# Ground-Truth Labels Matter: A Deeper Look into Input-Label Demonstrations

**Kang Min Yoo**[*#†‡§], **Junyeob Kim**[*§], **Hyuhng Joon Kim**[§], **Hyunsoo Cho**[§], **Hwiyeol Jo**[‡], **Sang-Woo Lee**[†‡♮], **Sang-goo Lee**[§], **Taeuk Kim**[#¶]
[§]Seoul National University, [†]NAVER AI Lab, [‡]NAVER CLOVA
[♮]Korea Advanced Institute of Science and Technology, [¶]Hanyang University
{juny116,heyjoonkim,johyunsoo,sglee}@europa.snu.ac.kr
{hwiyeol.jo,sang.woo.lee,kangmin.yoo}@navercorp.com
kimtaeuk@hanyang.ac.kr

## Abstract

Despite recent explosion of interests in in-context learning, the underlying mechanism and the precise impact of the quality of demonstrations remain elusive. Intuitively, ground-truth labels should have as much impact in in-context learning (ICL) as supervised learning, but recent work reported that the input-label correspondence is significantly less important than previously thought. Intrigued by this counter-intuitive observation, we re-examine the importance of ground-truth labels in in-context learning. With the introduction of two novel metrics, namely Label-Correctness Sensitivity and Ground-truth Label Effect Ratio (GLER), we were able to conduct quantifiable analysis on the impact of ground-truth label demonstrations. Through extensive analyses, we find that the correct input-label mappings can have varying impacts on the downstream in-context learning performances, depending on the experimental configuration. Through additional studies, we identify key components, such as the verbosity of prompt templates and the language model size, as the controlling factor to achieve more noise-resilient ICL.

## 1 Introduction

Large-scale language models (Rae et al., 2021; Chowdhery et al., 2022; Smith et al., 2022; Thoppilan et al., 2022) have shaped the NLP scene by introducing in-context learning (ICL) (Brown et al., 2020) as a novel approach to adapt language models for downstream tasks without explicit fine-tuning. ICL enables language models to learn and predict from task-specific prompts that contain demonstrations in the natural language format,

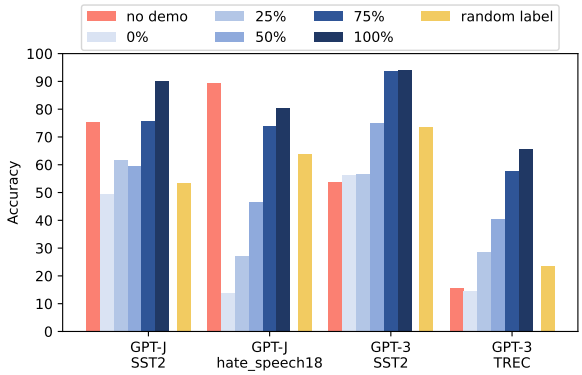



Figure 1: A demonstration of cases where the effect of the ground-truth label in in-context learning is much more significant than the aggregated results reported by Min et al. (2022b).

despite the language models were only trained to predict the next word token. Inspired by the new discovery, a flurry of recent work has investigated ways to explain and exploit the ICL mechanism (Schick and Schütze (2021a); Lu et al. (2022); *inter alia*), but it remains elusive.

Min et al. (2022b) have recently re-evaluated the role of input-label correspondence in demonstrations for ICL. Specifically, the authors have shown that the correct mapping between input and its label contributes less to the final performance than we thought compared to other aspects, including the format of demonstrations and the awareness of the input and label space. This finding is intriguing and has been sensational, as it is counter-intuitive to the expectation of how statistical learning typically works in supervised settings, and therefore it shows a potential of exploiting (few-shot) in-context learning given no real training data. For example, prior work established the strong impact of example ordering (Zhao et al., 2021), hence

[*]Equal contributions.
[#]Co-corresponding authors.



*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2422 - 2437
December 7-11, 2022 

in-context learning being less sensitive to the correctness of label demonstrations, which forms the basis of supervised learning, seems contradictory.

However, we encountered cases where the observation is inconsistent with the recent finding on the matter (Figure 1). Specifically, we found that the difference between the performance from the ground-truth label demonstration and that from entirely incorrect labels was as large as 80% (accuracy) for the hate speech dataset (de Gibert et al., 2018) on GPT-J (Wang and Komatsuzaki, 2021). Similar observations were found with the larger GPT-3 (Brown et al., 2020) model and other datasets (TREC (Li and Roth, 2002)). These cases illustrate how sensitive in-context learning can be to label demonstrations depending on the ICL settings. Thus, we cast a doubt on whether the trend can be generalized in diverse configurations, raising a call for an in-depth analysis of the phenomenon.

In this paper, we revisit the findings of Min et al. (2022b) and take a closer look into the importance of ground-truth labels for in-context learning. First, we point out limitations of the existing work. Then, we introduce novel metrics, namely Label-Correctness Sensitivity and Ground-Truth Label Effect Ratio (GLER), to reveal that the input-label correspondence plays a more vital role in contextual demonstration than previously considered. Furthermore, we show that the trend contradictory to the previous discovery becomes salient if we diverge the experimental settings (e.g., datasets, metrics, and templates) from the previous work. We observe the same trend in various language models, such as GPT-J and GPT-3 (Brown et al., 2020).

In addition, this paper uses statistics to provide a systematic and complementary perspective to the existing findings on the label-demonstration impact. To be specific, we combine linear regression and auxiliary metrics to conduct all-around and deeper analyses on how the ICL classification performance changes against label-demonstration corruption. To do so, we define the notion of sensitivity to quantify the degree to which the downstream classification performance changes when a model is subject to a fixed amount of label corruption. As a result, we demonstrate several noticeable patterns that support the claim that there is a considerable relationship between the performance and label correctness. It is worth noting that this trend was not clearly visible in the previous work, where the results of each dataset are macro-averaged rather than individually analyzed.

However, insensitivity, or robustness, towards the incorrectness of label-demonstrations is a useful property to have for many situations. For example, when augmenting an extremely small number of (e.g., less than four) examples using data augmentation techniques, exhibiting performance resilience towards prompt templates that consist of noisy synthetic examples as demonstrations is desirable. We further analyze how different factors of ICL, such as the inference method, the underlying language model, and the adoption of advanced ICL strategies, affect the performance sensitivity towards noises in input-label demonstrations, paving the way for a new approach to exploiting the demonstration insensitivity.

In summary, our contributions are as follows.

- We re-examine the recent findings on the phenomenon that the ICL performance is insensitive towards input-label demonstrations.
- We propose two new quantifiable metrics, sensitivity and GLER, to measure the impact of ground-truth label demonstrations on ICL.
- We conduct a thorough examination of how different components of ICL could impact the model's insensitivity towards label noises, allowing future work to exploit such property.

## 2 Looking Deeper into Ground-Truth Labels

Demonstrations of ground-truth labels[2], correctly paired with inputs, have been known to be a crucial factor of supervised learning, but a recent work by Min et al. (2022b) purportedly revealed the possibly counter-intuitive nature of label demonstrations in in-context learning (ICL). Specifically, the findings implied that the correctness of input-label correspondence in in-context demonstrations is not as important as we have thought. We name this phenomenon *input-label insensitivity*. Although the finding was supported by reasonably large-scale experiments, covering various experimental variables such as datasets, language models, in-context learning types, etc., we found that, through *deeper analysis* of the experiments, input-label insensitivity is not consistent across all experimental settings.

[2]Here, *label demonstrations* refer to the demonstration of input-label correspondence and not the demonstration of label space.

