datasets the required sam-

ples are as high as 6000 to outperform zero and few-shot setting, or up to 16000 to outperform the instruction-tuned models. The remaining characteristics have lower, less consistent impact. For example, the task type on binary datasets significantly affects the samples, with BERT on BoolQ not being able to outperform majority of the remaining approaches even with full dataset, but has almost no effect on multi-class datasets. At the same time, some datasets cause significant problems to specific models. For example instruction-tuning of Flan-T5 on CoLA dataset, prompting of Mistral-7B and LLaMA-2 models on TREC dataset, Zephyr-7B on BoolQ dataset, or Mistral-7B on MRPC dataset, or in-context learning of Flan-T5 model on DB-Pedia dataset.

Similarly to instruction-tuning, **general models show consistent performance regardless of the number of labelled samples.** Only in-context learning shows slight fluctuations in performance. However, this fluctuation is caused only indirectly by the number of labelled samples, as it affects the choice of in-context examples (i.e., what samples are available). As such, only the performance of fine-tuning consistently improves as more labelled samples are used. In addition, we **do not observe consistent improvement of in-context learning (few-shot setting) over prompting (zero-shot setting).** Instead, the performance difference is more dataset dependent. On the multi-class datasets, in-context learning is beneficial over prompting in majority of the cases (only exception is Flan-T5 on DB-Pedia and ChatGPT on AG News). On the binary datasets, the benefit of in-context learning strongly depends on the models. For example on BoolQ dataset, Zephyr-7B benefits from in-context learning while Mistral-7B does not. At the same time, the improvement of in-context learning over prompting is minor in some cases (such as the increase in Flan-T5 on the SNIPS dataset from 42.5% to 44.5%).

3.2 Impact of Performance Variance on Comparison

In this section, we answer following research question: ***RQ2:*** *How does the variance from repeated runs affect the comparison between specialised and general models?* Our goal is to determine how the number of required samples increases when the performance variance is taken into consideration. To accomplish this, we identify the second break-even points (average+variance) denoting "worst-case sizes" of training sets. They represent the point after which fine-tuning shows better performance, even when the variance is taken into consideration (i.e., there is low probability that the performance would be worse when using low number of runs). The results from this investigation are presented in Figure 2 (aggregated in Figure 1) and the comparison of first and second break-even points for the best performing models of different sizes in Table 2.

Performance variance has a significant effect on the break-even points between models, increasing the number of required labelled samples by a large amount. On average, fine-tuning requires $2-3$ times more labelled samples (increase of $100-200\%$) to outperform the remaining approaches when taking the variance into consideration. In specific cases, the impact of variance is significantly lower (negligible), as we observe a 0% increase in the number of required samples (e.g., RoBERTa and LLaMA2/ChatGPT prompting and in-context learning on AG News dataset). At the same time, the impact of variance is significantly higher in other cases, leading to an increase of up to 1500%, or an increase where even with the full labelled dataset the second break-even point is not achieved (e.g., BERT and RoBERTa in comparison to Mistral/Zephyr in-context learning on the CoLA dataset).

However, **the impact of the performance variance strongly depends on the dataset and the overall number of labelled samples required for the first break-even point.** Looking at the increase in absolute numbers instead of a percentual increase, an increase of $100-200\%$ represents the need to annotate only between $10-750$ more samples in some cases (e.g., most multi-class datasets, such as BERT or RoBERTa in comparison to Flan-T5 in-context learning on AG News dataset), while representing an increase of $2000-6500$ or more labelled samples in other cases (e.g., binary datasets, such as RoBERTa in comparison to Mistral/Zephyr in-context learning on BoolQ dataset, or BERT in comparison to LLaMA2/ChatGPT in-context learning on SST2 dataset).

In addition, **the second break-even point heavily depends on the variance of the approaches/models used.** In case of prompting, which shows the lowest performance variance, the highest observed increase is 233% (RoBERTa with LLaMA2/ChatGPT in-context learning on SST2 dataset) in relative numbers or 850 (BERT with

Table 2: Break-even points between fine-tuning and the remaining approaches on all investigated datasets. The required labelled samples are indicated only for the best performing model from the 3 groups of models of different sizes: 1) Flan-T5 (small); 2) Mistral/Zephyr (medium); and 3) LLaMA2/ChatGPT (large). The first break-even point (average performance) and the second one (average and variance) are separated with the "/" symbol, with superscript indicating the percentual increase and "NA" indicating no break-even point exists (fine-tuning never outperforms the other approach). We observe a **significant influence of the variance in results on the number of required labelled samples, leading to an increase of up to 1500%** (from 500 to 8000 samples).

