# Mind Your Format: Towards Consistent Evaluation of In-Context Learning Improvements

**Anton Voronov**
Yandex, HSE University, MIPT
voronov.ad@phystech.edu

**Lena Wolf**
Yandex, HSE University
e.a.volf@yandex.ru

**Max Ryabinin**
Together AI
mryabinin0@gmail.com

## Abstract

Large language models demonstrate a remarkable capability for learning to solve new tasks from a few examples. The *prompt template*, or the way the input examples are formatted to obtain the prompt, is an important yet often overlooked aspect of in-context learning. In this work, we conduct a comprehensive study of the template format's influence on the in-context learning performance. We evaluate the impact of the prompt template across 21 models (from 770M to 70B parameters) and 4 standard classification datasets. We show that a poor choice of the template can reduce the performance of the strongest models and inference methods to a random guess level. More importantly, the best templates do not transfer between different setups and even between models of the same family. Our findings show that the currently prevalent approach to evaluation, which ignores template selection, may give misleading results due to different templates in different works. As a first step towards mitigating this issue, we propose *Template Ensembles* that aggregate model predictions across several templates. This simple test-time augmentation boosts average performance while being robust to the choice of random set of templates.

## 1 Introduction

Pretrained language models have emerged as a dominant paradigm for solving many NLP problems in a unified framework (Brown et al., 2020; Chowdhery et al., 2022; Scao et al., 2023; Touvron et al., 2023a). In particular, these models can achieve impressive downstream results with just a few demonstrations given as a part of their input (Liu et al., 2021; Min et al., 2022c), which is often called *a prompt* in this case.

These few-shot or *in-context learning* (ICL) abilities (Brown et al., 2020) of large models are a subject of frequent study, as the primary factors behind them are not yet fully understood. For example, one line of work investigates in-context learning within different theoretical frameworks (Xie et al., 2022; Garg et al., 2022; Akyürek et al., 2023). In addition, multiple publications study the importance of different prompt attributes, such as the order of input demonstrations (Lu et al., 2022a) and their labels (Min et al., 2022d).

As shown in Zhao et al. (2021); Min et al. (2022a), the prompt format (i.e., a transformation from a set of examples to a natural language input) is also highly important. However, this aspect is often overlooked in most existing studies. Namely, works proposing modifications of ICL frequently present their results for a *specific* template without specifying the criteria guiding its selection. Furthermore, even when the results are averaged over a set of templates, they are compared to methods that were evaluated on a *different* set of templates. We illustrate this common discrepancy in Appendix A. Such inconsistency can lead to a misinterpretation of the reported results: the difference between the performance of two methods may be explained by the variation across prompt formats rather than the methods themselves.

In this work, we evaluate the template sensitivity of 21 models from 8 families, including state-of-the-art open-access models, such as Llama 2 (Touvron et al., 2023b) and Falcon (Almazrouei et al., 2023), as well as latest instruction-tuned models, such as Mistral (Jiang et al., 2023) and Llama 3 Instruct (AI@Meta, 2024). We show that this issue persists regardless of the model size and the number of demonstrations. Moreover, comparing various in-context learning enhancements while taking the template influence into account renders the superiority of one method over others less apparent. Therefore, it is likely that the gains reported for advanced prompting methods can often be attributed to a luckily chosen template.

*Findings of the Association for Computational Linguistics ACL 2024*, pages 6287–6310
August 11-16, 2024 ©2024 Association for Computational Linguistics

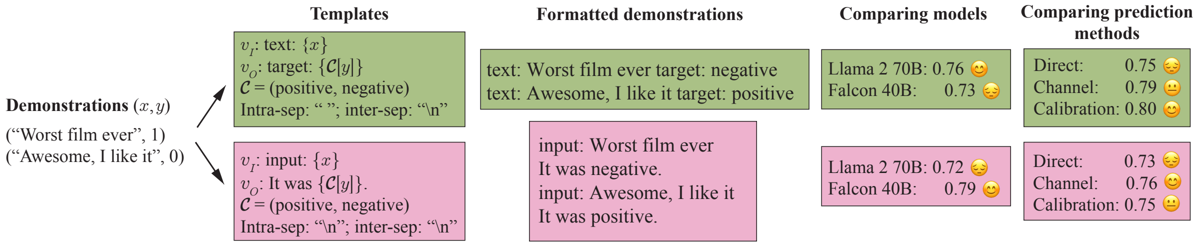

Templates
Formatted demonstrations
Comparing models
Comparing prediction methods
Demonstrations $(x, y)$
(“Worst film ever”, 1)
(“Awesome, I like it”, 0)
$v_I$: text: $\{x\}$
$v_O$: target: $\{\mathcal{C}[y]\}$
$\mathcal{C}$ = (positive, negative)
Intra-sep: “ ”; inter-sep: “\n”
$v_I$: input: $\{x\}$
$v_O$: It was $\{\mathcal{C}[y]\}$.
$\mathcal{C}$ = (positive, negative)
Intra-sep: “\n”; inter-sep: “\n”
text: Worst film ever target: negative
text: Awesome, I like it target: positive
input: Worst film ever
It was negative.
input: Awesome, I like it
It was positive.
Llama 2 70B: 0.76 😊
Falcon 40B: 0.73 😔
Llama 2 70B: 0.72 😔
Falcon 40B: 0.79 😊
Direct: 0.75 😔
Channel: 0.79 😐
Calibration: 0.80 😊
Direct: 0.73 😔
Channel: 0.76 😊
Calibration: 0.75 😐

Figure 1: An example template transformation for two demonstrations. Different prompt formats lead to different rankings both for models and ICL methods, and the best template for one method can be suboptimal for others.

Crucially, **there are no universally best templates** for a given task. The best performing demonstration format for a fixed evaluation setting (i.e., the dataset, the model, the demonstration set, and the prediction method) does not transfer consistently across models (even within the same family), demonstration sets, or different prediction methods. We find this concerning, as even the best template for a given setting can produce poor results after slight changes, which makes “tuning” the template a very difficult task.

As a first step towards addressing template sensitivity in a practical way, we propose **Template Ensembles** — a test-time augmentation approach that averages model predictions over several prompt formats. This method is easy to implement and increases the average performance across templates for multiple prompting methods while reducing the sensitivity of these methods to the template choice.

In summary, our contributions are as follows:

1. We conduct a broad evaluation[1] of prompt template sensitivity across 21 models and 4 datasets, showing that the performance gains similar to using in-context learning improvements can be achieved solely by selecting a proper template.

2. We show that the choice of the best template depends on a combination of factors and that it is not possible to transfer the best template between models or prompting methods without a negative impact on quality.

3. We propose Template Ensembles as a baseline solution for improving the template robustness for in-context learning.

## 2 Background and Related Work

### 2.1 In-Context Learning

An important property of LLMs is their ability to learn new tasks from only a few demonstrations (Radford et al., 2019; Brown et al., 2020). This capability, known as in-context learning, forms the focus of our work. We focus on sequence classification, as it is the most widely studied task for understanding and improving ICL performance.

Formally, classifying an input $x_{test}$ with in-context learning can be described as finding the class $c$ in the space of label tokens $\mathcal{C}$ that yields a sequence with the highest probability according to a language model. The input sequence consists of demonstration inputs and labels $(x_i, y_i)$ and a test input $(x_{test}, c)$; to obtain a natural language input, demonstrations are formatted with a template.

Each template consists of four components: input and output verbalizers $v_I(x)$ and $v_O(y, \mathcal{C})$ that transform $(x_i, y_i)$ into a natural language text, an intra-separator to divide input from output, and an inter-separator to join several demonstrations. Figure 1 shows an example of transforming a set of demonstrations into a context for ICL.

### 2.2 In-Context Learning Analysis

Recent work has shown that ICL can perform at levels comparable to finetuning (Chowdhery et al., 2022; Hoffmann et al., 2022). Still, in-context learning is known to be highly dependent on the way the model input is formed: a prompt is defined by several components, and altering any of them can lead to unpredictable changes in performance.

**Template Selection** There are multiple ways to construct a template for a task. The most straightforward approach is to use minimal templates ($v_I = \{x\}, v_O = \{\mathcal{C}[y]\}$) or universal verbalizers like “input/output”, as done in Wang et al. (2023) and Wei et al. (2023).

[1] Our code and results of all evaluations can be found at github.com/yandex-research/mind-your-format

Another strategy is to create task-specific templates. Jiang et al. (2020) generate paraphrases of templates for the relation extraction task. Authors show the sensitivity of masked language models to the prompt format and propose to ensemble predictions over the best templates. Compared to this method, our approach is task-agnostic and does not require evaluating all templates in advance.

Several studies aim to find templates that directly optimize in-context learning performance (Shin et al., 2020; Gao et al., 2021). Our work unifies the results of previous research, using the verbalizers proposed by Gao et al. (2021), as well as minimal and universal templates.

**Choice and Order of Demonstrations** The choice of examples for ICL is highly important, as they enable the model to condition on correct input and label distributions for the task (Wu et al., 2023; Nguyen and Wong, 2023; Min et al., 2022d; Chang and Jia, 2023). Furthermore, the order of examples also significantly affects the results and does not transfer between models even within the same family (Lu et al., 2022b; Zhao et al., 2021).

In this work, we analyze two recent methods for selecting demonstrations. Wang et al. (2023) propose learning latent concept variables for a task and using them to find examples that can best predict the task concept. We refer to this method as *Implicit Topic Models* or ITM. In turn, Z-ICL (Lyu et al., 2023) generates pseudo-demonstrations by retrieving most similar examples to the test sentence from an unlabeled dataset and assigning random labels to retrieved examples.

Crucially, both methods are evaluated on single templates that differ across two works. Therefore, it is unclear whether the reported performance gains arise from the methods themselves or from a particular combination of the example selection strategy, the model, and the chosen template.

**Prediction Methods** The standard approach for classification with LLMs is to compute the sequence probability with each of the possible labels and select the label with the highest probability. We refer to this method as DIRECT further on.

Alternatively, one can use more advanced prediction methods that aim to reduce the variance across prompt formats. The CALIBRATION method (Zhao et al., 2021) computes a correction factor based on the deviation of the model's predictions for a placeholder input from a uniform distribution over labels and applies this factor to test set predictions. Recent work has proposed multiple improvements of this method (Fei et al., 2023; Han et al., 2023; Zhou et al., 2024); to limit the scope of our study, we focus only on the base CALIBRATION approach in this work. Lastly, the CHANNEL prompting technique, proposed in Min et al. (2022b), maximizes $P(x|y)$ instead of $P(y|x)$.

Both of these methods aim to mitigate the issue of ICL sensitivity to the prompt template choice. However, as we show in Appendix A, these methods are evaluated on their own sets of templates. In this paper, we strive for a more unified view on the robustness of advanced prompting methods and compare their performance across a broader range of templates and models.

**Prompt and Template Robustness** Although the problem of prompt robustness is relatively well-known, until recently, the discussion of *template robustness* has been limited. Notably, Sclar et al. (2023) present a highly relevant study of prompt format sensitivity, reporting a significant performance variation across formats even for large models or minor template changes. While their experiments are conducted in the standard setup (randomly selected examples and default prompting), our work instead focuses on alternative prompting and example selection methods, several of which (Zhao et al., 2021; Min et al., 2022b) were proposed to improve the prompt robustness of ICL. Similarly to papers in other subfields of machine learning arguing for a more consistent methodology (Dacrema et al., 2019; Musgrave et al., 2020; Platonov et al., 2023), the goal of our work is to demonstrate that disparate experiment setups lead to an invalid comparison of competing methods.

Moreover, several works study prompt robustness in a broader sense by considering models that use natural language instructions instead of labeled demonstrations (Webson and Pavlick, 2022; Leidinger et al., 2023; Weber et al., 2023). Recently, Mizrahi et al. (2023) have shown that very similar instructions can lead to drastic differences in task performance for a variety of instruction-tuned models. Although we study a similar issue, the focus of our work is on in-context learning and the transfer of best prompts between evaluation setups. Still, we find that instruction-tuned models lack in-context robustness as well, which confirms previous observations and emphasizes the need for language model evaluation that takes prompt design into account.

| Model family | Parameters (B) |
|---|---|
| GPT-J (Wang and Komatsuzaki, 2021) | 6 |
| GPT-NeoX (Black et al., 2022) | 20 |
| BLOOM (Scao et al., 2023) | 1.7, 3, 7.1 |
| OPT (Zhang et al., 2022) | 6.7, 30, 66 |
| Pythia (Biderman et al., 2023) | 6.9, 12 |
| LLaMA (Touvron et al., 2023a) | 7, 13, 30, 65 |
| Llama 2 (Touvron et al., 2023b) | 13, 70 |
| Falcon (Almazrouei et al., 2023) | 1, 7, 40 |
| Llama 3 Instruct (AI@Meta, 2024) | 8 |
| Mistral v0.3 Instruct (Jiang et al., 2023) | 7 |

Table 1: Language models used in our work.

# 3 Setup & Methodology

## 3.1 Models and Data

We evaluate the robustness of in-context learning to template selection across a wide range of models on classification tasks. All models used in our work are listed in Table 1: we run experiments on model families frequently used in literature (such as OPT and BLOOM), as well as the latest models with the highest quality (such as Llama 2 and Falcon).

In preliminary experiments, we observed that the performance of some models in the few-shot regime lags behind their zero-shot results. Hence, we excluded these models from further investigation. Further details regarding this selection procedure can be found in Appendix B.

We experiment with 4 sequence classification datasets: SST-2 (Socher et al., 2013), DBPedia ontology classification task (Lehmann et al., 2015), AGNews (Zhang et al., 2015), and TREC Question Classification (Li and Roth, 2002). Although these datasets are frequently used in ICL studies, there is no consensus regarding the templates that should be used for each task.

One can construct an input for in-context learning from a set of demonstrations by using a template consisting of four parts, as illustrated in Figure 1. We present all options for verbalizers and separators for each dataset we study in Table 2. Any combination of these components results in a valid template. This set of options results in 216 possible prompt formats for SST-2 and 168 for DBPedia, AGNews and TREC. A single evaluation run of all models on 10 random templates in one setup takes 17–48 hours on a single NVIDIA A100-80GB GPU, depending on the dataset.

## 3.2 Methods

Along with studying the robustness of standard in-context learning, we consider its improvements proposed in prior work. We focus on two main directions of ICL enhancements mentioned in Section 2: example selection and prediction methods. For each setting, we aggregate the results over 3 random seeds for example selection, with 10 random templates used for each seed and report the mean and standard deviation of classification accuracy, unless specified otherwise.