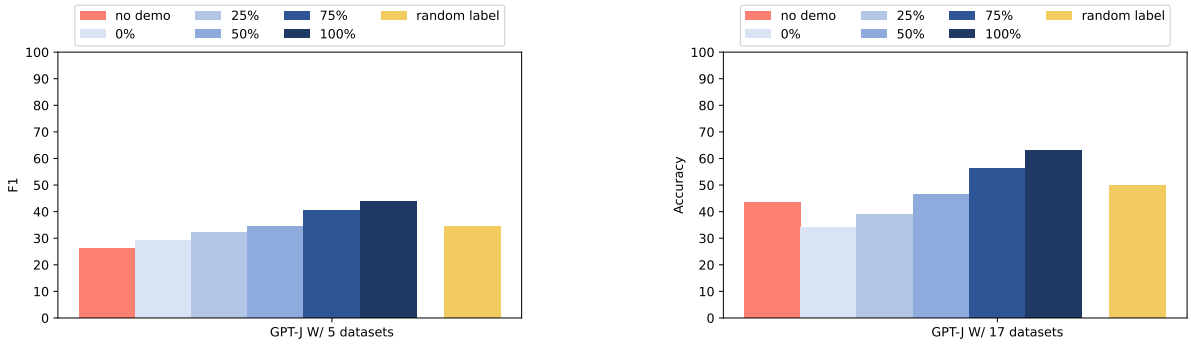



Figure 2: A counter-example of slightly varied but equally valid experimental settings is shown on the right, while the results from the prior experimental settings (Min et al., 2022b) is shown on the left. "No Demo" refers to the result without demonstrations and "Random Label" refers to the result with label demonstrations replaced with a random label uniformly sampled from the label space. Minor variations in the experimental settings could result in a large difference in the degree of which the ICL performance responds to the label corruption. More details on the experiment is described in Appendix A.

This section highlights the limitations of the existing work, proposes new metrics to quantify the impact of input-label correspondence, and finally presents deeper analyses of the ICL experiments utilizing the newly proposed metrics.

### 2.1 Limitations of the Existing Work

Min et al. (2022b) showed that replacing ground-truth labels in prompt demonstrations with incorrect labels marginally affects the *mean-aggregated overall* performance on selected datasets. Although the input-label insensitivity phenomenon was less prominent on GPT-J with the direct ICL method, the ICL still performed better when entirely incorrect labels were given than the absence of demonstrations (the zero-shot baseline), allegedly supporting the input-label sensitivity idea (Min et al., 2022b). However, we argue that there are mainly two limitations to the existing claim.

**Over-generalization** The existing claim suffers from over-generalization in two regards: (1) the mean-aggregated results fails to capture the insensitivity behavior in individual tasks and (2) the proposed experimental settings in the existing work is not general enough to be fully supportive of the claim. Mean-aggregation does not paint the full picture without the information on the variance. Furthermore, individual analyses on large-scale tasks are needed to obtain precise insight into input-label sensitivity. Our deeper analyses on the ICL experiments (§2.4) provide more evidence of this claim.

The second over-generalization is supported by the existence of a counter-example: higher input-label sensitivity observed from a slight varied but equally valid experimental settings (Figure 2). The subfigure on the left corresponds to the result of an existing set of experimental settings, where the Noisy Channel method (Min et al., 2022a) was used for ICL, the macro-F1 score for the evaluation metric, and the five classification datasets listed in the existing work. The subfigure on the right has been obtained using (*Direct*) method, the accuracy score as metric and results were aggregated from all 17 datasets listed in the existing work (see AppendixA).

**Lack of Quantification** Existing work relies on human judgement to determine the input-label sensitivity, which could be subjective. Furthermore, we are not only interested in *whether* the input-label insensitivity phenomenon exists but also *how* insensitive the ICL is towards the demonstrations, enabling us to exploit the phenomenon. Hence, a set of systematic quantification methods is needed to perform the deeper analyses.

### 2.2 Key Concepts

This subsection establishes key concepts and notations related to our analysis on the impact of input-label demonstrations and the downstream ICL performance. $x$ and $c$ denote the input and the label respectively. They exist in each respective input ($\mathcal{X}$) or label space ($\mathcal{C}$) associated with the dataset or task. A language model $P$ predicts the next token given the preceding tokens: $P(x_t|x_{<t})$. In ICL, a prompt $\mathcal{P}$ is designed to elicit particular behaviors from the language model. For exam-

ple, to utilize the language model as a text classifier, a prompt template $\mathcal{T}$ takes a set of examples $\mathcal{D}_{\text{ex}} = \{(x_1, c_1), ..., (x_k, c_k)\}$ and a test input $x$ to produce the prompt $\mathcal{P}$. The prompt is then fed into the language model to produce the most plausible continuation: $\text{argmax}_{x'} P(x'|\mathcal{P})$. A task-specific verbalizer $\mathcal{V}$ is designed to interpret the generated output $x'$ into the label space $\mathcal{C}$. We measure the performance $y$ of the language model $P$ and the prompt template $\mathcal{T}$ on a test set $\mathcal{D}_{\text{test}}$.

Our analyses mainly involve manipulating $\mathcal{T}$ and the example set $\mathcal{D}_{\text{ex}}$ to set-up baselines and conduct ablation studies. Key experimental set-ups include: **No Demo**, or denoted as "zero-shot", represents zero-shot predictions, where the prompt template $\mathcal{T}$ ignores $\mathcal{D}_{\text{ex}}$ and only uses the test input $x$: $P(c|x)$. The example set $\mathcal{D}_{\text{ex}}$ in $\alpha$**%-Correct** consists $k \times a/100$ correct input-label pairs and $k \times (1 - a/100)$ incorrect pairs where $(0 \leq a \leq 100)$. For **Random Label**, the labels $c$ in $\mathcal{D}_{\text{ex}}$ are replaced by uniform samples from the label space $\mathcal{C}$, and it is one of the key baselines of our studies. Additional details on the set-up variations are presented in Appendix A.

### 2.3 Metrics for Measuring the Impact of Input-Label Demonstrations

This section proposes two new metrics to quantify the impact of input-label demonstrations in ICL.

**Label-Correctness Sensitivity** We define label-correctness sensitivity, or **sensitivity** for short, as the degree of which the downstream classification performance changes when the model is subject to a fixed amount of label corruption. Sensitivity in the context of in-context learning demonstrations can be computed by conducting the single-scalar linear regression analysis on a performance metric (e.g., accuracy or F1-score) $y$ against the percentage of examples that are labelled correctly ($s$):

$$y = \beta_0 + \beta_1 s$$

where $\beta_0$ is the bias and $\beta_1$ is the coefficient of label correctness. The scalar value of the weight parameter $\beta_1$ is interpreted as the sensitivity measure. The data points for linear regression were obtained by following the experimental protocol proposed by Min et al. (2022b). The sensitivity measure can be interpreted as a linearly interpolated measure of performance degradation for each unit decrease in label correctness.

**Ground-Truth Label Effect Ratio (GLER)** Another way to understand the impact of labels, namely correct or ground-truth labels, is to quantify *how much the ground-truth labels improve the ICL performance compared to the random-label baseline*. The higher the gap, the bigger the impact the ground-truth labels have on the performance. The gap is then normalized by the performance difference between ground-truth labels and the absence-of-demonstration baseline (zero-shot):

$$\text{GLER} = \frac{y_{\text{GT}} - y_{\text{RL}}}{y_{\text{GT}} - y_{\emptyset}} \quad (1)$$

where $y_{\text{GT}}$ is the ground-truth label performance, $y_{\text{RL}}$ the random-label baseline (**Random-Label**), and $y_{\emptyset}$ the zero-shot performance. The denominator in Equation 1 is intended to allow the GLER metric to be compared across different tasks. Additionally, we clip GLER to be bounded between 0 and 1.

### 2.4 Deeper Analyses

This subsection performs deeper analyses using the aforementioned metrics to reveal additional insights into input-label insensitivity.

#### 2.4.1 Experimental Setup

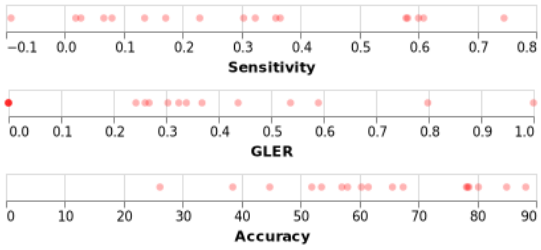
All of our experiments mentioned in the rest of the paper generally follows the experimental settings in Min et al. (2022b), where $\alpha$**%-Correct** is mainly utilized to conduct sensitivity analysis. However, there are key differences: (1) we do not employ label-length normalization (in our experiments length normalization does not always increase the performance), and there are minor template $\mathcal{T}$ design differences, including how the separator token interacts with the model and the dataset-specific implementation of data preprocessor; (2) we use accuracy, instead of F1-score, as the primary evaluation metrics for ICL performance. However, we do report the full results in Appendix A, along with the full details of the setup.