		Prompting			In-Context Learning			Instruction-Tuning	
		Flan-T5	Mistral/Zephyr	LLaMA2/ChatGPT	Flan-T5	Mistral/Zephyr	LLaMA2/ChatGPT	Flan-T5	Mistral/Zephyr
SST2	BERT	$90/125^{+39\%}$	$25/50^{+100\%}$	$500/1200^{+140\%}$	$125/225^{+\ 80\%}$	$75/150^{+100\%}$	$3500/10000^{+186\%}$	$16000/20000^{+25\%}$	$4000/20000^{+400\%}$
	RoBERTa	$80/100^{+25\%}$	$50/75^{+\ 50\%}$	$300/1000^{+233\%}$	$100/225^{+125\%}$	$75/175^{+133\%}$	$5500/\ 8000^{+\ 45\%}$	$13000/20000^{+54\%}$	$2500/20000^{+700\%}$
MRPC	BERT	$600/900^{+50\%}$	$450/700^{+55\%}$	$900/1750^{+94\%}$	$700/1100^{+57\%}$	$600/1250^{+108\%}$	$650/1250^{+92\%}$	$900/1750^{+\ 94\%}$	$1500/4000^{+167\%}$
	RoBERTa	$225/350^{+55\%}$	$175/300^{+71\%}$	$300/\ 550^{+83\%}$	$250/\ 400^{+60\%}$	$225/\ 400^{+\ 78\%}$	$225/\ 400^{+78\%}$	$300/\ 600^{+100\%}$	$400/1000^{+150\%}$
CoLA	BERT	$10/10^{+\ \ 0\%}$	$20/\ 75^{+275\%}$	$400/800^{+100\%}$	$50/100^{+100\%}$	$850/NA$	$350/2000^{+471\%}$	$10/10^{+\ \ 0\%}$	$3000/NA$
	RoBERTa	$10/25^{+150\%}$	$75/100^{+\ 33\%}$	$250/750^{+200\%}$	$75/150^{+100\%}$	$800/NA$	$250/1500^{+500\%}$	$10/15^{+50\%}$	$1750/NA$
BoolQ	BERT	NA/NA	$1800/3500^{+94\%}$	NA/NA	NA/NA	$500/8000^{+1500\%}$	NA/NA	NA/NA	NA/NA
	RoBERTa	$4250/6000^{+41\%}$	$800/1250^{+56\%}$	$4750/5000^{+5\%}$	$1750/4000^{+129\%}$	$1000/3000^{+\ 200\%}$	$6000/NA$	$4750/NA$	$3000/4000^{+33\%}$
AGNews	BERT	$50/60^{+20\%}$	$30/50^{+67\%}$	$50/60^{+20\%}$	$550/1100^{+100\%}$	$75/100^{+33\%}$	$50/100^{+100\%}$	$550/1100^{+100\%}$	$100/300^{+200\%}$
	RoBERTa	$40/50^{+25\%}$	$40/50^{+25\%}$	$50/50^{+\ \ 0\%}$	$350/1100^{+214\%}$	$50/\ 50^{+\ \ 0\%}$	$50/\ 50^{+\ \ 0\%}$	$350/1100^{+214\%}$	$75/250^{+233\%}$
TREC	BERT	$60/80^{+33\%}$	$30/40^{+33\%}$	$50/75^{+50\%}$	$90/125^{+39\%}$	$75/100^{+33\%}$	$60/90^{+50\%}$	$100/150^{+50\%}$	$100/175^{+75\%}$
	RoBERTa	$40/50^{+25\%}$	$30/35^{+17\%}$	$40/50^{+25\%}$	$55/\ 75^{+36\%}$	$50/\ 60^{+20\%}$	$40/60^{+50\%}$	$60/\ 90^{+50\%}$	$75/100^{+33\%}$
SNIPS	BERT	$25/40^{+60\%}$	$10/20^{+100\%}$	$40/50^{+25\%}$	$25/40^{+\ 60\%}$	$50/75^{+50\%}$	$50/70^{+40\%}$	$25/40^{+\ 60\%}$	$75/175^{+133\%}$
	RoBERTa	$25/40^{+60\%}$	$10/30^{+200\%}$	$30/40^{+33\%}$	$20/40^{+100\%}$	$40/50^{+25\%}$	$35/50^{+43\%}$	$20/40^{+100\%}$	$50/\ 60^{+\ 20\%}$
DBPedia	BERT	$50/70^{+40\%}$	$40/50^{+25\%}$	$70/80^{+14\%}$	$10/10^{+0\%}$	$100/200^{+100\%}$	$80/200^{+150\%}$	$125/225^{+\ 80\%}$	$150/250^{+\ 67\%}$
	RoBERTa	$45/45^{+\ \ 0\%}$	$40/45^{+13\%}$	$50/50^{+\ \ 0\%}$	$10/10^{+0\%}$	$70/100^{+\ 43\%}$	$50/100^{+100\%}$	$80/160^{+100\%}$	$100/225^{+125\%}$