As a baseline for demonstration selection, we choose the most straightforward approach of selecting $N$ random examples from the training dataset. Intuitively, selecting more relevant examples for ICL should yield better performance. Therefore, we investigate the template sensitivity of two demonstration selection methods described in Section 2.1: ITM and z-ICL. Specifically, we select $N = 4$ examples using official implementations of each method.

Importantly, ITM requires training a concept model before choosing the examples. For GPT-2 Large, this procedure takes approximately 30 hours on a single NVIDIA A100-80GB. Repeating it for each model would be infeasible, especially given that the largest model has 86 times more parameters. Therefore, we use the checkpoints of the GPT2 Large concept model provided by the authors to select demonstrations. Also, we reuse the same examples for all models, leveraging authors' observations that demonstrations chosen with ITM can be transferred between models.

| Dataset | Input verbalizer | Output verbalizer | Intra-separator | Inter-separator |
|---|---|---|---|---|
| SST-2 | “input: {}”,<br>“text: {}”,<br>“sentence: {}”,<br>“{}” | “output: {}”, “target: {}”, “label: {}”,<br>“emotion: {}”, “sentiment: {}”, “A {} one.”<br>“It was {}.”, “All in all {}.”, “A {} piece.” | “ ”,<br>“\n” | “ ”,<br>“\n”,<br>“\n\n” |
| DBPedia | | “output: {}”, “target: {}”, “label: {}”,<br>“Topic: {}.”, “Subject: {}.”,<br>‘This is about {}.”, “It is about {}.” | | |
| AGNews | | | | |
| TREC | | | | |

Table 2: Possible choices for all components of templates used in our work.

| Model | SST-2 | | DBPedia | | AGNews | | TREC | |
|---|---|---|---|---|---|---|---|---|
| | 2-shot | 4-shot | 2-shot | 4-shot | 2-shot | 4-shot | 2-shot | 4-shot |
| Falcon 1B | $0.65_{0.17}$ | $0.77_{0.15}$ | $0.36_{0.25}$ | $0.44_{0.23}$ | $0.52_{0.17}$ | $0.56_{0.19}$ | $0.26_{0.09}$ | $0.31_{0.09}$ |
| Falcon 7B | $0.77_{0.16}$ | $0.83_{0.16}$ | $0.40_{0.21}$ | $0.49_{0.18}$ | $0.51_{0.20}$ | $0.60_{0.19}$ | $0.32_{0.09}$ | $0.39_{0.11}$ |
| Falcon 40B | $0.79_{0.17}$ | $0.92_{0.07}$ | $0.42_{0.15}$ | $0.54_{0.06}$ | $0.64_{0.23}$ | $0.75_{0.09}$ | $0.36_{0.07}$ | $0.46_{0.10}$ |
| Llama 2 13B | $0.79_{0.17}$ | $0.92_{0.07}$ | $0.40_{0.15}$ | $0.51_{0.09}$ | $0.70_{0.15}$ | $0.76_{0.09}$ | $0.32_{0.09}$ | $0.41_{0.14}$ |
| Llama 2 70B | $0.83_{0.14}$ | $0.92_{0.09}$ | $0.46_{0.15}$ | $0.60_{0.05}$ | $0.76_{0.14}$ | $0.82_{0.05}$ | $0.41_{0.07}$ | $0.51_{0.06}$ |

Table 3: Classification accuracy in the baseline setting for 2 LLM families. Standard deviation across 30 runs (10 templates for 3 sets of demonstrations) is in underscript. The results for all base models are presented in Appendix C.

As discussed in Section 2, more advanced prediction techniques can improve in-context learning accuracy. Therefore, we compare Direct prompting with Channel (Min et al., 2022b) and Calibration (Zhao et al., 2021) prediction methods.

## 4 Evaluation

### 4.1 Baseline Results

We begin with analyzing the robustness of language models to the template choice in the baseline setup. Specifically, we evaluate models in zero-shot and few-shot settings, selecting 2/4 random demonstrations and using the Direct prediction method.

Our results in Table 3 show that even the most capable models such as Falcon and Llama 2 are highly sensitive to the prompt format; Appendix C contains the results for the full set of 19 base models, and Appendix L reports the results for instruction-tuned models. Although the variance caused by this sensitivity makes it harder to observe the increase in ICL performance with the model size or the number of demonstrations, both trends still persist. However, even the largest models have standard deviations of scores up to 35% of their mean values.

To mitigate this lack of robustness, we could remove consistently underperforming prompt formats from the template pool. We analyze the impact of separate template components in Appendix D and find that there are no specific parts (for example, verbalizers or separators) which could be excluded from evaluation. Furthermore, we observe that a combination of "suboptimal" parts may result in an optimal template.

### 4.2 Prediction Methods

Next, we aim to evaluate the performance of different prediction methods in a unified setting. Ideally, we would like these modifications to reduce the variance across templates, making the model behavior less dependent on the input format.

We evaluate Channel and Calibration methods in the 2-shot setting along with the Direct baseline. As depicted in Figure 2, both Channel and Calibration generally exhibit improved performance in comparison with Direct. Still, for a number of models and datasets, the range of scores for Direct substantially overlaps with those of advanced methods. This suggests that there are templates reaching the best performance with the Direct prediction method.

Additionally, Table 10 of Appendix E reveals that despite Calibration yielding the highest mean accuracy more often than other methods, it is more sensitive to the template choice than Channel. Similar findings for instruction-tuned models are contained in Appendix L. Therefore, the choice of the prediction method should likely rely on the downstream usage scenario and the target evaluation setting.

### 4.3 Example Selection Methods

Another area of ICL improvements that we evaluate on the matter of template sensitivity is the example selection strategy. We compare ITM and z-ICL methods to the Random baseline in 4-shot setting, since using 4 demonstration was the main evaluation setting in the works proposing these methods. We use Direct prediction method to evaluate the gains of advanced example selection strategies independently from other ICL modifications.

Results in Figure 3 and Table 14 illustrate that when taking template sensitivity into account, advanced example selection methods often perform worse than random choice baseline. ITM increases the average performance in most cases but still has a remarkably high standard deviation across templates. Examples selected using the z-ICL method lead to more consistent but worse performance.

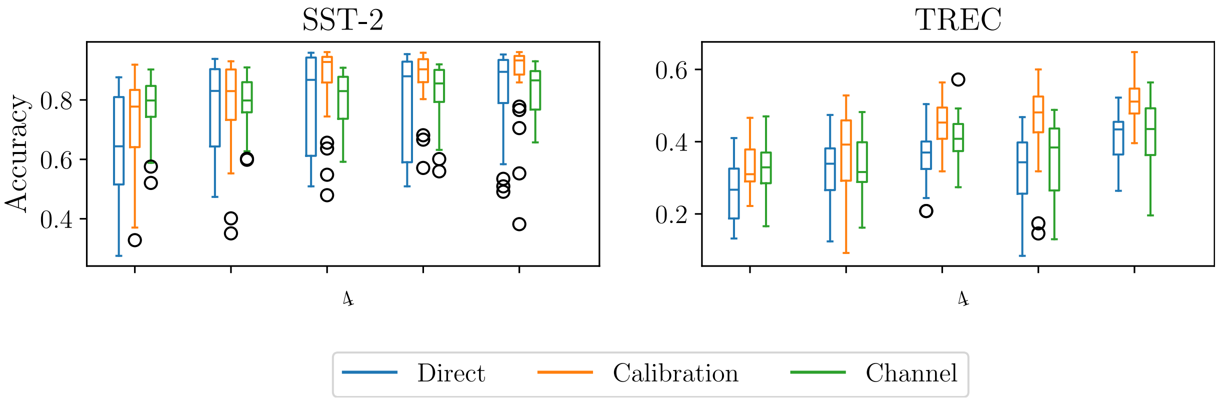


Figure 2: Comparison of in-context learning prediction methods in the 2-shot setting.

Note that our evaluation setup differs from those described in the original works, which might explain the discrepancy between our findings and the results reported by authors. Namely, we use the DIRECT prompting method and sample 10 random templates that may not include the templates used by authors of ITM and z-ICL. To confirm template instability for prediction methods in their original implementations, we reproduce both methods and report our findings in Appendix F. We observe high sensitivity to the prompt format, which raises a question of how much the reported gains of these methods can be attributed to the methods themselves and not to the template choice.

We conclude that the prompt format should be viewed as important as the example selection or the prediction method for ICL evaluation. However, the search space of possible templates is infinite, which makes exhaustive search for each combination of the dataset, the model and the number of examples impractical. Ideally, the best template for one setting would be optimal for all others or at least for similar settings. However, as we demonstrate in the following section, this is not the case.

## 5 Template Transfer Evaluation

### 5.1 Setup

We begin by defining a successful transfer between ICL settings. In order to do so, we evaluate how the quality of model predictions varies across 30 random templates from Table 2. The results described in Appendix H demonstrate that the top-10 template on average yields 90% of the best template score. Therefore, if a prompt format is present in top-10 for both of the two compared setups, we can consider this an instance of successful transfer.

To compare sets of the best templates for a pair of settings, we compute Intersection-over-Union (IoU), also known as the Jaccard similarity coefficient (Jaccard, 1912), for top-10 best templates in each setting. We also consider using the $\rho$ rank correlation coefficient (Spearman, 1904) as another measure of template transfer. However, its value can increase when low-performing templates have similar rankings in different ICL setups, while the transfer of efficient templates remains low. Still, we provide the results for this metric in Appendix I.

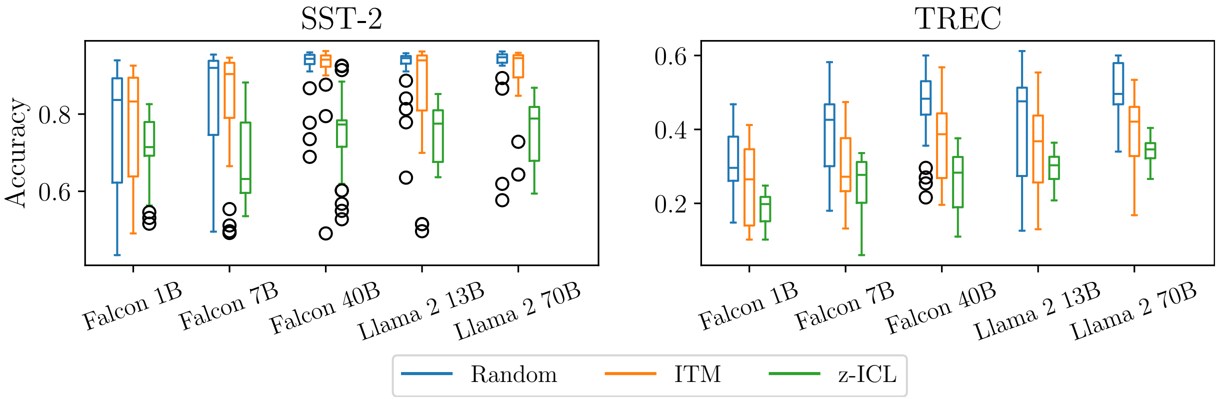


Figure 3: Comparison of the selection methods in the DIRECT 4-shot setting. For the evaluation results of other models and datasets, please refer to Appendix G.

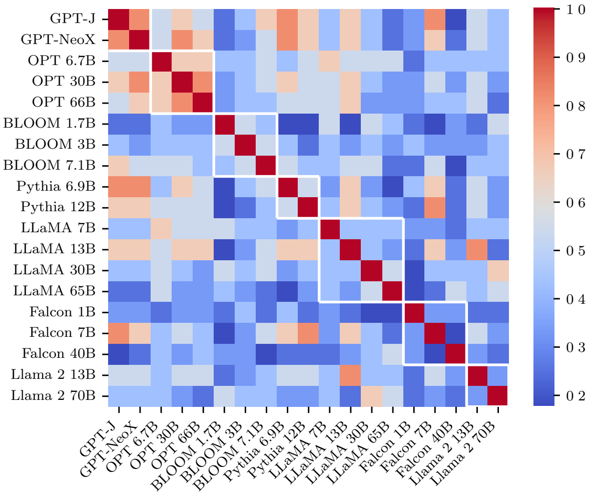


Figure 4: IoU of top-10 templates for all base models with 2 random demonstrations and the DIRECT prediction method on the DBPedia dataset.

### 5.2 Transfer Between Models

Next, we analyze the transfer of the best-performing templates between models in the baseline setup. Specifically, we collect the results of each model in the 2-shot learning setting with DIRECT prediction method and RANDOM demonstrations (fixed throughout the experiments) for 30 templates. A heatmap of IoU for the transfer of top-10 best templates between 19 base models on the DBPedia dataset is presented in Figure 4; for other datasets, please see Appendix J.

We observe that the IoU values exceed 0.5 only for a few model pairs on all datasets, meaning that the capacity for template transfer between models in the same setup is generally low. This is especially concerning for models within a single family: as these models are trained on the same data and have the same architecture, one would expect them to perform similarly on the same prompt formats.

These observations signify that comparing ICL methods across models with a single template can lead to incorrect conclusions: a template that is effective for one model can easily be one of the worst choices for another model.

### 5.3 Transfer Between Prediction Methods

As discussed in Section 4.2, no prediction method that we evaluate can consistently outperform others across all models and datasets. Therefore, to find an optimal setup for a new ICL improvement, one needs to evaluate every prediction technique in multiple templates. We investigate the possibility of finding a universally optimal prompt for different methods to reduce the total computational cost.

| | Direct ↔ Calibration | Direct ↔ Channel | Channel ↔ Calibration |
|---|---|---|---|
| SST-2 | $0.49_{0.17}$ | $0.30_{0.11}$ | $0.31_{0.08}$ |
| DBPedia | $0.54_{0.17}$ | $0.47_{0.15}$ | $0.45_{0.14}$ |
| AG News | $0.36_{0.11}$ | $0.25_{0.13}$ | $0.35_{0.14}$ |
| TREC | $0.31_{0.12}$ | $0.23_{0.09}$ | $0.28_{0.13}$ |

Table 4: Intersection-over-Union for pairs of prompting methods averaged over the results of 19 base models obtained in the RANDOM 2-shot setup. Standard deviations are in subscript.