#### 2.4.2 Label Correctness Does Affect Performance

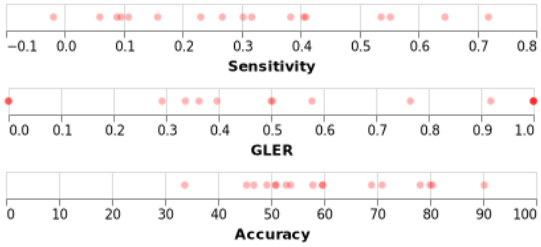
To analyze the overall sensitivity of performance under the variation of label correctness, we aggregate sensitivities across all 17 classification datasets and the results are shown in Table 1. The results show that the aggregated sensitivity is significantly high with good fit (in the range of 0.81-0.86) for all configurations. When tested on our specific setup,

| Method | Coefficient | Intercept | $R^2$ |
|---|---|---|---|
| GPT-NeoX Direct | 0.300 | 0.327 | 0.810 |
| GPT-J Direct | 0.309 | 0.291 | 0.861 |

Table 1: Aggregated linear regression analysis on the performance against the percentage of correct labels. "Ours" indicates that the data points for the linear regression analysis were obtained using our proposed experimental settings (Appendix A).




(a) GPT-NeoX




(b) GPT-J

Figure 3: Individual scatter-plots of the proposed metrics, sensitivity and GLER, across two models (GPT-NeoX and GPT-J) and 17 datasets. We also report the nominal ground-truth label accuracy values to further showcase the highly varied nature of the tasks.

the sensitivity was as high as 0.309, implying that, on average, there was a 0.309% drop in accuracy for each percentage drop in label correctness.

The trend of sensitivity, which is more apparent in our quantitative analysis, may have been overlooked due to the relative dwarfing effect from zero-shot (or "no demo") results in prior studies. The results also show that the sensitivity is lower in the Channel method,[3] suggesting that sensitivity can be significantly lowered with the employment of more advanced ICL methods.

#### 2.4.3 Label Demonstration Impact is Highly Varied Across Tasks and Settings

Although the aggregating analysis shows a general sensitive trend towards demonstration correctness,

[3]We hypothesize that this observation is attributed to the fact that, while generating longer sentences, prediction distribution from Channel model are more affected by the pretrained prior rather than the current context.

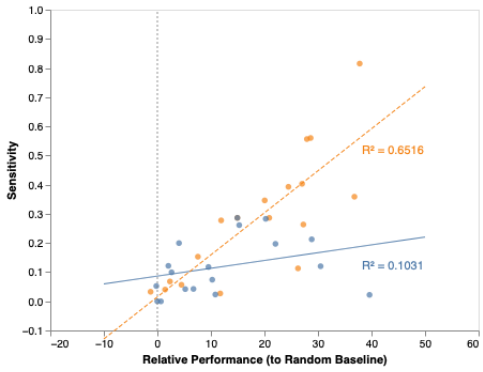



Figure 4: A scatter plot of sensitivities of 17 datasets against the corresponding task difficulties measured using the relative performance. The Direct approach is colored in orange and the Channel approach is colored in blue. The dashed vertical line indicates a neutral performance level where there is no difference with the random baselines. More details is found in Appendix C.

individual analyses shed deeper insight into the distribution of task sensitivities. Individual sensitivity plots are illustrated in Figure 3. Sensitivity can vary from small negative values (indicating increasing performance under increasing label corruption) to value as high as 0.815 (for the hate speech dataset), suggesting that summarizing the trend for all tasks and datasets may be difficult and that certain datasets may possess distributional properties that allow models to more easily exploit label demonstrations. This high-variance observation is valid for other metrics (GLER and the ground-truth label performance) as well. Further analyses are available in §3

#### 2.4.4 Sensitivity and Task Difficulty

Tasks where the model struggle to exploit in-context demonstrations may exhibit low sensitivity towards them, since understanding patterns in demonstrations is inherently linked with the ability to absorb demonstrative label-supervision. To confirm our theory, we conduct an analysis on the sensitivities of 17 datasets against the task difficulty. We define task difficulty as the relative performance of ground-truth label demonstrations compared to a baseline. Specifically, relative performance $y_{\text{rel}}$ is computed by $y_{\text{rel}} = y_{\text{GT}} - y_{\text{baseline}}$. We consider the *random baseline*.

Our analysis (Figure 11) shows that the model's performance sensitivity is strongly related to the difficulty of the task. The tasks, where the model exhibits low sensitivity (i.e. $< 0.1$), struggle to achieve meaningful classification performance.

This suggests that designing experiments with datasets that can be meaningfully solved using in-context learners may be more important than previously understood. Hence, the sensitivity measure by itself is insufficient for benchmarking the impact of input-label demonstrations.

## 3 When Do the Ground-Truth Labels Actually (Not) Matter?

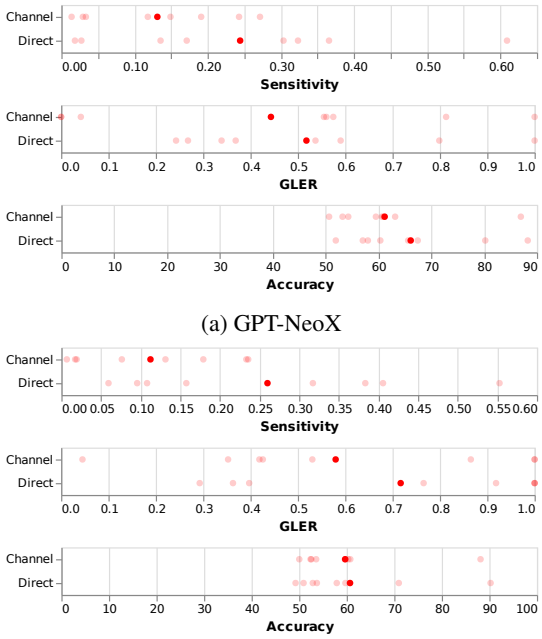
As revealed in our deeper analyses (§2.4), many factors including datasets and the choice of the ICL method can significantly affect the label-sensitivity. Gaining more understanding of the mechanism by which the input-label correspondence impacts the downstream ICL performance could enable us to systematically exploit the label-insensitivity phenomenon. For example, few-shot ICL models can be improved to tolerate label noises from synthetic data samples generated in the joint input and label spaces (Yoo et al., 2021).

To understand the conditions that reduce the label sensitivity, we conduct a series of experiments that investigate different factors contribute to the phenomenon quantified using the metrics proposed in §2.3. Namely, we consider the particular technical choice in carrying out ICL (whether to employ the noisy channel method (Min et al., 2022a) and the likelihood calibration (Zhao et al., 2021)), various properties of the prompt template (the number of in-context examples and the verbosity), and the model size.

**Sensitivity and GLER** Recall that the sensitivity measure is the nominal coefficient of the linear line fitted on the performance-versus-label-corruption data points. Since baselines can vary depending on the experimental setting, hyperparameters and the dataset[4], comparing the nominal sensitivity alone can be inconclusive, as the same degree of absolute improvement has different implications depending on the baseline level. To account for the variations in the *characteristics* of the task and the model, we consider GLER and the ground-truth label performance as the auxiliary measures in the following studies.

### 3.1 Techniques for In-context Learning

In-context learning, as first proposed by Brown et al. (2020), is a straightforward parameter-free approach, where the downstream task of interest is expressed as natural text demonstrations and used to conditionally generate from a language model. Recently, Min et al. (2022a) proposed Noisy Channel (denoted as *Channel*) that exploits the language generation capability of language models for discriminative tasks using the Bayes' Theorem. We compare the two ICL methods on all three (sensitivity, GLER, and the ground-truth label ICL accuracy) measures.




(a) GPT-NeoX



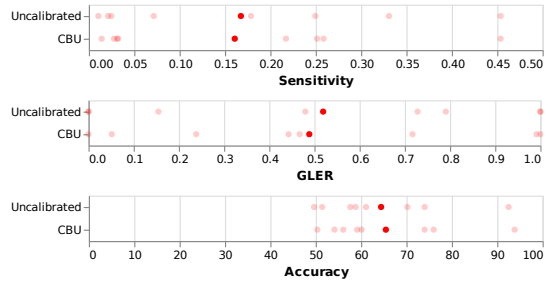

(b) GPT-J

Figure 5: The effect of switching the ICL inference method from *Direct* to *Channel*. Employing the Noisy Channel method improves insensitivity while improving the overall ICL performance.