LLaMA2/ChatGPT on MRPC dataset) in absolute numbers. On the other hand, the increase in the number of required labelled samples is significantly higher for in-context learning approaches (which show the highest variance across the runs) or specific cases in the instruction-tuning approaches (where the variance is more dependent on the model, with larger models showing higher variance), with the highest observed increase of 1500% (BERT with Mistral/Zephyr in-context learning on BoolQ dataset) in relative numbers or increase of 17500 labelled samples (RoBERTa with Mistral/Zephyr instruction-tuning on SST2 dataset) in absolute number. The variance in the fine-tuning approaches does not have as much of an effect, as it is strongly affected by the number of labelled samples – the variance is higher with low number of labelled samples and almost non-existent (lower than prompting) with majority of the dataset (more than $\sim 60\%$ of the labelled samples in the dataset), with the exception of binary dataset where even with all the labelled samples, the variance is quite high. In some cases, instruction-tuning leads to lowered variance of the model (in comparison to using it in zero/few shot setting), but in other cases leads to even larger variance in results by introducing additional sources of randomness in the optimisation process.

3.3 Additional Insights from the Investigation

In this section, we are mainly interested in how other factors such as the size of the model affect the comparison (both in terms of overall performance and variance).

Larger models do not consistently lead to better performance across different approaches. With the fine-tuning, we observe the expected behaviour. On lower number of labelled samples (up to $20 - 100$ depending on the dataset) the smaller BERT model is able to outperform the larger RoBERTa model. However, the larger model can achieve the highest performance on the different datasets using fewer labelled samples. As such, the individual break-even points appear earlier for the larger fine-tuning models. On the other hand, in case of prompting, in-context learning and instruction-tuning, the smaller models are often able to achieve better or similar performance than the larger counterparts even though the different in the number of parameters is large. For example, Flan-T5 in-context learning on the MRPC and CoLA dataset outperforms in-context learning with LLaMA2 and ChatGPT models, while on the BoolQ dataset the situation is opposite. Similarly, instruction-tuning Mistral-7B and Zephyr-7B models often leads to similar or lower performance than the instruction-tuning of Flan-T5 model. **This behaviour does not necessarily depend on the**

dataset characteristics. On some datasets that can be considered harder (longer inputs, more classes, harder tasks that require better language understanding), the Flan-T5 model in zero/few shot setting still outperforms the larger models, while on other datasets it underperforms them. The only impact of the dataset characteristics is the probability of a 'failure mode', where the smaller model does not perform at all at the task. For example, using in-context learning with Flan-T5 on the DBPedia dataset always leads to prediction of a single class for all of the samples, with this behaviour disappearing when reducing the number of classes, or using prompting. Similarly, in-context learning with Flan-T5 on the BoolQ dataset (which is characteristic with its longer input) shows a significantly higher variance. We believe the main culprit for this is the limited context size of the smaller model.

As such, the **number of required labelled samples is not consistently dependent on the size of the general models.** For prompting and in-context learning, we observe many cases when the break-even point for the smaller Flan-T5 requires similar or larger number of labelled samples than the larger models. For example, on the AG News dataset the number of required labelled samples for RoBERTa to outperform Flan-T5 in-context learning is 550 (or 1100 when taking the variance into consideration), while for the medium models it is 75 (or 100) and for large models it is 50 (or 100). Similar observations are present in instruction tuning as well, where the BERT model requires 16000 samples to outperform Flan-T5 instruction-tuning, but only 4000 to outperform medium sized models. Curiously, fine-tuning can often outperform the medium sized models (Mistral/Zephyr) with lower number of samples than the smallest Flan-T5 model. In addition, on the binary datasets, fine-tuning consistently requires the highest number of samples to outperform the largest models (LLaMA-2 and ChatGPT).