To answer this question, we calculate the IoU between top 10 performing templates for each method for a fixed set of demonstrations. Results in Table 4 display that similarly to the models, the transfer between prediction methods is also low. Consequently, the prompt format sensitivity issue creates a burden on authors of new ICL modifications; they must tune templates for every prediction method they want to combine with their own approach.

### 5.4 Transfer Between Demonstration Selection Methods

Having found that the best-performing templates are specific both to the model and the prediction method, we now aim to find whether the best formats would be the same for different demonstration sets in the same setup. Similarly to previous experiments, we calculate IoU for 10 templates that yield the highest scores for each method.

Results in Figure 5 illustrate that simply adding demonstrations, even if they were obtained with the same method, can significantly alter the ranking of the best templates. This justifies the necessity to evaluate example selection methods on a range of templates to avoid misinterpretation of the results.

### 5.5 Discussion

Based on the above findings, we conclude that the results of evaluation of various ICL improvements without consideration of template sensitivity issue are hardly reliable for several reasons. First, as the best templates do not transfer between models even within the same family, scoring a method across several models using the same format will inevitably lead to underestimation of the method for all models except the one for which the format was tuned. Next, as there is little evidence of transfer between setups, the format selection procedure needs to be precisely described and applied in all evaluated settings for a fair comparison.

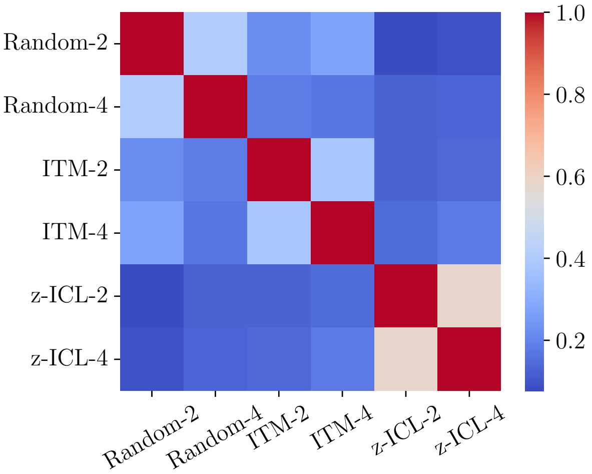


Figure 5: IoU of 10 best templates for example selection methods on the AG News dataset. METHOD-N indicates that METHOD was used to select $N$ examples.

In summary, we find that there are no universally well-performing prompt formats. Therefore, the results of in-context learning evaluation can be reliable only if they are **aggregated over several templates** or if **each setting is evaluated in its best-performing template**. The former approach requires accounting for the variance of the scores and makes comparison less apparent, while the latter can be computationally expensive.

## 6 Template Ensembles

To reduce the variance in performance caused by the template choice, we propose to ensemble model predictions across multiple templates. This approach is widely used in machine learning (Ho, 1995; Lakshminarayanan et al., 2017) for improving the predictive performance of the model, as well as its robustness, and can be viewed as a form of test-time augmentation (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015). Prior work on prompt ensembling has shown significant gains by training a boosting algorithm on model outputs (Hou et al., 2023); by contrast, our method needs only the pretrained model predictions without additional training.

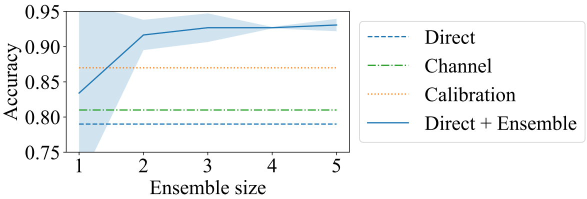


Figure 6: Template ensemble accuracy as a function of its size for Falcon 40B on the SST-2 dataset in the 2-shot learning setup. Dashed lines depict the results of baseline methods averaged over 10 templates.

Formally, our method computes label probabilities across predictions for each of $N$ templates, where $N$ is the ensemble size, and outputs the label with the highest average probability. In early experiments, we tried selecting the most common label among the predictions; however, we found this voting strategy to perform poorly on tasks with a large number of classes. It is also important to note that ensembling $N$ predictions involves running the model $N$ times more compared to single-format evaluation, which makes this approach more computationally expensive. We view template ensembles as a baseline solution for the problem of prompt format sensitivity and leave the exploration of more efficient methods to future work.

We begin with determining the minimal ensemble size that consistently reduces variance while increasing the average performance. We observe that for the majority of models and prediction methods, an ensemble achieves the best accuracy when its size reaches 4 or 5 (see an example in Figure 6), with further expansion being less effective. We also found that smaller ensembles may demonstrate unstable behavior, with the possibility of a drop in

| Model | Direct | | Channel | | Calibration | |
|---|---|---|---|---|---|---|
| | Single | Ensemble | Single | Ensemble | Single | Ensemble |
| LLaMA 2 13B | $0.79_{0.17}$ | $0.85_{0.09}$ | $0.82_{0.10}$ | $0.90_{0.02}$ | $0.88_{0.09}$ | $\mathbf{0.93}_{0.03}$ |
| LLaMA 2 70B | $0.83_{0.14}$ | $\mathbf{0.95}_{0.01}$ | $0.83_{0.08}$ | $0.92_{0.01}$ | $0.88_{0.13}$ | $0.94_{0.03}$ |
| Falcon 1B | $0.65_{0.17}$ | $0.74_{0.07}$ | $0.77_{0.10}$ | $\mathbf{0.89}_{0.01}$ | $0.71_{0.17}$ | $\mathbf{0.89}_{0.01}$ |
| Falcon 7B | $0.77_{0.16}$ | $0.81_{0.00}$ | $0.78_{0.09}$ | $0.90_{0.00}$ | $0.79_{0.15}$ | $\mathbf{0.93}_{0.02}$ |
| Falcon 40B | $0.79_{0.17}$ | $0.93_{0.01}$ | $0.81_{0.09}$ | $0.91_{0.01}$ | $0.87_{0.13}$ | $\mathbf{0.95}_{0.00}$ |

Table 5: Comparison of 2-shot learning performance on the SST-2 dataset using ensembles of 5 templates and a single template. Standard deviations over 5 random seeds are in subscript, best accuracy for each model is in bold.

performance if a suboptimal template is sampled. Therefore, we report results for ensembles of size 5 and average the results over 5 random seeds.

Next, we evaluate the performance gains of Template Ensembles for different prediction methods. Our findings in Tables 5 and 20 and Appendices K and L indicate that ensembles increase the accuracy for all evaluated models and prediction methods. Most importantly, they also significantly reduce the variance caused by the template choice for most setups. Therefore, we conclude that template ensembling allows to preserve the increase in accuracy provided by ICL modifications while mitigating the template sensitivity issue.

## 7 Conclusion

In this work, we study the inconsistencies in the evaluation of in-context learning advancements introduced by the template sensitivity of large language models. Specifically, we find that ICL improvements exhibit high variation across template formats and that it is not possible to reuse the same template across different modifications. This aspect is often overlooked in prior work, despite the fact that the impact of template selection on prediction accuracy may be comparable with the choice of demonstrations or prompting methods.

While we propose Template Ensembles as an initial solution to this problem, the general sensitivity of language models to minor prompt variations is yet to be addressed. Consequently, we believe that the research community should take this problem into account when developing new models, evaluation benchmarks, or in-context learning methods.

## Limitations

Due to limited computational resources and the high cost for evaluation on a large range of models, we only focus on four classification datasets. Moreover, we only compare two example selection methods to a random baseline, potentially overlooking other effective approaches.

Additionally, the space of templates could be expanded for more comprehensive experimentation. For example, we did not explore label mapping, including random labels, which is an important aspect of the template.

We would like to notice that our study focuses on a template selection impact on a performance and a degree of template transfer between different setups but not on templates themselves. Future work should further analyze not only which templates lead to a change in performance but also on why they affect it.

## References

AI@Meta. 2024. Llama 3 model card.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model. *Preprint*, arXiv:2204.06745.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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, Toronto, Canada. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek

Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. 2019. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 101–109, New York, NY, USA. Association for Computing Machinery.

Yu Fei, Yifan Hou, Zeming Chen, and Antoine Bosselut. 2023. Mitigating label biases for in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14014–14031, Toronto, Canada. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2022. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*.

Zhixiong Han, Yaru Hao, Li Dong, Yutao Sun, and Furu Wei. 2023. Prototypical calibration for few-shot learning of language models. In *The Eleventh International Conference on Learning Representations*.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *Preprint*, arXiv:2203.15556.

Bairu Hou, Joe O'Connor, Jacob Andreas, Shiyu Chang, and Yang Zhang. 2023. PromptBoosting: Black-box text classification with ten forward passes. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 13309–13324. PMLR.

Jaccard. 1912. The distribution of the flora of the alpine zone. In *New Phytologist*, volume 11, pages 37–50.

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*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Preprint*, arXiv:1612.01474.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021. What makes good in-context examples for gpt-3? In *Workshop on Knowledge Extraction and Integration for Deep Learning Architectures; Deep Learning Inside Out*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022a. 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. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022b. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Preprint*, arXiv:2104.08786.

Xinxi Lyu, Sewon Min, Iz Beltagy, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Z-icl: Zero-shot in-context learning with pseudo-demonstrations. *Preprint*, arXiv:2212.09865.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022a. Noisy channel language model prompting for few-shot text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5316–5330, Dublin, Ireland. Association for Computational Linguistics.

Sewon Min, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Noisy channel language model prompting for few-shot text classification. *Preprint*, arXiv:2108.04106.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022c. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022d. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what art? a call for multi-prompt llm evaluation. *Preprint*, arXiv:2401.00595.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Computer Vision – ECCV 2020*, pages 681–699, Cham. Springer International Publishing.

Tai Nguyen and Eric Wong. 2023. In-context example selection with influences. *Preprint*, arXiv:2302.11042.

Oleg Platonov, Denis Kuznedelev, Michael Diskin, Artem Babenko, and Liudmila Prokhorenkova. 2023. A critical look at the evaluation of GNNs under heterophily: Are we really making progress? In *The Eleventh International Conference on Learning Representations*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia

Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Karen Fort, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

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. *Preprint*, arXiv:2310.11324.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

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. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *Preprint*, arXiv:2301.11916.

Lucas Weber, Elia Bruni, and Dieuwke Hupkes. 2023. The icl consistency test. *Preprint*, arXiv:2312.04945.

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

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, and Tengyu Ma. 2023. Larger language models do in-context learning differently. *Preprint*, arXiv:2303.03846.

Zhiyong Wu, Yaoxiang Wang, Jiacheng Ye, and Lingpeng Kong. 2023. Self-adaptive in-context learning: An information compression perspective for in-context example selection and ordering. *Preprint*, arXiv:2212.10375.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

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.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. *CoRR*, abs/2102.09690.

Han Zhou, Xingchen Wan, Lev Proleev, Diana Mincu, Jilin Chen, Katherine A Heller, and Subhrajit Roy. 2024. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*.

## A Templates from Prior Work

Tables 6 to 9 provide a comparison of all the templates used in the works presenting all methods we evaluate. Noticeably, prompt formats (and the choice of label words for some formats and datasets) used in works proposing investigated methods have no intersection. This is also concerning, since the original papers proposing these methods refer to each other. For instance, CHANNEL prompting outperforming CALIBRATION in Min et al. (2022b) might be explained by selecting a more favorable set of templates for the method proposed in the paper rather than by the advantages of the method itself.

## B Model Selection

Our initial evaluation pool consisted of 23 models. We evaluated each of them in 0-shot and 2-shot settings with three prediction methods on four datasets, resulting in 12 runs. For each run in both 0-shot and 2-shot setups, we compare the model performance averaged over 10 random templates.

Based on the results presented in Table 10, we restricted the final pool of models for evaluation to those that have a consistent increase in performance in the 2-shot setting, in other words, to those demonstrating a performance boost from ICL. More specifically, we kept the models that had 8 or more wins in 2-shot evaluation against 0-shot.

## C Full Baseline Results

Table 11 shows the results of evaluation of all 19 models in the default setting with a varying number of few-shot examples. These results illustrate that the template sensitivity issue is present in all models regardless of their size, and is not efficiently mitigated with the increase in the number of demonstrations.

## D Template Parts Analysis

In addition to studying prompt format sensitivity in general, we analyze how each part of a template impacts model performance. For instance, it could be possible that the inclusion of a certain verbalizer in a template consistently leads to a decline in accuracy, irrespective of the other components.

To find that out, we decompose all templates into their parts and measure the distribution of scores for different variations of each component separately. The results presented in Figure 7 illustrate that even for state-of-the-art models, such as Llama 2 70B and Falcon 40B, many components exhibit high variance; also, the variance differs between two models. In other words, even if a certain template yields good performance and low variance for a given setup, it is not guaranteed to work consistently well in other setups, and changing a single component could have detrimental effects.

Along with the non-transferability of whole templates, we notice that individual components also do not transfer both between models and prediction methods. For instance, while "It was { }" ranks highest among output verbalizers for Llama 2 70B with the DIRECT prediction method, it is one of the worst for Falcon 40B.

Moreover, while a combination of best verbalizers is often a well-performing template, it is not necessarily the best one; the same is applicable for "bad" verbalizers too. For example, "input: { }\n sentiment: { }\n\n" is the best template for Falcon 40B with the DIRECT method, even though "sentiment: { }" is one of the "worst" output verbalizers for that model.

In summary, there is a complex interaction between the components of a template and their influence on model performance. We hypothesize that the transfer of both whole prompt templates and their parts is limited and requires further analysis.

## E Prediction Methods

We provide the results of advanced prediction methods evaluation for all models in 0-shot and 2-shot setting with random demonstrations in Table 10. We conclude from this comparison that neither of the advanced prediction strategies do not decrease prompt format sensitivity consistently across models and datasets. Moreover, when accounting for the spread in accuracy scores caused by this issue, the advantages of these methods over DIRECT become less apparent.