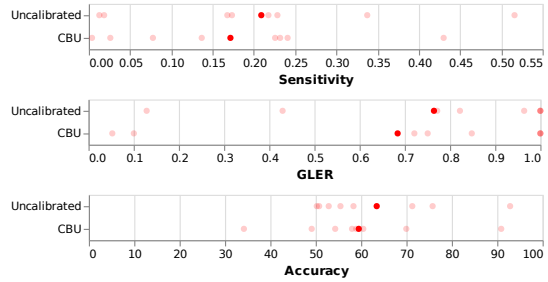
Results (Figure 5) show that Channel reduces the label-sensitivity on average compared to the original Direct method while maintaining the Accuracy on similar levels. The label insensitivity effect is observed in both GPT-NeoX and GPT-J.

Another recent advance in ICL, namely Calibrate Before Use (CBU), involves calibrating the output likelihood of the word tokens that correspond to the labels (Zhao et al., 2021). We conduct the same set of experiments with CBU applied and report all three metrics. As shown in Figure 6, the calibration technique reduces the label sensitivity while generally improving the ICL performance on both GPT-J and GPT-NeoX. Applying CBU can be an effective way to reduce label sensitivity while not sacrificing the performance.

[4]For example, under the same conditions (GPT-J and Direct inference), the random-label accuracy baseline is 28.08 for TREC and 53.58 for SST2.




(a) GPT-NeoX




(b) GPT-J

Figure 6: The effect of applying Calibrate Before Use (CBU) (Zhao et al., 2021). Label sensitivity decreases but the ground-truth label accuracy improves, making CBU ideal for sensitivity reduction. This trend is more apparent in the larger GPT-variant, GPT-NeoX (20B).

### 3.2 Prompt Templates

Various design choices in in-context prompt templates have significant impact on the downstream ICL performance (Reynolds and McDonell, 2021). A well-designed and verbose prompt template (e.g., a prompt with detailed description of the task) could allow in-context label demonstrations to have relatively less impact on ICL, thereby reducing the label-demonstration sensitivity.

This section mainly explores (1) the number of in-context examples and (2) the level of task description details. To quantify the impact of the number of in-context examples, we conduct the same set of experiments with varying number of in-context examples, ranging from 1 to 16. Results (Figure 7a) unsurprisingly show that the number of prompt examples is positively linked to all three metrics. Although sensitivity rises with the number of examples, this is due to the final ICL performance and the impact of ground-truth labels improving with more demonstration examples.

We also hypothesize that the level of task details contained in the prompt template also serves to relatively weaken the label demonstration impact. Results in Figure 7b confirm our hypothesis.

### 3.3 Model Sizes

The scale of the language model could influence how susceptible the model is to label noises within input-label demonstrations. The larger the model is, the more prior knowledge the model could leverage to reduce label sensitivity. To study whether this is the case, we analyze five different sizes of GPT-style language models, ranging from GPT-2 XL to GPT-3[5]. The choice of models and the corresponding number of parameters are listed in Figure 8. Results show that sensitivity is generally correlated with the model size, but we also observe a plateauing phenomenon after the GPT-J 6B scale. However, the results on the ICL performance with ground-truth label demonstrations shows that the performance scales well beyond the 6B mark,

## 4 Discussion

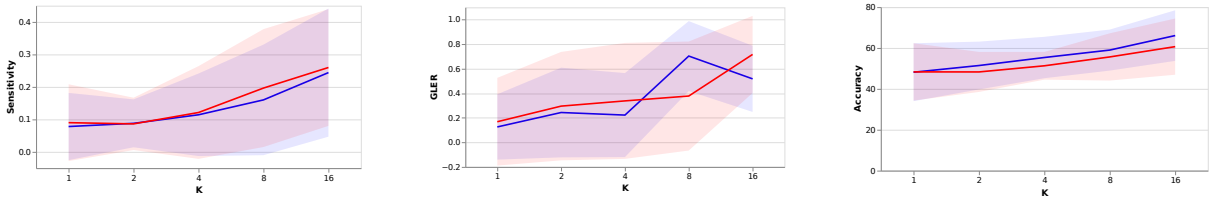
This section provides additional evidence that the demonstration of ground-truth labels can be more important than the previous finding suggests and that existing interpretation of the experimental results may have been obfuscated by the entanglement of various aspects of demonstrations.

### 4.1 The Complementary Relationship between Input-label Correspondence and Label-space Demonstrations

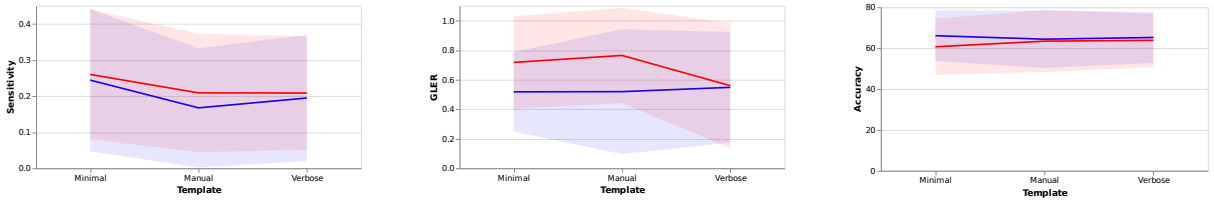
Input-label correspondence is just one of the aspects of possible in-context label demonstrations, the others including label-space demonstration. However, it is unclear whether label-space and input-label correspondence can complement each other in the absence of explicit demonstration of the other. For example, pretrained language models may be able to deduce sentence-sentiment mappings from the mentions of sentiment labels alone through inductive bias.

Prior work (Min et al., 2022b) showed significant performance degradation in the absence of both aspects of label demonstration, but the results beg the question: could the significant degradation have been caused by *complete lack of label demonstration*? To find out, we conduct additional ablation studies to study the performance under the demonstration of input-label pairings but not of the explicit label space which we call *prior-free* label experiments.

[5]Note that the general trend along the model scale persists with mixed language model architectures, as reported by Srivastava et al. (2022)




(a) Number of in-context examples




(b) Prompt template verbosity

Figure 7: Results for varying prompt sizes and prompt verbosity. The sensitivity, impact ratio, and final ground-truth label performance are all positively correlated with the number of prompt examples. For template verbosity, the sensitivity and the impact ratio decreases with the increase in versbosity, but the performance does not deterioriate. Results for GPT-NeoX (20B) are colored blue, while GPT-J (6B) is colored red.

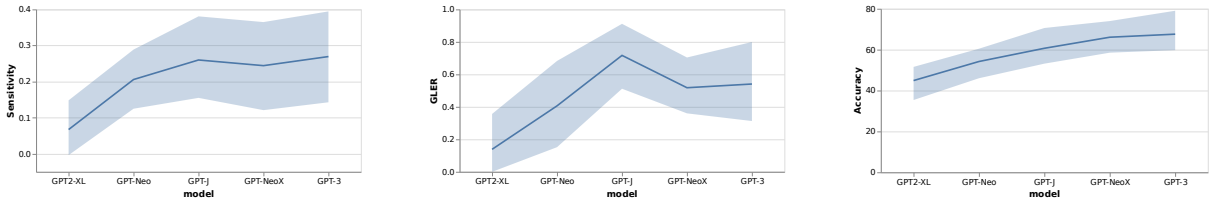



Figure 8: Comparison of sensitivity, GLER, and the ground-truth label ICL accuracy across different model sizes, ranging from GPT2-XL (1.5B) to GPT-3 (175B). Sensitivity and GLER plateau from the six-billion scale (GPT-J) while the ground-truth label performance continues to improve as the model size scales up.

Specifically, we study the case where class labels are replaced with *prior-free* labels while maintaining the correspondence between the input and the labels. For example, "positive" and "negative" labels in sentiment analysis can be replaced with "0" and "1" labels respectively, which do not reveal the information about the labels themselves. However, language models can still capture mild label-associations in abstract symbols through inductive bias (Ouyang et al., 2022). To diversity "prior-free" choices, we consider (1) random tokens from the language model's word space, (2) alphabets, and (3) numerical labels[6].