In addition, **using larger models does not necessarily lead to lower performance variance.** With the fine-tuning models, we observe the expected behaviour again. Larger model shows lower variance when using enough labelled data, but shows larger variance than the smaller model when dealing with limited data. The performance variance in prompting is similar across all the models, as there are almost no sources of randomness that could affect it. On the other hand, the variance in in-context learning is not as consistent. Comparing the smallest model (Flan-T5) and the largest models (LLaMA-2 and ChatGPT), we observe similar performance variance on majority of the datasets. Only on specific datasets the variance of Flan-T5 is significantly higher (e.g., BoolQ) or significantly lower (AG News). At the same time, the variance of the medium sized models (Mistral-7B and Zephyr-7B) is often significantly higher than their smaller or larger counterparts. Finally, with instruction-tuning, we observe that the Mistral-7B and Zephyr-7B models often show larger variance when compared to Flan-T5. Although this may be due to the different type of instruction-tuning (full tuning in Flan-T5 vs. LoRA tuning in Mistral/Zephyr), we observe similar behaviour of these models in the zero/few shot setting as well.

3.4 Recommendations Based on the Findings

Based on the findings, we argue that **using general large language models in a zero or few shot setting is a preferable solution only in specific cases**. First, if the task **does not have well defined classes** (e.g., planning or answering free text questions) or the task **requires some kind of generation** (e.g., translation, summarisation). Even though the instruction-tuning would provide benefit for such setting, the general models are often already tuned for this behaviour and further tuning would require extensive dataset. As providing in-context examples to allow for in-context learning could be problematic in such a case, prompting is more efficient option. Second, if **faced with a limited annotation and computation budget**, where we either cannot obtain large enough labelled dataset for fine-tuning or we are not able to instruction-tune the general models. Third, if we are **interested in quick prototyping and approximation of the overall performance** we can achieve on the task, before deciding whether to dedicate more time and budget for specialising the models. This recommendation is based on our main finding (specialised models require only small number of labelled samples to outperform general models) and is **in contrast to the common practice of evaluating prompting and in-context learning on classification tasks** (Liu et al., 2023).

Fine-tuning of smaller models is preferable if we have large annotation budget, but small computation budget (and a classification task), or if the task is not well designed for generation and would require additional modifications (e.g., com-

plicated mapping of the generated text to classes, or when multiple classes are possible for each sample). Even when fine-tuning small model (BERT), such approach can outperform all the general models if the annotation budget is large enough. In this case, training the largest possible model (allowed by the computation budget) should be opted for, to reduce the number of required labelled samples as much as possible, to achieve the highest possible performance and to prevent any shortcomings on tasks that require better language understanding. Based on our findings, we **need to annotate only approximately** 1000 **samples, with larger models requiring fewer samples.**

Instruction-tuning of the larger language models is the optimal solution as it provides consistent benefits for all the other cases (large computation budget with either small or large annotation budget). Even with small number of labelled samples (10), instruction-tuning slightly larger general language model (i.e., models that were already pre-trained for prompting and in-context learning such as Flan-T5) can lead to the best overall performance. Instruction tuning provides the best trade-off for the performance and the required resources, as increasing the number of labelled samples does not have a significant impact on the performance, while increasing the size of the instruction-tuned model has only low impact on performance. However, if interested in best overall performance on the task, using the largest general model and instruction-tuning it with large number of labelled samples should provide the most benefit even at the cost of significant increase in effort (and strong diminishing returns).

Finally, **when comparing between different approaches and models, the performance variance should be taken into consideration**, as it has a strong impact on the comparison. If only a single run (or low number of runs) is used, one of the models can be incorrectly denoted as better only based on a random chance. The use of multiple runs is especially important when using in-context learning, instruction-tuning or fine-tuning on low number of labelled samples.