## F Reproduction of Results for Advanced Selection Methods

To evaluate the sensitivity of example selection methods to the template choice, we compare how the results reported in original works on these methods change when evaluated on a set of random templates instead of a predefined single one. For an accurate reproduction of original setups, we evaluate both Z-ICL and ITM using CHANNEL prompting with corresponding templates from Tables 6 and 7.

| Method | Input verbalizer | Output verbalizer | Intra-sep | Inter-sep | Label words |
|---|---|---|---|---|---|
| ITM | “sentence: {}” | “{}” | “ " | “ ” | negative, positive |
| z-ICL | “Review: {}” | “Sentiment: {}” | “\n” | “\n\n\n” | terrible, great |
| Channel | “{}” | “A {} one” | “ " | “ ” | terrible, great |
|  | “{}” | “It was {}.” | “ ” | “ ” | terrible, great |
|  | “{}” | “All in all {}.” | “ ” | “ ” | terrible, great |
|  | “{}” | “A {} piece.” | “ ” | “ ” | terrible, great |
| Calibration | “Review: {}” | “Answer: {}” | “\n” | “\n\n” | Negative, Positive |
|  | “Review: {}” | “Answer: {}” | “\n” | “\n\n” | bad, good |
|  | “Review: {}” | “Positive review? {}” | “\n” | “\n\n” | No, Yes |
|  | “Input: {}” | “Sentiment: {}” | “\n” | “\n\n” | Negative, Positive |
|  | “Review: {}” | “Positive: {}” | “\n” | “\n\n” | False, True |
|  | “My review for last night’s film: {}” | “The critics agreed that this movie was {}” | “ ” | “\n\n” | bad, good |
|  | “One of our critics wrote {}” | “Her sentiment towards the film was {}” | “ ” | “\n\n” | Negative, Positive |
|  | “In a contemporary review, Roger Ebert wrote {}.” | “Entertainment Weekly agreed, and the overall critical reception of the film was {}” | “ ” | “\n\n” | bad, good |
|  | “Review: {}” | “Question: Is the sentiment of the above review Positive or Negative? \nAnswer: {}” | “\n” | “\n\n” | Negative, Positive |
|  | “Review: {}” | “Question: Did the author think that the movie was good or bad?\nAnswer: {}” | “\n” | “\n\n” | bad, good |
|  | “Question: Did the author of the following tweet think that the movie was good or bad?\nTweet: {}” | “Answer: {}” | “\n” | “\n\n” | bad, good |
|  | “{}” | “My overall feeling was that the movie was {}” | “ ” | “\n\n” | bad, good |
|  | “{}” | “I {} the movie.” | “ ” | “\n\n” | hated, liked |
|  | “{}” | “My friend asked me if I would give the movie 0 or 5 stars, I said {}” | “ ” | “\n\n” | 0, 5 |

Table 6: All templates used in methods we evaluate for SST-2 dataset.

| Method | Input verbalizer | Output verbalizer | Intra-sep | Inter-sep | Label words |
|---|---|---|---|---|---|
| Channel | “{}” | “Topic: {}.” | “ ” | “ ” | Company, Educational Institution, Artist, Athlete, Office Holder, Building, Natural Place, Village, Animal, Plant, Album, Film, Written Work, Mean of Transportation |
| | “{}” | “Subject: {}.” | “ ” | “ ” | |
| | “{}” | “This is about {}.” | “ ” | “ ” | |
| | “{}” | “It is about {} one.” | “ ” | “ ” | |
| ITM | “{}” | “{}” | “ ” | “ ” | Same as above |
| Calibration | “Classify the documents based on whether they are about a [Label words] \n\n Article: {}” | “Answer: {}” | “\n” | “\n\n” | Company, School, Artist, Athlete, Politician, Building, Nature, Village, Animal, Plant, Album, Film, Book, Transportation |

Table 7: All templates used in methods we evaluate for DBPedia dataset.

| Method | Input verbalizer | Output verbalizer | Intra-sep | Inter-sep | Label words |
|---|---|---|---|---|---|
| Channel | “{}” | “{}” | “ ” | “ ” | Description, Entity, Expression, Human, Location, Number |
| | “{}” | “Q: {}." | “ ” | “ ” | |
| | “{}” | “Why {}?" | “ ” | “ ” | |
| | “{}” | “Answer: {}” | “ ” | “ ” | |
| Calibration | “Classify the questions based on whether their answer type is a [Label words]\n\n Question: {}” | “Answer Type: {}” | “\n” | “\n\n” | Number, Location, Person, Description, Entity, Abbreviation |

Table 8: All templates used in methods we evaluate for TREC dataset.

| Method | Input verbalizer | Output verbalizer | Intra-sep | Inter-sep | Label words |
|---|---|---|---|---|---|
| Channel | “{}” | “Topic: {}." | “ ” | “ ” | World, Sports, Business, Technology |
| | “{}” | “Subject: {}." | “ ” | “ ” | |
| | “{}” | “This is about {}." | “ ” | “ ” | |
| | “{}” | “It is about {} one." | “ ” | “ ” | |
| Calibration | “Article: {}” | “Answer: {}” | “\n” | “\n\n” | Same as above. |

Table 9: All templates used in methods we evaluate for AG News dataset.