As shown in 9, results on *prior-free* labels outperform that of the random labels (with random input-label mappings), indicating that language models are capable of capturing the input-label correspondence even in the absence of label-space demonstrations. Among the prior-free results, we note that the alphabetical and numerical labels outperform random-token labels. This could be explained by the fact that, since random word tokens may introduce unintended biases through misleading association with unrelated word semantics, abstract labels provide better prior-free environment.

## 4.2 Change in label distribution may result the higher sensitivity.

The distribution of labels in demonstration is one of the critical factor for the prediction (Zhao et al., 2021). When data imbalance exists, corrupting the labels cause distributional shift which may lead performance change regardless of the input-label mappings. High sensitivity in imbalanced dataset may be due to this unintentional distributional shift. To analyze the impact of distributional shift, we conducted additional experiments using label balanced demonstrations for imbalanced dataset (hate_speech18, ethos-race, ethos-national_origin, ethos-religion).

As shown in 10, using balanced demonstrations degrade the performance and sensitivity when com-

[6]We exclude "0" since it is often associated with the state of nil

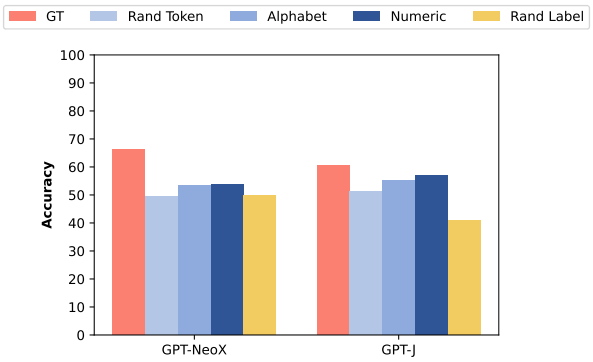
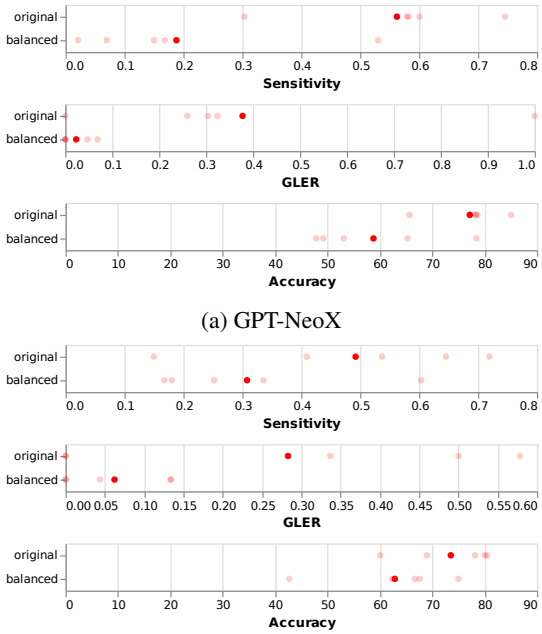



Figure 9: The results of "label prior-free" experiments (on 8 text classification datasets), where we control the prior information of the class labels. Here, the labels are replaced with tokens that are unrelated to the label semantics while still maintaining the input-label mappings. The replacement tokens include alphabet tokens, numeric tokens, and random word tokens from the language model's word space ("rand token"). The baselines obtained from the ground-truth labels and random labels are denoted as "GT" and "rand label" respectively. Results strongly suggest that language models are still able to utilize input-label demonstrations without access to label priors.

pared to demonstrations sampled from data distributions which supports our suspicion. On the other hand, average sensitivity are 0.189 and 0.308 (for GPT-NeoX and GPT-J respectively) even in balanced demonstrations setting which supports the importance of input-label demonstrations.

## 5 Related Work

As the scale of language models becomes larger (Rae et al., 2021; Chowdhery et al., 2022; Smith et al., 2022; Thoppilan et al., 2022), fine-tuning becomes prohibitively expensive due to the space and time complexities. As an alternative, in-context learning (ICL) (Brown et al., 2020) has shown to be an effective parameter-free learning strategy by prompting language models with task-specific prompt templates. Since then, a plethora of works has investigated both the properties of the learning mechanism (Schick and Schütze, 2021b; Reynolds and McDonell, 2021; Kim et al., 2021; Zhao et al., 2021; Lu et al., 2022; Min et al., 2022b). Although numerous efficient fine-tuning strategies have been proposed in the past (Li and Liang, 2021; Hu et al., 2022; Lester et al., 2021), the absence of an explicit training step in ICL has enabled it to retain its own class of adapting large-scale language models.




(a) GPT-NeoX




(b) GPT-J

Figure 10: The effect of using label balanced demonstrations in 5 imbalanced datasets. Employing the balanced demonstrations degrade all metrics due to the distributional shift in label demonstrations. However, sensitivity is still significant which supports the importance of input-label demonstrations.

## 6 Conclusion and Future Work

In this work, we took a closer look at how input-label relationships affect the in-context learning performance. To quantitatively analyze the impact of input-label mappings in in-context learning, we proposed novel metrics, GLER and input-label sensitivity. Through extensive experiments, we found that the integrity of the input-label mapping is a crucial factor in performing ICL. We also conducted ablation studies to reveal various conditions that allow ICL to improve insensitivity towards label corruptions (while still maintaining a healthy performance). For future work, based on the current findings, we will investigate whether we could exploit data augmentation for extremely low-resource situations for ICL.

## Limitations

**PLMs are over sensitive to the choice of prompts.** As it is widely known that performance of the PLMs is highly sensitive to the choice of the prompts (Brown et al., 2020; Lu et al., 2022; Zhao et al., 2021). Prompt engineering to find the optimal prompt was not feasible considering

the amount of datasets and settings that we experimented. The findings from this work may differ depending on the choice of prompts. However, to minimize this limitations the templates and prompts are adopted from well studied previous works as much as possible.

**Ground-truth label demonstrations are just one piece of the puzzle.** According the full analysis from Min et al. (2022b), other components of demonstrations not covered in this paper (e.g., input-space demonstrations) exhibit even stronger impacts on ICL. Although our experiments were designed to analyze solely the impact of input-label correspondence, disentangling diverse aspects of demonstrations is highly difficult as mentioned in section 4. Other factors such as label distribution may have unexpectedly influenced the results.

**Huggingface Implementation.** We use Huggingface implementation of GPT-NeoX. To our knowledge, current version of GPT-NeoX in Huggingface under performs when compared to the original implementations from Black et al. (2022).

## Acknowledgement

This work was mainly supported by SNU-NAVER Hyperscale AI Center and partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [No.2020-0-01373, Artificial Intelligence Graduate School Program (Hanyang University), No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]. Last but not least, we would like to express gratitude to Yejin Choi for the insightful discussions and feedback.

## References

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, ..., and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts. *arXiv:2202.01279*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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. In *Proceedings of the ACL Workshop on Challenges & Perspectives in Creating Large Language Models*.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, ..., and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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, ..., and Noah Fiedel. 2022. PaLM: Scaling language modeling with pathways. *arXiv:2204.02311*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, MLCW'05, page 177–190, Berlin, Heidelberg. Springer-Verlag.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Boseop Kim, HyoungSeok Kim, Sang-Woo Lee, Gichang Lee, Donghyun Kwak, Jeon Dong Hyeon, Sunghyun Park, Sungju Kim, Seonhoon Kim, Dongpil Seo, Heungsub Lee, Minyoung Jeong, Sungjae Lee, Minsub Kim, Suk Hyun Ko, Seokhun Kim, Taeyong Park, Jinuk Kim, Soyoung Kang, ..., and Nako Sung. 2021. What changes can large-scale language models bring? intensive study on HyperCLOVA: Billions-scale Korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3405–3424, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, ..., and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. 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 4582–4597, Online. Association for Computational Linguistics.

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

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

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *J. Assoc. Inf. Sci. Technol.*, 65(4):782–796.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. 2020. Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3458–3465, New York, NY, USA. Association for Computing Machinery.

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, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022b. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv:2022.12837*.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. Ethos: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. *OpenAI Blog*.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John

Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, ..., and Geoffrey Irving. 2021. Scaling language models: Methods, analysis & insights from training gopher. *arXiv:2112.11446*.