4 Conclusion

In this paper, we investigate the impact of changing the number of available labelled training samples on the performance and comparison of different, currently popular, approaches for data-efficient learning in NLP. The main focus of the investigation is to determine how many labelled training samples are needed for the specialised small language models (obtained through fine-tuning or instruction-tuning) to outperform their general larger counterparts used in zero and few-shot settings. Based on the break-even points from 8 representative text classification datasets of various characteristics, we find that the number of required training samples is quite small, but strongly dependent on the dataset and task characteristics, with more labelled samples needed for binary datasets and tasks that require better language understanding. In addition, we identify a significant influence of the performance variance, stemming from the sensitivity of the different approaches especially in-context learning and fine-tuning with few labelled samples, on the overall comparison and the number of required labelled samples. We can conclude that the specialised smaller models are a strong contender on the text classification tasks, with the general language models showing their benefits for specific cases, such as quick prototyping or when faced with extremely limited annotation or computation budget. Finally, based on findings of our experiments, we provide recommendations that allow for better decision making regarding what approach to use for different settings.

Limitations

The investigation is done on a set of English classification datasets with various characteristics (number of classes, input size, task type, etc.). This choice may limit the generalisation of our findings to other datasets, tasks and languages, such as the generation tasks, which are more representative for the general models and on which the fine-tuning cannot be used as easily (e.g., question answering, summarisation, translation). However, we explicitly discuss this in Section 3.4.

In addition, we perform the investigation across a lower number of smaller models. We work with the base versions of the smaller BERT, RoBERTa and Flan-T5 models. The larger models used, specifically Mistral-7B, Zephyr-7B and LLaMA-2 13B, are all used with the 4 bit quantisation. The largest model we use is the ChatGPT (3.5 version). We explicitly choose these models to provide a more contrasted comparison between the performance while limiting the required computation costs and to provide more in-depth analysis of their behaviour (in-

cluding how the performance variance affects the comparisons). Even though the fine-tuning models consistently outperforms the larger counterparts with low number of labelled samples, the choice of smaller models may limit the generalisation of our findings. However, we publicly release the source code of our investigation to allow for its extension to other models and datasets.

Due to the significant computation costs incurred by the inference of larger language models (namely LLaMA-2, Mistral-7B and Zephyr-7B), instruction-tuning of the medium sized models (Mistral/Zephyr) and the additional other costs of the ChatGPT model, we evaluate the models in a limited setting. The evaluation is done on a randomly selected set of 1 000 samples (following the practice in prior works (Gao et al., 2021; Chang and Jia, 2023; Sclar et al., 2024; Li and Qiu, 2023; Köksal et al., 2023)). As the majority of the datasets is smaller (e.g., MRPC, BoolQ, CoLA, SNIPS, TREC), this decision may not be as problematic, as the used 60:20:20 split leads to approximately the 1000 test samples. However, on the larger datasets (SST2, AG News and DB Pedia) this decision may skew the results. In addition, we run the ChatGPT prompting and in-context learning only 10 times instead of 20 and only for the 1 000 labelled training samples setting, due to its costs.

Finally, we use the same prompt for all the models, which is a result of a prompt engineering on the Mistral-7B model. The prompt was created based on dataset description, the prompts used in related works and the formats recommended for different models (e.g., taking inspiration from (Sun et al., 2023)). As such, this may not represent the optimal format for all the models (as identified in previous works (Sclar et al., 2023)) and performing the investigation on multiple different prompts may improve the overall model performance and affect the findings. However, we opted for using only a single optimised prompt format to reduce the computation costs and provide more in-depth analysis on larger number of models.

Ethical Considerations

The experiments in this paper work with publicly available benchmark datasets GLUE and SuperGLUE and other publicly available datasets (AG News, TREC, SNIPS and DB Pedia), citing the original authors. As we were not able to determine the license for the tasks used, we opted to use them in as limited form as possible, adhering to the terms of use (no annotation of the test set) defined by the GLUE and SuperGLUE and applying it to other datasets as well. We do not work with any personally identifiable information or offensive content and perform no crowdsourcing for further data annotation. In addition, we are not aware of any potential ethical harms or negative societal impacts of our work, apart from the ones related to the advancement of the field of machine learning. Finally, we follow the license terms for all the models we use (such as the one required for the use of the LLaMA-2 model). It is possible the large language models we use contain biases and potentially offensive or harmful content. However, the original authors of these models reduce this bias as much as possible. At the same time, we do not release any output of the models which should further reduce the potential bias and negative impact.