| Model | N | SST-2 | | | DBPedia | | | AGNews | | | TREC | | | 2-shot wins |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Direct | Channel | Calib. | Direct | Channel | Calib. | Direct | Channel | Calib. | Direct | Channel | Calib. | |
| GPT-2 Large | 0 | $0.65_{0.09}$ | $0.72_{0.04}$ | $0.70_{0.06}$ | $0.34_{0.13}$ | $0.40_{0.07}$ | $0.49_{0.09}$ | $0.48_{0.16}$ | $0.56_{0.04}$ | $0.64_{0.14}$ | $0.25_{0.06}$ | $0.28_{0.11}$ | $0.30_{0.08}$ | 6/12 |
| | 2 | $0.59_{0.10}$ | $0.70_{0.12}$ | $0.62_{0.08}$ | $0.14_{0.08}$ | $0.53_{0.10}$ | $0.58_{0.10}$ | $0.32_{0.12}$ | $0.58_{0.07}$ | $0.57_{0.09}$ | $0.26_{0.09}$ | $0.29_{0.08}$ | $0.30_{0.06}$ | |
| GPT-2 XL | 0 | $0.76_{0.04}$ | $0.73_{0.05}$ | $0.70_{0.09}$ | $0.40_{0.05}$ | $0.43_{0.08}$ | $0.54_{0.09}$ | $0.52_{0.09}$ | $0.56_{0.05}$ | $0.65_{0.08}$ | $0.23_{0.03}$ | $0.24_{0.07}$ | $0.26_{0.05}$ | 5/12 |
| | 2 | $0.58_{0.11}$ | $0.71_{0.09}$ | $0.63_{0.11}$ | $0.15_{0.09}$ | $0.54_{0.09}$ | $0.50_{0.15}$ | $0.40_{0.20}$ | $0.61_{0.08}$ | $0.56_{0.15}$ | $0.26_{0.07}$ | $0.34_{0.09}$ | $0.33_{0.08}$ | |
| GPT-J | 0 | $0.71_{0.09}$ | $0.68_{0.08}$ | $0.68_{0.08}$ | $0.41_{0.07}$ | $0.44_{0.06}$ | $0.57_{0.08}$ | $0.61_{0.08}$ | $0.64_{0.03}$ | $0.64_{0.07}$ | $0.32_{0.05}$ | $0.20_{0.07}$ | $0.33_{0.04}$ | 9/12 |
| | 2 | $0.65_{0.14}$ | $0.77_{0.11}$ | $0.68_{0.11}$ | $0.25_{0.16}$ | $0.68_{0.06}$ | $0.71_{0.16}$ | $0.47_{0.19}$ | $0.67_{0.09}$ | $0.73_{0.11}$ | $0.26_{0.07}$ | $0.32_{0.09}$ | $0.33_{0.06}$ | |
| GPT-NeoX | 0 | $0.71_{0.08}$ | $0.67_{0.06}$ | $0.70_{0.09}$ | $0.48_{0.04}$ | $0.42_{0.05}$ | $0.60_{0.07}$ | $0.67_{0.06}$ | $0.56_{0.04}$ | $0.58_{0.07}$ | $0.30_{0.08}$ | $0.22_{0.06}$ | $0.32_{0.05}$ | 9/12 |
| | 2 | $0.69_{0.15}$ | $0.82_{0.06}$ | $0.79_{0.12}$ | $0.32_{0.19}$ | $0.67_{0.05}$ | $0.72_{0.16}$ | $0.52_{0.22}$ | $0.67_{0.10}$ | $0.70_{0.13}$ | $0.31_{0.08}$ | $0.32_{0.07}$ | $0.36_{0.08}$ | |
| OPT 1.3B | 0 | $0.78_{0.07}$ | $0.68_{0.07}$ | $0.79_{0.07}$ | $0.41_{0.05}$ | $0.33_{0.08}$ | $0.57_{0.12}$ | $0.49_{0.07}$ | $0.60_{0.01}$ | $0.67_{0.07}$ | $0.27_{0.04}$ | $0.17_{0.06}$ | $0.24_{0.03}$ | 6/12 |
| | 2 | $0.69_{0.15}$ | $0.80_{0.06}$ | $0.71_{0.16}$ | $0.21_{0.11}$ | $0.58_{0.08}$ | $0.61_{0.12}$ | $0.48_{0.24}$ | $0.66_{0.06}$ | $0.61_{0.11}$ | $0.27_{0.08}$ | $0.38_{0.09}$ | $0.35_{0.08}$ | |
| OPT 6.7B | 0 | $0.79_{0.07}$ | $0.67_{0.07}$ | $0.80_{0.07}$ | $0.46_{0.04}$ | $0.49_{0.05}$ | $0.61_{0.06}$ | $0.59_{0.08}$ | $0.61_{0.06}$ | $0.64_{0.07}$ | $0.24_{0.04}$ | $0.27_{0.09}$ | $0.33_{0.02}$ | 9/12 |
| | 2 | $0.67_{0.16}$ | $0.81_{0.06}$ | $0.72_{0.19}$ | $0.27_{0.14}$ | $0.69_{0.05}$ | $0.71_{0.17}$ | $0.45_{0.17}$ | $0.69_{0.09}$ | $0.70_{0.14}$ | $0.27_{0.08}$ | $0.34_{0.08}$ | $0.34_{0.07}$ | |
| OPT 30B | 0 | $0.79_{0.06}$ | $0.72_{0.05}$ | $0.77_{0.08}$ | $0.48_{0.04}$ | $0.48_{0.07}$ | $0.61_{0.08}$ | $0.64_{0.05}$ | $0.60_{0.06}$ | $0.65_{0.11}$ | $0.24_{0.03}$ | $0.26_{0.05}$ | $0.31_{0.01}$ | 8/12 |
| | 2 | $0.64_{0.17}$ | $0.79_{0.09}$ | $0.73_{0.17}$ | $0.34_{0.21}$ | $0.73_{0.05}$ | $0.78_{0.14}$ | $0.55_{0.19}$ | $0.69_{0.09}$ | $0.76_{0.11}$ | $0.31_{0.06}$ | $0.35_{0.09}$ | $0.33_{0.06}$ | |
| OPT 66B | 0 | $0.73_{0.12}$ | $0.73_{0.07}$ | $0.74_{0.10}$ | $0.41_{0.03}$ | $0.48_{0.07}$ | $0.61_{0.09}$ | $0.64_{0.07}$ | $0.58_{0.06}$ | $0.62_{0.07}$ | $0.26_{0.03}$ | $0.23_{0.06}$ | $0.31_{0.05}$ | 8/12 |
| | 2 | $0.65_{0.15}$ | $0.81_{0.08}$ | $0.76_{0.16}$ | $0.34_{0.16}$ | $0.77_{0.06}$ | $0.81_{0.15}$ | $0.45_{0.17}$ | $0.74_{0.05}$ | $0.70_{0.14}$ | $0.28_{0.07}$ | $0.38_{0.08}$ | $0.34_{0.07}$ | |
| BLOOM 1.7B | 0 | $0.68_{0.11}$ | $0.67_{0.06}$ | $0.68_{0.11}$ | $0.47_{0.03}$ | $0.47_{0.06}$ | $0.47_{0.07}$ | $0.61_{0.08}$ | $0.53_{0.04}$ | $0.58_{0.06}$ | $0.27_{0.04}$ | $0.24_{0.08}$ | $0.33_{0.03}$ | 9/12 |
| | 2 | $0.66_{0.12}$ | $0.75_{0.06}$ | $0.71_{0.10}$ | $0.27_{0.19}$ | $0.62_{0.08}$ | $0.57_{0.13}$ | $0.43_{0.19}$ | $0.59_{0.08}$ | $0.61_{0.11}$ | $0.31_{0.09}$ | $0.39_{0.07}$ | $0.37_{0.08}$ | |
| BLOOM 3B | 0 | $0.71_{0.10}$ | $0.71_{0.06}$ | $0.70_{0.08}$ | $0.39_{0.06}$ | $0.40_{0.07}$ | $0.48_{0.05}$ | $0.66_{0.02}$ | $0.48_{0.06}$ | $0.60_{0.08}$ | $0.22_{0.06}$ | $0.20_{0.07}$ | $0.20_{0.06}$ | 9/12 |
| | 2 | $0.72_{0.14}$ | $0.77_{0.09}$ | $0.77_{0.10}$ | $0.27_{0.21}$ | $0.67_{0.06}$ | $0.57_{0.14}$ | $0.45_{0.19}$ | $0.62_{0.07}$ | $0.67_{0.13}$ | $0.34_{0.09}$ | $0.35_{0.08}$ | $0.36_{0.09}$ | |
| BLOOM 7.1B | 0 | $0.72_{0.09}$ | $0.71_{0.06}$ | $0.68_{0.06}$ | $0.44_{0.05}$ | $0.45_{0.08}$ | $0.51_{0.08}$ | $0.64_{0.06}$ | $0.56_{0.04}$ | $0.64_{0.10}$ | $0.35_{0.07}$ | $0.22_{0.08}$ | $0.32_{0.04}$ | 9/12 |
| | 2 | $0.69_{0.15}$ | $0.76_{0.09}$ | $0.76_{0.11}$ | $0.26_{0.18}$ | $0.70_{0.06}$ | $0.67_{0.14}$ | $0.43_{0.17}$ | $0.69_{0.06}$ | $0.68_{0.12}$ | $0.33_{0.08}$ | $0.34_{0.07}$ | $0.36_{0.06}$ | |
| Pythia 6.9B | 0 | $0.75_{0.08}$ | $0.72_{0.05}$ | $0.69_{0.11}$ | $0.45_{0.05}$ | $0.43_{0.04}$ | $0.63_{0.09}$ | $0.58_{0.14}$ | $0.59_{0.04}$ | $0.64_{0.08}$ | $0.31_{0.07}$ | $0.21_{0.07}$ | $0.32_{0.03}$ | 8/12 |
| | 2 | $0.63_{0.12}$ | $0.78_{0.09}$ | $0.77_{0.11}$ | $0.28_{0.16}$ | $0.67_{0.08}$ | $0.68_{0.14}$ | $0.43_{0.17}$ | $0.68_{0.09}$ | $0.69_{0.14}$ | $0.34_{0.09}$ | $0.37_{0.07}$ | $0.38_{0.06}$ | |
| Pythia 12B | 0 | $0.73_{0.07}$ | $0.71_{0.08}$ | $0.69_{0.10}$ | $0.43_{0.05}$ | $0.43_{0.04}$ | $0.51_{0.18}$ | $0.61_{0.09}$ | $0.57_{0.05}$ | $0.65_{0.09}$ | $0.33_{0.06}$ | $0.23_{0.05}$ | $0.32_{0.03}$ | 8/12 |
| | 2 | $0.63_{0.13}$ | $0.79_{0.10}$ | $0.74_{0.12}$ | $0.29_{0.15}$ | $0.68_{0.07}$ | $0.71_{0.14}$ | $0.53_{0.18}$ | $0.68_{0.08}$ | $0.70_{0.12}$ | $0.29_{0.09}$ | $0.35_{0.06}$ | $0.33_{0.08}$ | |
| LLaMA 7B | 0 | $0.77_{0.08}$ | $0.70_{0.07}$ | $0.74_{0.12}$ | $0.46_{0.04}$ | $0.53_{0.05}$ | $0.55_{0.10}$ | $0.72_{0.05}$ | $0.65_{0.06}$ | $0.66_{0.06}$ | $0.34_{0.04}$ | $0.25_{0.07}$ | $0.30_{0.03}$ | 8/12 |
| | 2 | $0.72_{0.17}$ | $0.83_{0.07}$ | $0.83_{0.11}$ | $0.38_{0.21}$ | $0.76_{0.06}$ | $0.73_{0.13}$ | $0.61_{0.24}$ | $0.75_{0.08}$ | $0.72_{0.13}$ | $0.29_{0.10}$ | $0.38_{0.07}$ | $0.38_{0.11}$ | |
| LLaMA 13B | 0 | $0.81_{0.03}$ | $0.69_{0.07}$ | $0.77_{0.08}$ | $0.42_{0.04}$ | $0.52_{0.07}$ | $0.65_{0.11}$ | $0.74_{0.03}$ | $0.62_{0.07}$ | $0.73_{0.04}$ | $0.34_{0.04}$ | $0.18_{0.05}$ | $0.34_{0.03}$ | 10/12 |
| | 2 | $0.75_{0.17}$ | $0.83_{0.08}$ | $0.82_{0.14}$ | $0.38_{0.17}$ | $0.75_{0.06}$ | $0.80_{0.12}$ | $0.68_{0.15}$ | $0.75_{0.07}$ | $0.80_{0.08}$ | $0.35_{0.09}$ | $0.36_{0.08}$ | $0.42_{0.09}$ | |
| LLaMA 30B | 0 | $0.76_{0.08}$ | $0.71_{0.07}$ | $0.76_{0.08}$ | $0.51_{0.03}$ | $0.47_{0.09}$ | $0.67_{0.08}$ | $0.75_{0.04}$ | $0.66_{0.05}$ | $0.74_{0.06}$ | $0.33_{0.08}$ | $0.21_{0.05}$ | $0.30_{0.04}$ | 10/12 |
| | 2 | $0.78_{0.17}$ | $0.81_{0.10}$ | $0.83_{0.14}$ | $0.43_{0.19}$ | $0.76_{0.09}$ | $0.80_{0.09}$ | $0.65_{0.22}$ | $0.74_{0.13}$ | $0.78_{0.07}$ | $0.34_{0.11}$ | $0.41_{0.08}$ | $0.43_{0.13}$ | |
| LLaMA 65B | 0 | $0.78_{0.10}$ | $0.71_{0.05}$ | $0.75_{0.10}$ | $0.45_{0.05}$ | $0.49_{0.07}$ | $0.62_{0.08}$ | $0.74_{0.06}$ | $0.61_{0.07}$ | $0.74_{0.03}$ | $0.31_{0.06}$ | $0.19_{0.07}$ | $0.31_{0.03}$ | 11/12 |
| | 2 | $0.82_{0.17}$ | $0.84_{0.09}$ | $0.87_{0.13}$ | $0.45_{0.17}$ | $0.78_{0.06}$ | $0.80_{0.13}$ | $0.68_{0.20}$ | $0.78_{0.08}$ | $0.82_{0.05}$ | $0.38_{0.08}$ | $0.38_{0.09}$ | $0.45_{0.11}$ | |
| Falcon 1B | 0 | $0.72_{0.08}$ | $0.72_{0.03}$ | $0.73_{0.07}$ | $0.54_{0.03}$ | $0.55_{0.04}$ | $0.62_{0.10}$ | $0.68_{0.04}$ | $0.64_{0.06}$ | $0.63_{0.08}$ | $0.24_{0.04}$ | $0.25_{0.04}$ | $0.31_{0.02}$ | 9/12 |
| | 2 | $0.65_{0.17}$ | $0.77_{0.10}$ | $0.71_{0.17}$ | $0.36_{0.25}$ | $0.72_{0.05}$ | $0.74_{0.14}$ | $0.52_{0.17}$ | $0.72_{0.08}$ | $0.77_{0.08}$ | $0.26_{0.09}$ | $0.33_{0.06}$ | $0.33_{0.06}$ | |
| Falcon 7B | 0 | $0.72_{0.09}$ | $0.68_{0.05}$ | $0.73_{0.08}$ | $0.50_{0.06}$ | $0.51_{0.13}$ | $0.66_{0.06}$ | $0.75_{0.06}$ | $0.64_{0.03}$ | $0.72_{0.06}$ | $0.31_{0.04}$ | $0.21_{0.07}$ | $0.29_{0.03}$ | 10/12 |
| | 2 | $0.77_{0.16}$ | $0.78_{0.09}$ | $0.79_{0.15}$ | $0.40_{0.21}$ | $0.76_{0.06}$ | $0.80_{0.17}$ | $0.51_{0.20}$ | $0.76_{0.07}$ | $0.73_{0.12}$ | $0.32_{0.09}$ | $0.33_{0.08}$ | $0.37_{0.11}$ | |
| Falcon 40B | 0 | $0.76_{0.05}$ | $0.68_{0.07}$ | $0.74_{0.11}$ | $0.45_{0.03}$ | $0.57_{0.07}$ | $0.69_{0.08}$ | $0.75_{0.07}$ | $0.62_{0.07}$ | $0.72_{0.08}$ | $0.31_{0.07}$ | $0.27_{0.10}$ | $0.27_{0.02}$ | 11/12 |
| | 2 | $0.79_{0.17}$ | $0.81_{0.09}$ | $0.87_{0.13}$ | $0.42_{0.15}$ | $0.83_{0.06}$ | $0.85_{0.12}$ | $0.64_{0.23}$ | $0.79_{0.09}$ | $0.80_{0.06}$ | $0.36_{0.07}$ | $0.41_{0.06}$ | $0.45_{0.06}$ | |
| Llama 2 7B | 0 | $0.70_{0.12}$ | $0.59_{0.08}$ | $0.62_{0.16}$ | $0.35_{0.04}$ | $0.29_{0.09}$ | $0.21_{0.22}$ | $0.68_{0.04}$ | $0.45_{0.10}$ | $0.41_{0.23}$ | $0.30_{0.06}$ | $0.13_{0.05}$ | $0.15_{0.16}$ | 5/12 |
| | 2 | $0.66_{0.13}$ | $0.69_{0.10}$ | $0.66_{0.16}$ | $0.14_{0.11}$ | $0.17_{0.14}$ | $0.16_{0.18}$ | $0.37_{0.14}$ | $0.40_{0.11}$ | $0.42_{0.20}$ | $0.26_{0.09}$ | $0.29_{0.09}$ | $0.21_{0.19}$ | |
| Llama 2 13B | 0 | $0.77_{0.09}$ | $0.71_{0.04}$ | $0.74_{0.10}$ | $0.45_{0.03}$ | $0.59_{0.05}$ | $0.63_{0.10}$ | $0.75_{0.07}$ | $0.66_{0.05}$ | $0.76_{0.05}$ | $0.33_{0.03}$ | $0.27_{0.07}$ | $0.34_{0.03}$ | 9/12 |
| | 2 | $0.79_{0.17}$ | $0.82_{0.10}$ | $0.88_{0.09}$ | $0.40_{0.15}$ | $0.79_{0.06}$ | $0.83_{0.11}$ | $0.70_{0.15}$ | $0.76_{0.09}$ | $0.81_{0.05}$ | $0.32_{0.09}$ | $0.35_{0.11}$ | $0.46_{0.11}$ | |
| Llama 2 70B | 0 | $0.85_{0.05}$ | $0.68_{0.08}$ | $0.77_{0.10}$ | $0.48_{0.04}$ | $0.53_{0.06}$ | $0.72_{0.13}$ | $0.78_{0.06}$ | $0.64_{0.05}$ | $0.77_{0.06}$ | $0.34_{0.03}$ | $0.18_{0.07}$ | $0.34_{0.04}$ | 11/12 |
| | 2 | $0.83_{0.14}$ | $0.83_{0.08}$ | $0.88_{0.13}$ | $0.46_{0.15}$ | $0.79_{0.10}$ | $0.84_{0.12}$ | $0.76_{0.14}$ | $0.78_{0.09}$ | $0.81_{0.09}$ | $0.41_{0.07}$ | $0.41_{0.10}$ | $0.51_{0.06}$ | |
| Highest mean, % | | **37.5** | 32.5 | 30.0 | 5.0 | 12.5 | **82.5** | 32.5 | 10.0 | **57.5** | 25.0 | 15.0 | **60.0** | |
| Lowest std, % | | 12.5 | **80.0** | 7.5 | 42.5 | **55.0** | 2.5 | 20.0 | **62.5** | 17.5 | 17.5 | 25.0 | **57.5** | |

Table 10: Evaluation of advanced prediction methods for all models on 4 datasets in 0-shot and 2-shot with random demonstrations. Models that were removed from further evaluation are highlighted in gray. "Calib." stands for the Calibration prompting method.

.

| Model | SST-2 | | | DBPedia | | | AGNews | | | TREC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 2 | 4 | 0 | 2 | 4 | 0 | 2 | 4 | 0 | 2 | 4 |
| GPT-J | $0.71_{0.09}$ | $0.65_{0.14}$ | $0.66_{0.14}$ | $0.41_{0.07}$ | $0.25_{0.16}$ | $0.34_{0.18}$ | $0.61_{0.08}$ | $0.47_{0.19}$ | $0.48_{0.23}$ | $0.32_{0.05}$ | $0.26_{0.07}$ | $0.36_{0.11}$ |
| GPT-NeoX | $0.71_{0.08}$ | $0.69_{0.15}$ | $0.82_{0.12}$ | $0.48_{0.04}$ | $0.32_{0.19}$ | $0.37_{0.21}$ | $0.67_{0.06}$ | $0.52_{0.22}$ | $0.48_{0.22}$ | $0.30_{0.08}$ | $0.31_{0.08}$ | $0.40_{0.10}$ |
| OPT 6.7B | $0.79_{0.07}$ | $0.67_{0.16}$ | $0.80_{0.14}$ | $0.46_{0.04}$ | $0.27_{0.14}$ | $0.33_{0.18}$ | $0.59_{0.08}$ | $0.45_{0.17}$ | $0.47_{0.22}$ | $0.24_{0.04}$ | $0.27_{0.08}$ | $0.34_{0.10}$ |
| OPT 30B | $0.79_{0.06}$ | $0.64_{0.17}$ | $0.79_{0.14}$ | $0.48_{0.04}$ | $0.34_{0.21}$ | $0.39_{0.19}$ | $0.64_{0.05}$ | $0.55_{0.19}$ | $0.61_{0.18}$ | $0.24_{0.03}$ | $0.31_{0.06}$ | $0.34_{0.10}$ |
| OPT 66B | $0.73_{0.12}$ | $0.65_{0.15}$ | $0.84_{0.13}$ | $0.41_{0.03}$ | $0.34_{0.16}$ | $0.40_{0.16}$ | $0.64_{0.07}$ | $0.45_{0.17}$ | $0.53_{0.19}$ | $0.26_{0.03}$ | $0.28_{0.07}$ | $0.33_{0.09}$ |
| BLOOM 1.7B | $0.68_{0.11}$ | $0.66_{0.12}$ | $0.67_{0.13}$ | $0.47_{0.03}$ | $0.27_{0.19}$ | $0.31_{0.20}$ | $0.61_{0.08}$ | $0.43_{0.19}$ | $0.42_{0.19}$ | $0.27_{0.04}$ | $0.31_{0.09}$ | $0.36_{0.11}$ |
| BLOOM 3B | $0.71_{0.10}$ | $0.72_{0.14}$ | $0.76_{0.12}$ | $0.39_{0.06}$ | $0.27_{0.21}$ | $0.33_{0.21}$ | $0.66_{0.02}$ | $0.45_{0.19}$ | $0.46_{0.22}$ | $0.22_{0.06}$ | $0.34_{0.09}$ | $0.39_{0.12}$ |
| BLOOM 7.1B | $0.72_{0.09}$ | $0.69_{0.15}$ | $0.74_{0.15}$ | $0.44_{0.05}$ | $0.26_{0.18}$ | $0.32_{0.21}$ | $0.64_{0.06}$ | $0.43_{0.17}$ | $0.41_{0.21}$ | $0.35_{0.07}$ | $0.33_{0.08}$ | $0.38_{0.10}$ |
| Pythia 6.9B | $0.75_{0.08}$ | $0.63_{0.12}$ | $0.77_{0.14}$ | $0.45_{0.05}$ | $0.28_{0.16}$ | $0.35_{0.19}$ | $0.58_{0.14}$ | $0.43_{0.17}$ | $0.43_{0.20}$ | $0.31_{0.07}$ | $0.34_{0.09}$ | $0.38_{0.13}$ |
| Pythia 12B | $0.73_{0.07}$ | $0.63_{0.13}$ | $0.81_{0.13}$ | $0.43_{0.05}$ | $0.29_{0.15}$ | $0.35_{0.16}$ | $0.61_{0.09}$ | $0.53_{0.18}$ | $0.46_{0.22}$ | $0.33_{0.06}$ | $0.29_{0.09}$ | $0.35_{0.12}$ |
| LLaMA 7B | $0.77_{0.08}$ | $0.72_{0.17}$ | $0.85_{0.10}$ | $0.46_{0.04}$ | $0.38_{0.21}$ | $0.46_{0.20}$ | $0.72_{0.05}$ | $0.61_{0.24}$ | $0.64_{0.18}$ | $0.34_{0.04}$ | $0.29_{0.10}$ | $0.39_{0.15}$ |
| LLaMA 13B | $0.81_{0.03}$ | $0.75_{0.17}$ | $0.86_{0.14}$ | $0.42_{0.04}$ | $0.38_{0.17}$ | $0.52_{0.11}$ | $0.74_{0.03}$ | $0.68_{0.15}$ | $0.74_{0.13}$ | $0.34_{0.04}$ | $0.35_{0.09}$ | $0.42_{0.14}$ |
| LLaMA 30B | $0.76_{0.08}$ | $0.78_{0.17}$ | $0.87_{0.16}$ | $0.51_{0.03}$ | $0.43_{0.19}$ | $0.53_{0.16}$ | $0.75_{0.04}$ | $0.65_{0.22}$ | $0.71_{0.19}$ | $0.33_{0.08}$ | $0.34_{0.11}$ | $0.42_{0.16}$ |
| LLaMA 65B | $0.78_{0.10}$ | $0.82_{0.17}$ | $0.92_{0.10}$ | $0.45_{0.05}$ | $0.45_{0.17}$ | $0.52_{0.14}$ | $0.74_{0.06}$ | $0.68_{0.20}$ | $0.71_{0.17}$ | $0.31_{0.06}$ | $0.38_{0.08}$ | $0.47_{0.09}$ |
| Falcon 1B | $0.72_{0.08}$ | $0.65_{0.17}$ | $0.77_{0.15}$ | $0.54_{0.03}$ | $0.36_{0.25}$ | $0.44_{0.23}$ | $0.68_{0.04}$ | $0.52_{0.17}$ | $0.56_{0.19}$ | $0.24_{0.04}$ | $0.26_{0.09}$ | $0.31_{0.09}$ |
| Falcon 7B | $0.72_{0.09}$ | $0.77_{0.16}$ | $0.83_{0.16}$ | $0.50_{0.06}$ | $0.40_{0.21}$ | $0.49_{0.18}$ | $0.75_{0.06}$ | $0.51_{0.20}$ | $0.60_{0.19}$ | $0.31_{0.04}$ | $0.32_{0.09}$ | $0.39_{0.11}$ |
| Falcon 40B | $0.76_{0.05}$ | $0.79_{0.17}$ | $0.92_{0.07}$ | $0.45_{0.03}$ | $0.42_{0.15}$ | $0.54_{0.06}$ | $0.75_{0.07}$ | $0.64_{0.23}$ | $0.75_{0.09}$ | $0.31_{0.07}$ | $0.36_{0.07}$ | $0.46_{0.10}$ |
| Llama 2 13B | $0.77_{0.09}$ | $0.79_{0.17}$ | $0.92_{0.07}$ | $0.45_{0.03}$ | $0.40_{0.15}$ | $0.51_{0.09}$ | $0.75_{0.07}$ | $0.70_{0.15}$ | $0.76_{0.09}$ | $0.33_{0.03}$ | $0.32_{0.09}$ | $0.41_{0.14}$ |
| Llama 2 70B | $0.85_{0.05}$ | $0.83_{0.14}$ | $0.92_{0.09}$ | $0.48_{0.04}$ | $0.46_{0.15}$ | $0.60_{0.05}$ | $0.78_{0.06}$ | $0.76_{0.14}$ | $0.82_{0.05}$ | $0.34_{0.03}$ | $0.41_{0.07}$ | $0.51_{0.06}$ |