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

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

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

Emily Sheng and David Uthus. 2020. Investigating societal biases in a poetry composition system. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 93–106, Barcelona, Spain (Online). Association for Computational Linguistics.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv:2201.11990*.

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.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, ..., and Ziyi Wu. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv:2206.04615*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, ..., and Quoc Le. 2022. Lamda: Language models for dialog applications. *arXiv:2201.08239*.

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

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

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

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. 2021. CrossFit: A few-shot learning challenge for cross-task generalization in NLP. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7163–7189, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. 2021. GPT3Mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

# A Details on Our Experimental Settings

## A.1 Model

We mainly experiment with GPT-Neox 20B (Black et al., 2022) and GPT-J 6B (Wang and Komatsuzaki, 2021) which are publicly released, decoder-only, dense LMs. However, in Section 3.3 we also include GPT2-XL 1.5B (Radford et al., 2019), GPT-Neo 2.7B (Black et al., 2021), GPT-3 175B (Brown et al., 2020).

## A.2 Full Dataset

We evaluate on 17 text classification datasets covering diverse tasks including sentiment analysis, paraphrase detection, natural language inference, hate speech detection and diverse domains including science, social media, finance, and more. All datasets are from Huggigface datasets (Lhoest et al., 2021). Full list and details about the datasets are provided in Table 2.

As mentioned in Section 2.4.4, sensitivity highly depends on relative performance. In order to effectively capture correlation between sensitivity and diverse factors in Section 3, we evaluate on subset of 8 datasets, datasets with high relative performance, in Section 3. 8 datasets include glue-sst2, glue-rte, super_glue-cb, trec, financial_phrasebank, medical_questions_pairs, sick, and tweet_eval-hate. Due to limited resources, we only run experiments on 6 datasets in Section 3.3.

## A.3 Metric

We use accuracy as our primary metric. Accuracy is commonly used metric in multi-class classification which intuitively show how well the model performs. F1 score takes into account how the data is distributed thus it is useful when you have data with imbalance classes. However, F1 is less intuitive since it measures the *trade-off* between precision and recall. Moreover, F1 score can vary regarding the averaging method in multi-class classification.

## A.4 Template

We use 3 types of templates regarding engineering cost and verbosity of templates. First, as a baseline template we used minimal template following (Ye et al., 2021; Min et al., 2022b). We use minimal template throughout the paper. For ablation 3.2, we also evaluate manual templates and Verbose template. Templates are adopted from prior works (Brown et al., 2020; Zhao et al., 2021; Min et al., 2022b; Bach et al., 2022) if possible. Details and examples regarding the templates are in Table 3. Additionally, for Section 3.1 CBU experiment we use Manual template as the baseline since in our preliminary experiments, applying CBU in Minimal template degrade the performance in some cases.

Even though we use the same minimal template as Min et al. (2022b), there are minor difference in dataset-specific implementation of data preprocessor. (e.g., input sentences of glue_mrpc dataset used in Min et al. (2022b) have prefix "sentence1: ") Therefore, LMs may have slightly different behavior with same the dataset.

## A.5 Other details

Unless otherwise specified, we use k = 16 examples as demonstrations which are sampled at uniform from the training data. We run all experiments 5 times using different seeds. Due to limited resources, we only run experiments once for GPT-3. For all models expect for GPT-3, we used implementation and models from Huggingface transformers library (Wolf et al., 2020). For GPT-3 we used OpenAI API, assuming that model "davinci" is GPT-3 175B. When calculating the probability of label tokens, we do not normalize the score by the length of the tokens unlike in Min et al. (2022b). Our implementation is available at https://github.com/juny116/ICL-DeepSpeed.

## A.6 Corrupting input-label mapping

To see the detail impact of the ground truth input-label mapping, we revisit the experiments from Min et al. (2022b) Specifically, we replace fix amount of correct labels to incorrect labels in demonstrations and compare the end task performance.

- **No demonstrations** is a zero-shot prediction made via $argmax_{y \in C} P(y|x)$, where $x$ is the test input and $C$ is a small discrete set of possible labels. Verbalizers are used for mapping tokens to class.
- **Demonstrations w/ $a\%$ correct labels** consist $k \times a/100$ correct pairs and $k \times (1-a/100)$ incorrect pairs where $(0 \leq a \leq 100)$. A concatenation of k input-label pairs where $a\%$ labels are correct is used to make a prediction via $argmax_{y \in C} P(y|x_1, y_1, ..., x_k, y_k, x)$.
- **Demonstrations w/ random label** is formed with replacing correct labels to random labels

| Dataset | Train | Eval | Class |
|---|---|---|---|
| glue-sst2 (Socher et al., 2013) | 67,349 | 872 | 2 |
| glue-rte (Dagan et al., 2005) | 2,490 | 277 | 2 |
| glue-mrpc (Dolan and Brockett, 2005) | 3,668 | 408 | 2 |
| glue-wnli (Levesque et al., 2012) | 635 | 71 | 2 |
| super_glue-cb (de Marneffe et al., 2019) | 250 | 56 | 3 |
| trec (Voorhees and Tice, 2000) | 5,452 | 500 | 5 |
| financial_phrasebank (Malo et al., 2014) | 1,181 | 453 | 3 |
| poem_sentiment (Sheng and Uthus, 2020) | 843 | 105 | 3 |
| medical_questions_pairs (McCreery et al., 2020) | 2,438 | 610 | 2 |
| sick (Marelli et al., 2014) | 4,439 | 495 | 3 |
| hate_speech18 (de Gibert et al., 2018) | 8,562 | 2,141 | 4 |
| ethos-national_origin (Mollas et al., 2022) | 346 | 87 | 2 |
| ethos-race (Mollas et al., 2022) | 346 | 87 | 2 |
| ethos-religion (Mollas et al., 2022) | 346 | 87 | 2 |
| tweet_eval-hate (Barbieri et al., 2020) | 9,000 | 1,000 | 2 |
| tweet_eval-stance_atheism (Barbieri et al., 2020) | 461 | 52 | 3 |
| tweet_eval-stance_feminist (Barbieri et al., 2020) | 597 | 67 | 3 |

Table 2: Datasets used for the experiment.

that are randomly sampled at uniform from $C$. Since the labels are sampled at uniform from $C$, the distribution of labels in demonstration may change from sampled inputs.

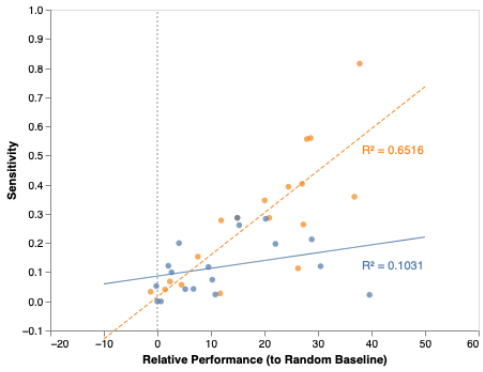
- **Demonstrations w/ shuffled label** is formed with randomly shuffling correct labels to other labels within the sampled k inputs. The distribution of labels in demonstration does not change from sampled inputs.
- **Majority class baseline** is a ratio of majority class within the test data. Since there are some datasets that have distributional imbalance, this can be a good indicator of how well the in-context learning is working.

## B Full Results

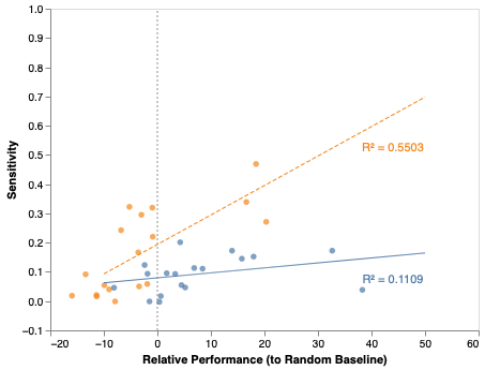
Full experiment results on 17 datasets with GPT-NeoX are in Table 4 and results on 17 datasets with GPT-NeoX are in Table 5.