Impact Statement: CO2 Emissions Related to Experiments The experiments presented in this paper used significant compute resources as they required multiple training and evaluation runs of multiple models (to deal with variance in results), as well as using large language models that requires a lot of computation even just for the inference. Overall, the experiments were conducted using a private infrastructure, which has a carbon efficiency of 0.432 kgCO_2eq/kWh (default value used as the actual efficiency of our HW instance was not measured). A cumulative of 4000 hours of computation was performed on hardware of type A100 PCIe 40GB (TDP of 250W). Total emissions are estimated to be 432 kgCO_2eq of which 0 percents were directly offset. This does not include the compute used by the ChatGPT model behind API as we are not able to estimate these resources. These estimations were conducted using the MachineLearning Impact calculator presented in (Lacoste et al., 2019). Whenever possible, we tried to reduce the compute resources used as much as possible. The most compute resources were used by the large language models – LLaMA-2, ChatGPT, Mistral-7B and Zephyr-7B. As the prompting and in-context learning with these models resulted in quite stable results, we decided to evaluate them only on a single setting (using 1 000 labelled training samples) and using a fraction of the whole test set (1 000 test samples). In addition, for the ChatGPT model, we evaluate only on 10 runs. Even in their reduced

evaluation, these experiments used large fraction of the GPU hours. The most significant contributor was the instruction-tuning, especially with the medium sized models (Mistral/Zephyr), where we opted to reduce the number of steps and epochs used for the training.

Acknowledgements

This work was partially supported by the projects funded by the European Union under the Horizon Europe: *DisAI*, GA No. 101079164, *AI-CODE*, GA No. 101135437; and by *Modermed*, a project funded by the Slovak Research and Development Agency, GA No. APVV-22-0414.

Part of the research results was obtained using the computational resources procured in the national project *National competence centre for high performance computing* (project code: 311070AKF2) funded by European Regional Development Fund, EU Structural Funds Informatization of Society, Operational Program Integrated Infrastructure.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144. ACL.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. 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 2924–2936, Minneapolis, Minnesota. ACL.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

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. *arXiv preprint arXiv:2002.06305*.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. ACL.

Odd Erik Gundersen, Kevin Coakley, Christine Kirkpatrick, and Yolanda Gil. 2022. Sources of irreproducibility in machine learning: A review. *arXiv preprint arXiv:2204.07610*.

Odd Erik Gundersen, Saeid Shamsaliei, Håkon Sletten Kjærnli, and Helge Langseth. 2023. On reporting robust and trustworthy conclusions from model comparison studies involving neural networks and randomness. In *Proceedings of the 2023 ACM Conference on Reproducibility and Replicability*, pages 37–61.

Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, Mutsumi Nakamura, Arindam Mitra, Santosh Mashetty, and Chitta Baral. 2023. Instruction tuned models are quick learners. *arXiv preprint arXiv:2306.05539*.

Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, Emily Alsentzer, et al. 2023. Do we still need clinical language models? In *Conference on Health, Inference, and Learning*, pages 578–597. PMLR.

SU Hongjin, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. 2022. Selective annotation makes language models better few-shot learners. In *The Eleventh International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Abdullatif Köksal, Timo Schick, and Hinrich Schuetze. 2023. MEAL: Stable and active learning for few-shot prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 506–517, Singapore. ACL.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. ACL.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick van Kleef, Sören Auer, and Christian Bizer. 2015. DBpedia – a large-scale, multilingual knowledge base extracted from wikipedia. *Semant. Web*, 6(2):167–195.

Xiaonan Li and Xipeng Qiu. 2023. Finding support examples for in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6219–6235, Singapore. ACL.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

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

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting down on prompts and parameters: Simple few-shot learning with language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. ACL.

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

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10572–10601, Singapore. ACL.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines. In *International Conference on Learning Representations*.

Marius Mosbach, Tiago Pimentel, Shauli Ravfogel, Dietrich Klakow, and Yanai Elazar. 2023. Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12284–12314, Toronto, Canada. ACL.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Branislav Pecher, Ivan Srba, and Maria Bielikova. 2023. On the effects of randomness on stability of learning with limited labelled data: A systematic literature review. *arXiv preprint arXiv:2312.01082*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1339–1384, Singapore. ACL.

Timo Schick and Hinrich Schütze. 2021. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. ACL.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. ACL.

Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, et al. 2023. Pushing the limits of chatgpt on nlp tasks. *arXiv preprint arXiv:2306.09719*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. ACM.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. ACL.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Yiming Zhang, Shi Feng, and Chenhao Tan. 2022. Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9134–9148, Abu Dhabi, United Arab Emirates. ACL.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.