Table 11: Classification accuracy of all models on all datasets in the default setting. The results are aggregated over 10 templates for each set of random demonstrations, i.e., 10 runs for 0-shot and 30 runs for few-shot learning.

| Model | SST 2 | | | DBPedia 14 | | |
|---|---|---|---|---|---|---|
| | Paper | Reproduced | Random | Paper | Reproduced | Random |
| GPT2-Large | $0.86_{0.01}$ | $0.84_{0.02}$ | $0.80_{0.11}$ | $0.57_{0.03}$ | $0.62_{0.05}$ | $0.63_{0.07}$ |
| GPT2-XL | $0.83_{0.04}$ | $0.73_{0.13}$ | $0.75_{0.12}$ | $0.59_{0.03}$ | $0.61_{0.02}$ | $0.62_{0.09}$ |
| GPT-J | $0.88_{0.02}$ | $0.87_{0.02}$ | $0.86_{0.06}$ | $0.60_{0.04}$ | $0.54_{0.05}$ | $0.71_{0.05}$ |
| OPT-6.7B | $0.74_{0.03}$ | $0.86_{0.03}$ | $0.87_{0.04}$ | $0.29_{0.02}$ | $0.62_{0.04}$ | $0.72_{0.06}$ |
| LLaMA-7B | $0.61_{0.05}$ | $0.81_{0.12}$ | $0.87_{0.07}$ | $0.17_{0.01}$ | $0.68_{0.02}$ | $0.80_{0.06}$ |

Table 12: Mean accuracy and standard deviation of 4-shot learning with ITM demonstrations using the CHANNEL prediction method.

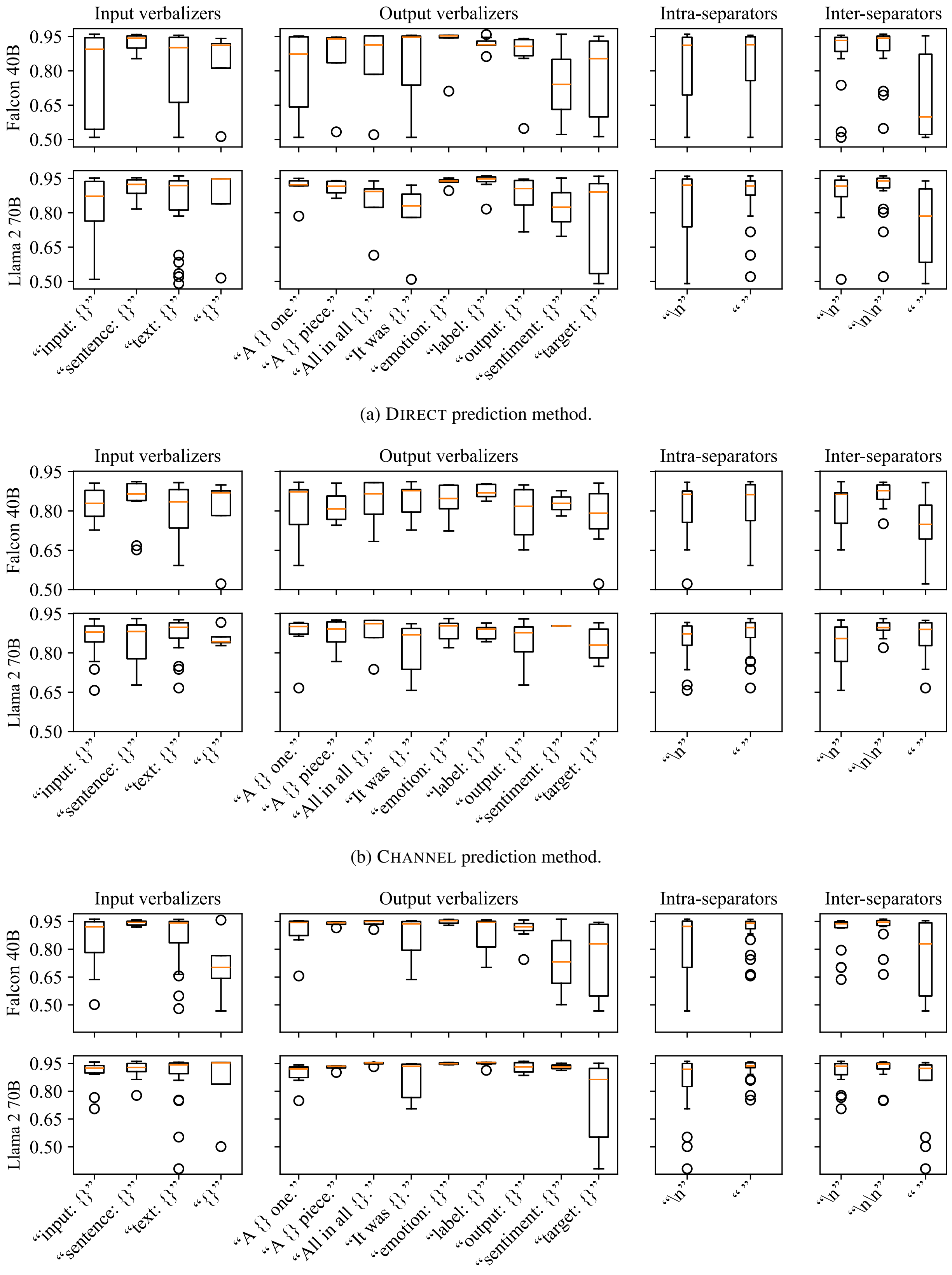
In Tables 12 and 13, we present the scores for a single template used in original implementations in the column "Reproduced", aggregating results over 5 example selection seeds. The "Random" column shows average scores for fixed demonstrations on a set of 10 random templates.

We observe that the results obtained in our code differ from the ones reported in the papers presenting both methods (the "Paper" column). The cause of this discrepancy is presumably the difference in tokenization during preprocessing of the datasets. Both methods use the same codebase with tokenization specific to GPT tokenizers, which results in a significant drop in quality for OPT and LLaMA models. By contrast, our tokenization approach is more general and preserves the ICL performance of these models.

| Model | Paper | Reproduced | Random |
|---|---|---|---|
| GPT-J | $0.83_{0.02}$ | $0.86_{0.00}$ | $0.61_{0.06}$ |
| GPT-NeoX | $0.79_{0.00}$ | $0.83_{0.00}$ | $0.80_{0.04}$ |

Table 13: Mean accuracy and standard deviation of 4-shot learning with Z-ICL demonstrations on SST-2 dataset using the CHANNEL prediction method.

From these results, we conclude that both methods are not robust to the template choice, as the mean performance decreases for multiple models while the standard deviation across seeds increases. Therefore, the gains from advanced example selection methods are caused to a certain degree by the choice of a proper prompt format rather than the retrieved demonstrations.



(a) DIRECT prediction method.

(b) CHANNEL prediction method.

(c) CALIBRATION prediction method.

Figure 7: Accuracy for evaluation of templates with fixed parts on the SST-2 dataset with RANDOM 2-shot for all prediction methods.

## G Example Selection Methods

Full results of evaluation of demonstration selection techniques in 4-shot learning using the DIRECT prediction method are presented in Table 14. The results highlight that advanced example selection techniques often perform comparably to the random choice baseline when evaluated on multiple templates. One might argue that the prompt format choice is inseparable from the method itself and thus such a comparison is invalid. However, since the best-performing formats do not transfer between models or demonstration sets of different sizes selected with the same method, a proper evaluation would require finding the best template for each setup. This procedure both is computationally expensive and difficult to accomplish, as authors of example selection methods frequently omit the description of their format selection algorithm.

## H Accuracy as a Function of Template Rank

We plot the dependence of accuracy on the rank of the template in Figure 8. The results are aggregated across 19 models. Each model was evaluated on 30 random templates with the DIRECT prediction method and the same set of 2 randomly selected demonstrations. We observe that for SST-2 and AGNews datasets, the mean quality of the tenth-best template is within 0.9 of the best template score, which we consider a successful transfer. Despite the more rapid decay for DBPedia and TREC, taking variation across models into account, we still count first 10 formats as performing on par with the best one.

## I Transfer Evaluation with Spearman Rank Correlation

One of the possible means to evaluate template transfer is to calculate the Spearman rank correlation between scores of all templates. As can be seen from Figure 9, this method yields higher correlations than IoU over 10 best formats, but the capacity for transfer is still far from perfect (for example, for SST-2 and TREC datasets).

## J IoU Transfer For All Datasets

Similarly to Figure 4, we provide Intersection-over-Union of 10 best prompt formats for all 19 models and all datasets explored in our work in Figure 10. These heatmaps illustrate that the transfer of best-performing templates between models is remarkably low for all datasets.

## K Additional Results For Template Ensembles

Tables 15 and 16 show the results of Template Ensembles evaluation on a broader set of models and datasets. For most setups, ensemble of size 5 exhibit better performance than a single template.

## L Evaluation of Instruct Models

We validate that our findings hold true even for the latest instruction-tuned models, such as Llama 3 8B Instruct and Mistral v0.3 7B Instruct. First, as we show in Table 17, in the baseline setting, an increase in number of demonstrations generally leads to a better performance of instruct models but does not significantly decrease variance of their final scores, similarly to what we observe in base models.

Second, Table 18 demonstrates that, after adjusting to template robustness, the default DIRECT prediction method performs on par with more advanced methods, e.g. CHANNEL and CALIBRATION, while sometimes having a less variance of the model's scores.