## C More Results on the Sensitivity vs Task Difficulty Plot

Figure 11 shows scatter plots of sensitivities of 17 datasets against the corresponding task difficulties measured using the relative performance with respect to accuracy and F-1 scores. The Direct approach is colored in orange and the Channel approach is colored in blue. The dashed vertical line indicates a neutral performance level where there is no difference with the random baselines. The best-fit linear lines show a general trend of increasing sensitivity with less task difficulty. Low sensitivity is strongly related to high task difficulty. Also, the Channel approach helps in alleviating hypersensitivity towards task difficulty.




(a) Accuracy




(b) F1-score

Figure 11: Scatter plots of sensitivities of 17 datasets against the corresponding task difficulties measured using the relative performance with respect to each metrics.

## D Label-Correctness Correlation

The first step of understanding the interaction between performance and input-label demonstration is quantifying the correlation between the two variables. Although we considered this metric as one of the foundation quantifying measures, we omit the analyses results due to space constraints. The Pearson correlation analysis on GPT-J and the Direct approach (Figure 12) shows that the label-correctness correlation is strong (i.e. larger than 0.9) for most tasks on all performance measures. The macro-average correlation across 18 tasks is 0.895 with a p-value of 0.057, strongly supporting the linkage.

| Dataset | Manual Template | Verbalizer |
|---|---|---|
| glue-sst2 | Review: a smile on your face<br>Sentiment: | negative, positive |
| meddical_questions_pairs | The DVD-CCA then appealed to the state Supreme Court .<br>The question is: The DVD CCA appealed that decision to the U.S. Supreme Court . True or False?<br>answer: | False, True |
| glue-rte | Oil prices fall back as Yukos oil threat lifted<br>The question is: Oil prices rise. True or False?<br>answer: | True, False |
| super_glue-cb | That was then, and then's gone. It's now now. I don't mean I 've done a sudden transformation.<br>The question is: she has done a sudden transformation True or False?<br>answer: | True, False, Not sure |
| trec | Question: How can I find a list of celebrities ' real names ?<br>Type: | description, entity, expression, human, number, location |
| sick | The young boys are playing outdoors and the man is smiling nearby<br>The question is: The kids are playing outdoors near a man with a smile True or False?<br>answer: | True, Not sure, False |
| tweet_eval-hate | Tweet: Hundreds of Syrian refugees return home from Lebanon - ABC News<br>Sentiment: | favor, against |

| Dataset | Verbose Template | Verbalizer |
|---|---|---|
| glue-sst2 | Question : Is the following review positive or negative? a smile on your face<br>Answer: | negative, positive |
| meddical_questions_pairs | Question: Does the following two sentences mean the similar thing? True or False?<br>The DVD-CCA then appealed to the state Supreme Court .<br>The DVD CCA appealed that decision to the U.S. Supreme Court .<br>Answer: | False, True |
| glue-rte | Question: Does the first sentence entails the second sentence? True or False?<br>Oil prices fall back as Yukos oil threat lifted<br>Oil prices rise.<br>Answer: | True, False |
| super_glue-cb | Question: Does the first sentence entails the second sentence? True, False, or Neither?<br>That was then, and then's gone. It's now now. I don't mean I 've done a sudden transformation.<br>she has done a sudden transformation<br>Answer: | True, False, Neither |
| trec | Question: Which category best describes the following sentence?<br>How can I find a list of celebrities ' real names ?<br>Answer: | description, entity, expression, human, number, location |
| sick | Question: Does the first sentence entails the second sentence? True, False, or Not sure?<br>The young boys are playing outdoors and the man is smiling nearby<br>The kids are playing outdoors near a man with a smile<br>Answer: | True, Not sure, False |
| tweet_eval-hate | Question: Does the tweet convey the author's hatred towards something or someone? True or False?<br>Hundreds of Syrian refugees return home from Lebanon - ABC News<br>Answer: | True, False |

Table 3: Examples of Manual and Verbose templates. Texts in blue are manual templates.

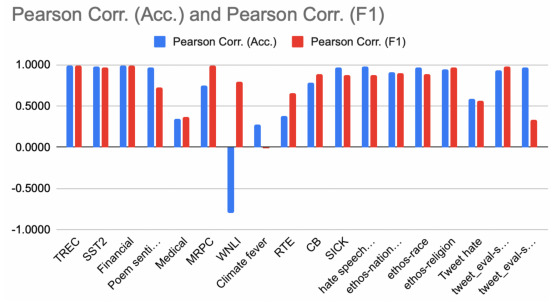



Figure 12: Pearson correlation analysis on all 18 tasks. A strong positive correlation is observed for all tasks and metrics, except for outliers.

| Dataset | Metric | no demo | 0% | 25% | 50% | 75% | 100% | random label | shuffled label |
|---|---|---|---|---|---|---|---|---|---|
| glue | Accuracy | 74.54 | 70.67±7.40 | 70.880±7.40 | 70.80±10.61 | 78.44±18.24 | 88.26±5.07 | 84.93±11.29 | 84.31±9.43 |
| sst2 | F1 | 73.94 | 67.34±10.13 | 66.96±14.15 | 74.50±13.19 | 74.78±23.46 | 88.06±5.33 | 84.09±13.03 | 83.59±10.59 |
| glue | Accuracy | 52.71 | 54.80±3.49 | 55.38±5.19 | 52.42±4.36 | 55.52±3.78 | 57.04±7.17 | 55.88±5.11 | 56.68±3.56 |
| rte | F1 | 34.52 | 47.52±7.80 | 48.50±9.19 | 45.33±5.71 | 51.80±5.51 | 48.81±14.17 | 47.97±10.65 | 51.43±6.01 |
| glue | Accuracy | 68.38 | 31.86±1.54 | 30.54±0.81 | 32.60±4.55 | 44.36±16.72 | 53.58±20.06 | 35.64±8.54 | 45.20±18.49 |
| mrpc | F1 | 40.61 | 26.79±4.62 | 23.86±0.13 | 28.24±6.82 | 34.75±11.75 | 35.89±10.53 | 29.79±9.66 | 35.15±11.24 |
| glue | Accuracy | 56.34 | 50.99±6.86 | 55.77±6.55 | 48.17±5.58 | 45.35±7.41 | 44.79±5.93 | 51.83±6.09 | 48.17±6.33 |
| wnli | F1 | 36.04 | 45.10±8.76 | 50.02±10.19 | 40.65±9.21 | 35.99±8.15 | 32.67±5.76 | 42.60±8.98 | 38.89±7.16 |
| super_glue | Accuracy | 8.93 | 21.07±14.58 | 41.43±6.36 | 46.79±7.72 | 54.29±7.32 | 60.36±11.39 | 30.00±9.48 | 49.29±6.51 |
| cb | F1 | 5.56 | 16.38±8.88 | 31.03±7.53 | 38.23±4.68 | 41.73±6.07 | 49.02±9.35 | 24.48±5.55 | 31.34±4.18 |
| trec | Accuracy | 21.20 | 34.84±8.34 | 41.20±12.44 | 43.92±8.89 | 56.84±6.21 | 67.44±6.04 | 42.60±10.79 | 36.16±5.59 |
| | F1 | 11.85 | 21.26±6.16 | 24.85±11.79 | 31.45±8.37 | 41.72±6.86 | 52.98±5.50 | 25.64±11.94 | 24.45±3.59 |
| financial | Accuracy | 21.85 | 25.03±2.31 | 37.62±8.53 | 42.38±14.95 | 79.51±5.70 | 80.22±8.58 | 33.60±8.01 | 57.57±9.43 |
| phrasebank | F1 | 17.50 | 26.78±8.21 | 42.88±8.40 | 47.72±11.33 | 78.31±4.48 | 75.72±6.01 | 41.36±8.81 | 43.56±10.26 |
| poem | Accuracy | 21.90 | 35.81±24.35 | 20.95±2.61 | 40.95±40.94 | 59.05±5.75 | 61.52±8.45 | 44.19±7.87 | 50.67±14.37 |
| sentiment | F1 | 22.62 | 19.25±6.99 | 18.11±5.99 | 27.05±14.71 | 35.84±6.20 | 35.62±9.43 | 31.689±8.10 | 30.87±2.74 |
| medical | Accuracy | 49.51 | 49.34±1.98 | 49.34±1.80 | 48.92±1.05 | 50.98±1.15 | 51.93±2.76 | 51.11±1.62 | 49.87±0.86 |
| questions_pairs | F1 | 33.11 | 38.60±5.63 | 38.64±7.90 | 42.25±8.02 | 43.69±6.19 | 49.06±3.13 | 41.49±9.38 | 40.34±7.87 |
| sick | Accuracy | 56.57 | 32.97±5.58 | 45.29±7.67 | 54.06±2.58 | 55.88±10.26 | 65.62±4.18 | 47.31±14.14 | 50.34±12.12 |
| | F1 | 24.96 | 26.14±6.16 | 33.10±3.44 | 38.63±6.95 | 42.19±13.93 | 49.80±11.15 | 33.53±13.01 | 33.95±12.21 |
| hate_speech18 | Accuracy | 89.49 | 13.20±3.70 | 35.13±18.19 | 38.30±32.00 | 77.94±23.46 | 85.01 ±9.90 | 71.28±21.44 | 89.49±0.02 |
| | F1 | 47.23 | 12.59±4.19 | 27.56±11.32 | 39.02±15.88 | 44.49±6.34 | 47.26±0.21 | 42.69±6.81 | 47.22±0.01 |
| ethos | Accuracy | 21.84 | 24.37±9.03 | 29.20±11.57 | 47.13±23.03 | 65.97±21.45 | 78.39±4.70 | 63.68±16.12 | 75.17±11.82 |
| national_origin | F1 | 22.99 | 22.43±8.12 | 27.04±9.59 | 36.04±13.34 | 46.07±7.46 | 49.28±5.04 | 45.55±3.77 | 52.17±11.04 |
| ethos | Accuracy | 26.44 | 23.68±3.51 | 27.59±7.13 | 48.28±17.66 | 68.74±7.77 | 78.16±0.00 | 61.38±14.35 | 78.39±0.51 |
| race | F1 | 242.76 | 20.61±5.15 | 25.33±8.50 | 42.13±15.77 | 45.29±4.75 | 43.87±0.00 | 45.90±1.85 | 44.93±2.36 |
| ethos | Accuracy | 21.84 | 22.76±3.19 | 24.60±5.30 | 37.24±14.93 | 58.39±22.49 | 78.62±2.89 | 31.38±15.41 | 69.43±18.44 |
| religion | F1 | 20.57 | 20.67±4.30 | 22.89±6.45 | 35.16±14.76 | 41.11±7.43 | 44.00±0.92 | 44.15±1.38 | 42.34±3.24 |
| tweet_eval | Accuracy | 42.70 | 43.08±2.59 | 45.48±4.78 | 47.40±5.81 | 49.52±4.42 | 58.00±3.86 | 52.36±5.41 | 52.46±7.38 |
| hate | F1 | 29.92 | 35.38±5.73 | 40.90±7.22 | 43.30±8.89 | 43.68±8.22 | 56.38±5.45 | 44.70±8.61 | 50.45±9.29 |
| tweet_eval | Accuracy | 53.85 | 18.46±1.72 | 20.38±3.49 | 22.31±3.99 | 21.54±4.59 | 26.15±10.84 | 18.85±2.51 | 22.31±4.63 |
| stance_atheism | F1 | 41.50 | 14.51±2.93 | 18.27±4.14 | 20.24±4.52 | 17.77±3.76 | 22.40±12.80 | 16.31±4.66 | 17.52±6.31 |
| tweet_eval | Accuracy | 49.25 | 28.06±4.99 | 31.64±5.11 | 29.25±6.38 | 30.78±9.12 | 38.51±5.42 | 29.96±4.30 | 35.22±4.79 |
| feminist | F1 | 34.97 | 20.79±5.90 | 24.75±5.94 | 25.25±4.81 | 24.78±10.70 | 24.95±5.83 | 21.03±5.43 | 20.70±4.96 |