Finally, we observe that the examples retrieved by z-ICL method turn out to be consistently worse than two other methods. This is also in line with our observation for base models. However, in

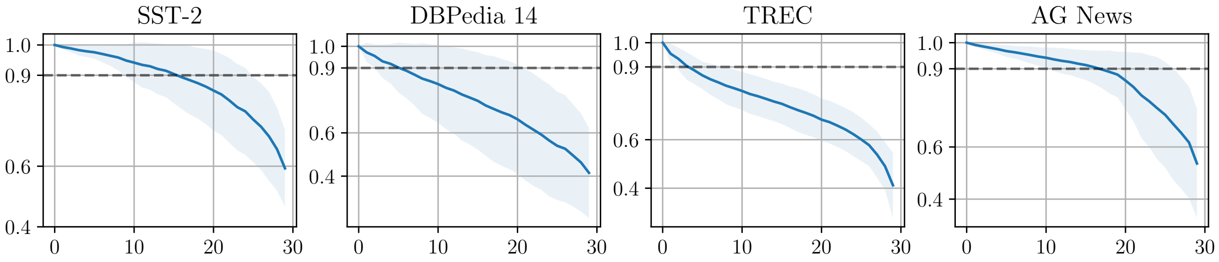


Figure 8: Relative quality of templates sorted by their classification accuracy. The shaded area indicates the standard deviation across 19 models.

| Model | SST-2 | | | DBPedia | | | AGNews | | | TREC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | ITM | z-ICL | Random | ITM | z-ICL | Random | ITM | z-ICL | Random | ITM | z-ICL |
| GPT-J | $0.66_{0.14}$ | $0.73_{0.16}$ | $0.66_{0.05}$ | $0.34_{0.18}$ | $0.46_{0.17}$ | $0.24_{0.10}$ | $0.48_{0.23}$ | $0.59_{0.12}$ | $0.45_{0.11}$ | $0.36_{0.11}$ | $0.25_{0.10}$ | $0.21_{0.04}$ |
| GPT-NeoX | $0.82_{0.12}$ | $0.73_{0.17}$ | $0.77_{0.09}$ | $0.37_{0.21}$ | $0.45_{0.19}$ | $0.25_{0.11}$ | $0.48_{0.22}$ | $0.58_{0.17}$ | $0.46_{0.12}$ | $0.40_{0.10}$ | $0.28_{0.09}$ | $0.25_{0.07}$ |
| OPT 6.7B | $0.80_{0.14}$ | $0.77_{0.18}$ | $0.72_{0.09}$ | $0.33_{0.18}$ | $0.46_{0.17}$ | $0.24_{0.11}$ | $0.47_{0.22}$ | $0.49_{0.17}$ | $0.45_{0.14}$ | $0.34_{0.10}$ | $0.23_{0.07}$ | $0.22_{0.05}$ |
| OPT 30B | $0.79_{0.14}$ | $0.80_{0.16}$ | $0.79_{0.12}$ | $0.39_{0.19}$ | $0.50_{0.16}$ | $0.27_{0.11}$ | $0.61_{0.18}$ | $0.67_{0.14}$ | $0.54_{0.13}$ | $0.34_{0.10}$ | $0.25_{0.08}$ | $0.25_{0.06}$ |
| OPT 66B | $0.84_{0.13}$ | $0.83_{0.15}$ | $0.75_{0.08}$ | $0.40_{0.16}$ | $0.48_{0.14}$ | $0.25_{0.10}$ | $0.53_{0.19}$ | $0.61_{0.15}$ | $0.48_{0.12}$ | $0.33_{0.09}$ | $0.25_{0.09}$ | $0.21_{0.04}$ |
| BLOOM 1.7B | $0.67_{0.13}$ | $0.69_{0.14}$ | $0.64_{0.09}$ | $0.31_{0.20}$ | $0.39_{0.19}$ | $0.20_{0.07}$ | $0.42_{0.19}$ | $0.46_{0.16}$ | $0.45_{0.08}$ | $0.36_{0.11}$ | $0.26_{0.07}$ | $0.19_{0.06}$ |
| BLOOM 3B | $0.76_{0.12}$ | $0.72_{0.15}$ | $0.62_{0.07}$ | $0.33_{0.21}$ | $0.43_{0.18}$ | $0.19_{0.08}$ | $0.46_{0.22}$ | $0.50_{0.16}$ | $0.43_{0.12}$ | $0.39_{0.12}$ | $0.29_{0.10}$ | $0.23_{0.03}$ |
| BLOOM 7.1B | $0.74_{0.15}$ | $0.71_{0.15}$ | $0.62_{0.08}$ | $0.32_{0.21}$ | $0.44_{0.19}$ | $0.20_{0.08}$ | $0.41_{0.21}$ | $0.50_{0.17}$ | $0.42_{0.10}$ | $0.38_{0.10}$ | $0.32_{0.08}$ | $0.21_{0.05}$ |
| Pythia 6.9B | $0.77_{0.14}$ | $0.72_{0.17}$ | $0.70_{0.10}$ | $0.35_{0.19}$ | $0.47_{0.17}$ | $0.22_{0.10}$ | $0.43_{0.20}$ | $0.56_{0.16}$ | $0.46_{0.12}$ | $0.38_{0.13}$ | $0.31_{0.08}$ | $0.22_{0.05}$ |
| Pythia 12B | $0.81_{0.13}$ | $0.72_{0.18}$ | $0.79_{0.10}$ | $0.35_{0.16}$ | $0.46_{0.14}$ | $0.24_{0.10}$ | $0.46_{0.22}$ | $0.57_{0.16}$ | $0.44_{0.13}$ | $0.35_{0.12}$ | $0.30_{0.09}$ | $0.22_{0.04}$ |
| LLaMA 7B | $0.85_{0.10}$ | $0.84_{0.16}$ | $0.70_{0.09}$ | $0.46_{0.20}$ | $0.51_{0.10}$ | $0.27_{0.08}$ | $0.64_{0.18}$ | $0.66_{0.15}$ | $0.52_{0.13}$ | $0.39_{0.15}$ | $0.28_{0.07}$ | $0.27_{0.05}$ |
| LLaMA 13B | $0.86_{0.14}$ | $0.85_{0.13}$ | $0.73_{0.11}$ | $0.52_{0.11}$ | $0.52_{0.10}$ | $0.31_{0.06}$ | $0.74_{0.13}$ | $0.77_{0.08}$ | $0.55_{0.08}$ | $0.42_{0.14}$ | $0.32_{0.13}$ | $0.29_{0.04}$ |
| LLaMA 30B | $0.87_{0.16}$ | $0.88_{0.11}$ | $0.67_{0.08}$ | $0.53_{0.16}$ | $0.57_{0.15}$ | $0.27_{0.06}$ | $0.71_{0.19}$ | $0.77_{0.08}$ | $0.45_{0.07}$ | $0.42_{0.16}$ | $0.30_{0.11}$ | $0.29_{0.07}$ |
| LLaMA 65B | $0.92_{0.10}$ | $0.91_{0.08}$ | $0.73_{0.10}$ | $0.52_{0.14}$ | $0.51_{0.13}$ | $0.30_{0.06}$ | $0.71_{0.17}$ | $0.84_{0.07}$ | $0.57_{0.08}$ | $0.47_{0.09}$ | $0.36_{0.09}$ | $0.34_{0.05}$ |
| Falcon 1B | $0.77_{0.15}$ | $0.77_{0.15}$ | $0.71_{0.09}$ | $0.44_{0.23}$ | $0.59_{0.18}$ | $0.23_{0.06}$ | $0.56_{0.19}$ | $0.58_{0.15}$ | $0.46_{0.10}$ | $0.31_{0.09}$ | $0.25_{0.10}$ | $0.19_{0.04}$ |
| Falcon 7B | $0.83_{0.16}$ | $0.82_{0.16}$ | $0.68_{0.11}$ | $0.49_{0.18}$ | $0.58_{0.09}$ | $0.27_{0.09}$ | $0.60_{0.19}$ | $0.63_{0.15}$ | $0.52_{0.12}$ | $0.39_{0.11}$ | $0.29_{0.09}$ | $0.25_{0.08}$ |
| Falcon 40B | $0.92_{0.07}$ | $0.92_{0.09}$ | $0.75_{0.11}$ | $0.54_{0.06}$ | $0.54_{0.07}$ | $0.28_{0.08}$ | $0.75_{0.09}$ | $0.80_{0.06}$ | $0.55_{0.11}$ | $0.46_{0.10}$ | $0.37_{0.11}$ | $0.26_{0.08}$ |
| Llama 2 13B | $0.92_{0.07}$ | $0.86_{0.14}$ | $0.75_{0.07}$ | $0.51_{0.09}$ | $0.53_{0.09}$ | $0.25_{0.06}$ | $0.76_{0.09}$ | $0.82_{0.05}$ | $0.54_{0.08}$ | $0.41_{0.14}$ | $0.34_{0.11}$ | $0.29_{0.04}$ |
| Llama 2 70B | $0.92_{0.09}$ | $0.91_{0.07}$ | $0.75_{0.09}$ | $0.60_{0.05}$ | $0.59_{0.09}$ | $0.28_{0.05}$ | $0.82_{0.05}$ | $0.85_{0.05}$ | $0.60_{0.07}$ | $0.51_{0.06}$ | $0.39_{0.10}$ | $0.34_{0.03}$ |
| Highest mean, % | 75.0 | 25.0 | 0.0 | 20.0 | 80.0 | 0.0 | 0.0 | 100.0 | 0.0 | 100.0 | 0.0 | 0.0 |
| Lowest std, % | 10.0 | 10.0 | 80.0 | 5.0 | 5.0 | 90.0 | 5.0 | 20.0 | 75.0 | 0.0 | 0.0 | 100.0 |

Table 14: Evaluation of advanced selection methods for all 19 models using DIRECT prediction method in 4-shot. Results are aggregated over 10 random templates for each of the three demonstrations selection seeds.

.

| Model | Direct | | Channel | | Calibration | |
|---|---|---|---|---|---|---|
| | Single | Ensemble | Single | Ensemble | Single | Ensemble |
| LLaMA 7B | $0.73_{0.17}$ | $0.81_{0.16}$ | $0.83_{0.06}$ | $0.89_{0.03}$ | $0.82_{0.13}$ | $\mathbf{0.92}_{0.02}$ |
| LLaMA 13B | $0.76_{0.17}$ | $0.82_{0.15}$ | $0.83_{0.08}$ | $\mathbf{0.89}_{0.03}$ | $0.78_{0.18}$ | $0.87_{0.08}$ |
| LLaMA 30B | $0.79_{0.15}$ | $0.86_{0.08}$ | $0.84_{0.08}$ | $0.87_{0.04}$ | $0.82_{0.16}$ | $\mathbf{0.93}_{0.01}$ |
| LLaMA 65B | $0.86_{0.13}$ | $\mathbf{0.95}_{0.00}$ | $0.85_{0.07}$ | $0.90_{0.02}$ | $0.89_{0.12}$ | $\mathbf{0.95}_{0.01}$ |
| Llama 2 13B | $0.79_{0.17}$ | $0.85_{0.09}$ | $0.82_{0.10}$ | $0.90_{0.02}$ | $0.88_{0.09}$ | $\mathbf{0.93}_{0.03}$ |
| Llama 2 70B | $0.83_{0.14}$ | $\mathbf{0.95}_{0.01}$ | $0.83_{0.08}$ | $0.92_{0.01}$ | $0.88_{0.13}$ | $0.94_{0.03}$ |
| Falcon 1B | $0.65_{0.17}$ | $0.74_{0.07}$ | $0.77_{0.10}$ | $\mathbf{0.89}_{0.01}$ | $0.71_{0.17}$ | $\mathbf{0.89}_{0.01}$ |
| Falcon 7B | $0.77_{0.16}$ | $0.81_{0.00}$ | $0.78_{0.09}$ | $0.90_{0.00}$ | $0.79_{0.15}$ | $\mathbf{0.93}_{0.02}$ |
| Falcon 40B | $0.79_{0.17}$ | $0.93_{0.01}$ | $0.81_{0.09}$ | $0.91_{0.01}$ | $0.87_{0.13}$ | $\mathbf{0.95}_{0.00}$ |

Table 15: Comparison of 2-shot learning performance on the SST-2 dataset using ensembles of 5 templates and a single template. Results are averaged over 5 random seeds.

contrast to our main results, where there was no evident winner between two other methods, the demonstrations selected with ITM method turn out to be slightly more robust to template choice for instruction-tuned models, as we show in Table 19.

### L.1 Ensemble Results

Table 20 shows that applying Template Ensembles method to instruct models results in improved mean classification accuracy with significantly reduced variance across different templates in all configurations that we test.

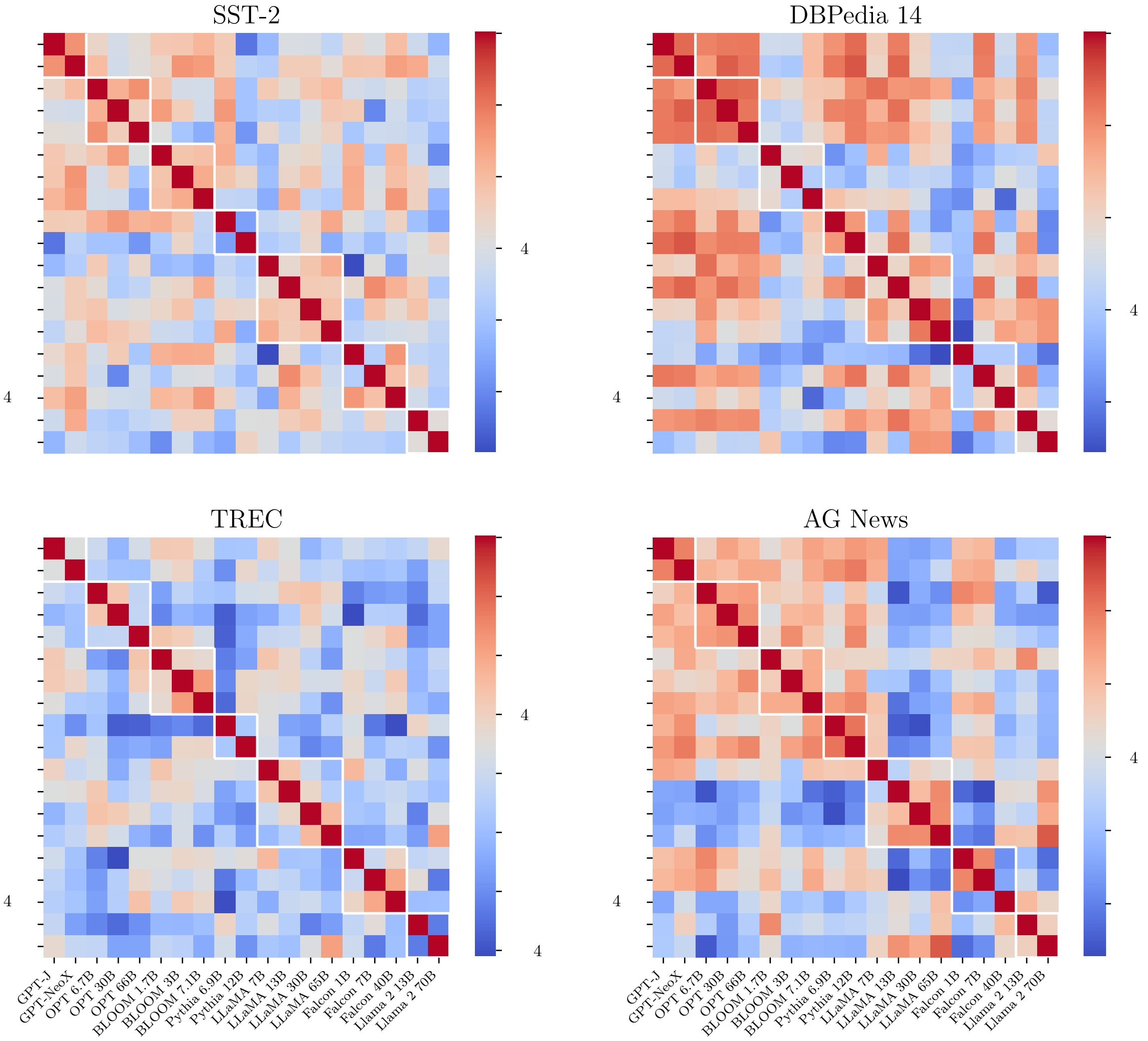


Figure 9: Spearman rank correlation over 30 templates for models evaluated in the DIRECT-RANDOM-2-shot setting.

| Model | Direct | | Channel | | Calibration | |
|---|---|---|---|---|---|---|
| | Single | Ensemble | Single | Ensemble | Single | Ensemble |
| LLaMA 7B | $0.29_{0.10}$ | $0.39_{0.04}$ | $0.38_{0.07}$ | $0.48_{0.10}$ | $0.38_{0.11}$ | $\mathbf{0.53}_{0.05}$ |
| LLaMA 13B | $0.35_{0.09}$ | $0.43_{0.09}$ | $0.36_{0.08}$ | $0.46_{0.03}$ | $0.42_{0.09}$ | $\mathbf{0.55}_{0.01}$ |
| LLaMA 30B | $0.34_{0.11}$ | $0.38_{0.04}$ | $0.41_{0.08}$ | $0.55_{0.02}$ | $0.43_{0.13}$ | $\mathbf{0.53}_{0.03}$ |
| LLaMA 65B | $0.38_{0.08}$ | $0.48_{0.07}$ | $0.38_{0.09}$ | $\mathbf{0.56}_{0.01}$ | $0.45_{0.11}$ | $0.55_{0.04}$ |
| Llama 2 13B | $0.32_{0.09}$ | $0.45_{0.11}$ | $0.35_{0.11}$ | $0.45_{0.02}$ | $0.46_{0.11}$ | $\mathbf{0.58}_{0.05}$ |
| Llama 2 70B | $0.41_{0.07}$ | $0.42_{0.09}$ | $0.41_{0.10}$ | $0.46_{0.03}$ | $0.51_{0.06}$ | $\mathbf{0.57}_{0.02}$ |
| Falcon 1B | $0.26_{0.09}$ | $0.33_{0.06}$ | $0.33_{0.06}$ | $0.43_{0.06}$ | $0.33_{0.06}$ | $\mathbf{0.44}_{0.01}$ |
| Falcon 7B | $0.32_{0.09}$ | $0.36_{0.05}$ | $0.33_{0.08}$ | $\mathbf{0.49}_{0.12}$ | $0.37_{0.11}$ | $0.48_{0.06}$ |
| Falcon 40B | $0.36_{0.07}$ | $0.39_{0.06}$ | $0.41_{0.06}$ | $0.52_{0.03}$ | $0.45_{0.06}$ | $\mathbf{0.53}_{0.05}$ |

Table 16: Comparison of 2-shot learning performance on the TREC dataset using ensembles of 5 templates and a single template. Results are averaged over 5 random seeds.

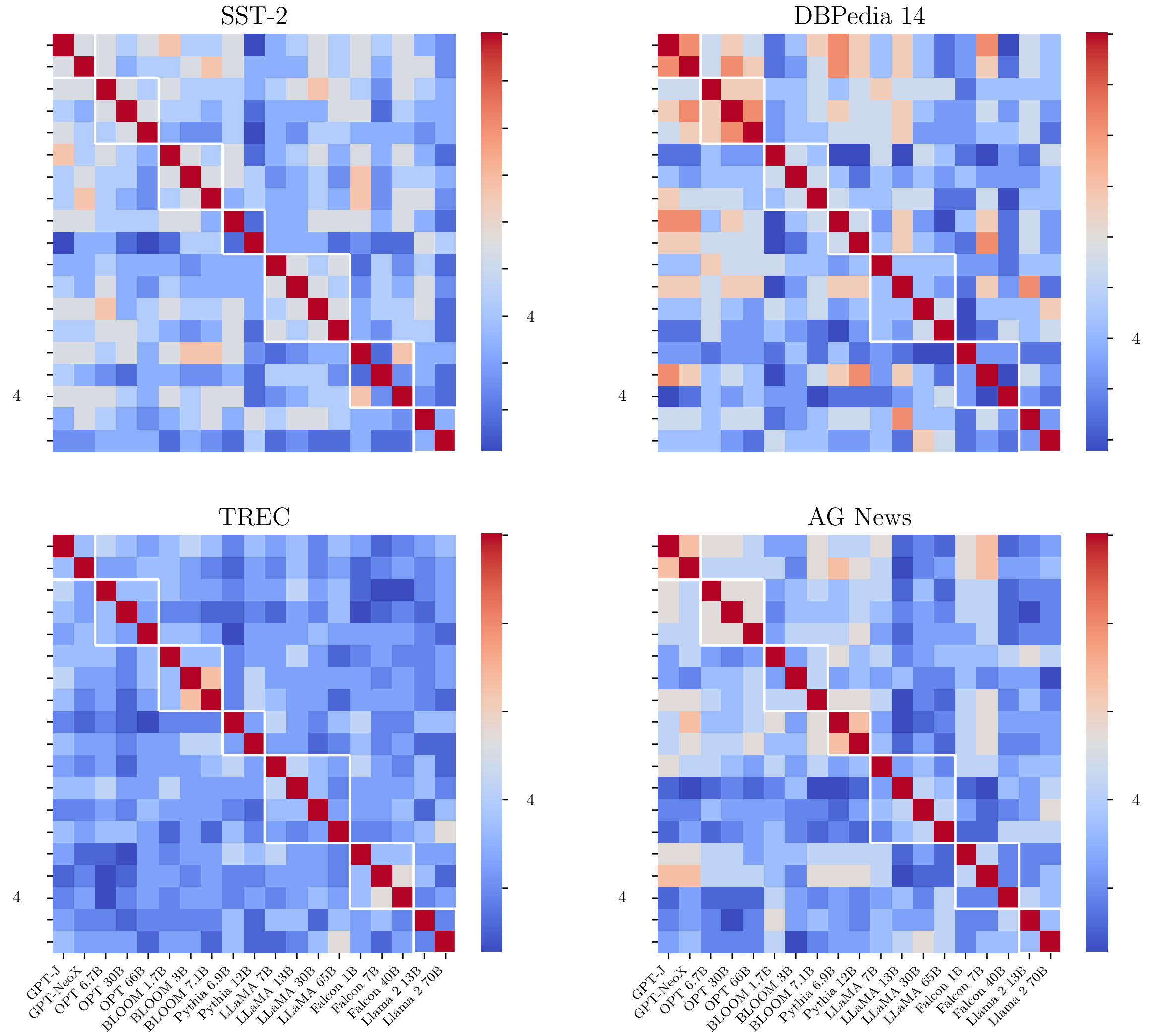


Figure 10: IoU of top-10 templates for all models evaluated in DIRECT-RANDOM-2-shot setting.

| Model | SST-2 | | | DBPedia | | | AGNews | | | TREC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 2 | 4 | 0 | 2 | 4 | 0 | 2 | 4 | 0 | 2 | 4 |
| Mistral v0.3 7B Instruct | $0.83_{0.07}$ | $0.90_{0.09}$ | $0.91_{0.12}$ | $0.50_{0.06}$ | $0.64_{0.05}$ | $0.65_{0.02}$ | $0.34_{0.05}$ | $0.40_{0.08}$ | $0.52_{0.08}$ | $0.71_{0.07}$ | $0.82_{0.07}$ | $0.76_{0.08}$ |
| Llama 3 8B Instruct | $0.77_{0.12}$ | $0.87_{0.13}$ | $0.91_{0.10}$ | $0.48_{0.02}$ | $0.55_{0.13}$ | $0.63_{0.05}$ | $0.24_{0.06}$ | $0.39_{0.09}$ | $0.48_{0.11}$ | $0.77_{0.06}$ | $0.77_{0.18}$ | $0.78_{0.10}$ |

Table 17: Classification accuracy of instruction-tuned models on all datasets in the default setting. The results are averaged over 10 templates for each set of random demonstrations (3 sets in total), i.e., 10 runs for 0-shot and 30 runs for few-shot learning.

| Model | N | SST-2 | | | DBPedia | | | AGNews | | | TREC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Direct | Channel | Calib. | Direct | Channel | Calib. | Direct | Channel | Calib. | Direct | Channel | Calib. |
| Mistral v0.3 7B Instruct | 0 | $0.83_{0.07}$ | $0.73_{0.05}$ | $0.83_{0.10}$ | $0.50_{0.06}$ | $0.47_{0.08}$ | $0.75_{0.09}$ | $0.34_{0.05}$ | $0.19_{0.08}$ | $0.30_{0.07}$ | $0.71_{0.07}$ | $0.59_{0.06}$ | $0.68_{0.06}$ |
| | 2 | $0.90_{0.09}$ | $0.83_{0.09}$ | $0.90_{0.12}$ | $0.64_{0.05}$ | $0.75_{0.04}$ | $0.91_{0.07}$ | $0.40_{0.08}$ | $0.41_{0.06}$ | $0.43_{0.06}$ | $0.82_{0.07}$ | $0.80_{0.02}$ | $0.80_{0.04}$ |
| Llama 3 8B Instruct | 0 | $0.77_{0.12}$ | $0.66_{0.04}$ | $0.74_{0.14}$ | $0.48_{0.02}$ | $0.48_{0.05}$ | $0.72_{0.08}$ | $0.24_{0.06}$ | $0.18_{0.08}$ | $0.22_{0.04}$ | $0.77_{0.06}$ | $0.57_{0.06}$ | $0.76_{0.06}$ |
| | 2 | $0.87_{0.13}$ | $0.78_{0.10}$ | $0.90_{0.09}$ | $0.55_{0.13}$ | $0.72_{0.04}$ | $0.85_{0.11}$ | $0.39_{0.09}$ | $0.36_{0.09}$ | $0.44_{0.06}$ | $0.77_{0.18}$ | $0.72_{0.10}$ | $0.80_{0.08}$ |

Table 18: Evaluation of advanced prediction methods for instruction-tuned models on 4 datasets in 0-shot and 2-shot with random demonstrations. "Calib." stands for the Calibration prompting method.

.

| Model | SST-2 | | | DBPedia | | | AGNews | | | TREC | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Random | ITM | z-ICL | Random | ITM | z-ICL | Random | ITM | z-ICL | Random | ITM | z-ICL |
| Mistral v0.3 7B Instruct | $0.91_{0.12}$ | $0.91_{0.10}$ | $0.86_{0.08}$ | $0.65_{0.02}$ | $0.65_{0.06}$ | $0.37_{0.05}$ | $0.52_{0.08}$ | $0.47_{0.10}$ | $0.37_{0.02}$ | $0.76_{0.08}$ | $0.83_{0.06}$ | $0.67_{0.06}$ |
| Llama 3 8B Instruct | $0.91_{0.10}$ | $0.92_{0.08}$ | $0.77_{0.08}$ | $0.63_{0.05}$ | $0.63_{0.08}$ | $0.26_{0.05}$ | $0.48_{0.11}$ | $0.45_{0.09}$ | $0.26_{0.04}$ | $0.78_{0.10}$ | $0.86_{0.04}$ | $0.53_{0.10}$ |

Table 19: Evaluation of advanced selection methods for instruction-tuned models using DIRECT prediction method in 4-shot. Results are aggregated over 10 random templates for each of the three demonstrations selection seeds.

.

| Model | Direct | | Channel | | Calibration | |
|---|---|---|---|---|---|---|
| | Single | Ensemble | Single | Ensemble | Single | Ensemble |
| | | | SST-2 | | | |
| Llama 3 8B Instruct | $0.87_{0.13}$ | $0.94_{0.01}$ | $0.78_{0.10}$ | $0.86_{0.02}$ | $0.90_{0.09}$ | $\mathbf{0.95}_{0.00}$ |
| Mistral v0.3 7B Instruct | $0.90_{0.09}$ | $0.94_{0.01}$ | $0.83_{0.09}$ | $0.91_{0.02}$ | $0.90_{0.12}$ | $\mathbf{0.95}_{0.00}$ |
| | | | DBPedia | | | |
| Llama 3 8B Instruct | $0.56_{0.12}$ | $0.64_{0.06}$ | $0.72_{0.04}$ | $0.79_{0.01}$ | $0.85_{0.11}$ | $\mathbf{0.95}_{0.02}$ |
| Mistral v0.3 7B Instruct | $0.64_{0.05}$ | $0.67_{0.02}$ | $0.75_{0.04}$ | $0.83_{0.02}$ | $0.91_{0.07}$ | $\mathbf{0.95}_{0.05}$ |
| | | | TREC | | | |
| Llama 3 8B Instruct | $0.39_{0.09}$ | $0.47_{0.01}$ | $0.36_{0.09}$ | $0.43_{0.03}$ | $0.44_{0.06}$ | $\mathbf{0.49}_{0.02}$ |
| Mistral v0.3 7B Instruct | $0.40_{0.08}$ | $\mathbf{0.48}_{0.02}$ | $0.41_{0.06}$ | $0.46_{0.04}$ | $0.43_{0.06}$ | $0.47_{0.03}$ |
| | | | AGNews | | | |
| Llama 3 8B Instruct | $0.77_{0.18}$ | $\mathbf{0.89}_{0.02}$ | $0.72_{0.10}$ | $0.81_{0.01}$ | $0.80_{0.08}$ | $0.84_{0.01}$ |
| Mistral v0.3 7B Instruct | $0.82_{0.07}$ | $\mathbf{0.86}_{0.02}$ | $0.80_{0.02}$ | $0.83_{0.01}$ | $0.80_{0.04}$ | $0.82_{0.02}$ |

Table 20: Comparison of 2-shot learning performance of Template Ensembles against other prediction methods on all datasets used in our work. We use ensembles of size 5 and average the results over 5 random seeds

.