Table 4: Full experiment results on GPT-NeoX.

| Dataset | Metric | no demo | 0% | 25% | 50% | 75% | 100% | random label | shuffled label |
|---|---|---|---|---|---|---|---|---|---|
| glue | Accuracy | 75.46 | 49.40±0.50 | 61.67±11.16 | 59.43±7.49 | 75.83±15.67 | 90.25±3.86 | 53.58±4.60 | 64.04±18.00 |
| sst2 | F1 | 75.31 | 33.73±1.04 | 54.18±16.94 | 51.56±11.75 | 72.35±22.11 | 90.20±3.93 | 41.69±8.58 | 55.68±25.16 |
| glue | Accuracy | 52.71 | 44.55±5.04 | 47.15±3.92 | 48.95±4.12 | 52.71±3.88 | 53.72±5.05 | 51.05±5.71 | 53.57±3.10 |
| rte | F1 | 34/52 | 38/79±4.09 | 42.52±6.72 | 38/34±6.47 | 48.80±7.75 | 48.56±8.88 | 43.63±5.27 | 48.18±5.01 |
| glue | Accuracy | 68.38 | 32.25±2.40 | 35.98±11.57 | 43.77±14.81 | 56.76±14.97 | 59.71±12.34 | 43.53±16.03 | 62.06±6.25 |
| mrpc | F1 | 40.61 | 27.51±5.58 | 29.10±7.72 | 35.84±9.65 | 44.44±7.97 | 43.60±2.99 | 36.11±15.93 | 43.32±3.11 |
| glue | Accuracy | 56.34 | 48.45±5.42 | 47.61±3.51 | 44.23±5.14 | 46.20±5.40 | 46.76±4.92 | 46.48±6.06 | 49.58±3.21 |
| wnli | F1 | 36.02 | 43.18±7.24 | 44.84±5.78 | 38.22±5.67 | 37.24±7.06 | 41.39±7.75 | 38.21±7.87 | 43.11±6.10 |
| super_glue | Accuracy | 17.86 | 13.21±5.73 | 21.07±4.62 | 40.71±10.37 | 43.21±12.53 | 52.86±12.40 | 20.71±13.87 | 50.71±8.24 |
| cb | F1 | 15.21 | 10.07±4.34 | 19.22±4.98 | 27.78±7.81 | 27.67±7.59 | 33.86±11.77 | 16.36±10.01 | 27.87±8.82 |
| trec | Accuracy | 21.60 | 17.92±7.60 | 30.20±16.30 | 39.00±15.04 | 46.88±13.35 | 49.24±11.47 | 28.08±5.32 | 30.44±12.99 |
| | F1 | 15.25 | 10.35±3.65 | 21.42±12.56 | 26.89±13.70 | 34.39±10.48 | 36.45±8.04 | 18.02±4.39 | 19.99±9.55 |
| financial | Accuracy | 29.58 | 18.28±4.51 | 23.31±4.51 | 23.66±8.88 | 56.16±12.31 | 70.95±5.84 | 20.22±4.78 | 44.81±18.02 |
| phrasebank | F1 | 34.92 | 17.32±8.45 | 17.32±8.45 | 20.07±9.39 | 41.98±9.71 | 55.11±12.43 | 19.06±8.04 | 27.07±4.36 |
| poem | Accuracy | 19.05 | 28.57±20.96 | 26.67±19.75 | 42.48±21.26 | 48.95±19.76 | 50.86±16.30 | 34.86±16.74 | 47.24±21.24 |
| sentiment | F1 | 19.23 | 17.49±7.04 | 19.68±8.33 | 27.39±12.40 | 26.40±7.73 | 30.48±7.85 | 23.50±11.50 | 30.25±8.53 |
| medical | Accuracy | 49.51 | 44.92±4.44 | 47.18±4.62 | 50.33±2.90 | 50.03±1.54 | 50.92±2.20 | 50.36±0.99 | 51.11±1.33 |
| questions_pairs | F1 | 33.11 | 36.08±3.17 | 39.66±6.01 | 40.11±8.56 | 38.63±5.83 | 42.22±8.49 | 37.51±4.04 | 38.17±6.85 |
| sick | Accuracy | 30.51 | 43.80±18.43 | 50.63±7.97 | 49.41±9.70 | 49.45±11.58 | 57.90±14.12 | 47.96±13.04 | 42.79±12.61 |
| | F1 | 24.42 | 22.39±6.24 | 26.71±2.99 | 27.90±4.82 | 34.82±12.47 | 46.76±19.92 | 26.31±6.83 | 29.44±6.83 |
| hate_speech18 | Accuracy | 89.49 | 13.96±6.71 | 27.09±25.05 | 46.66±31.79 | 73.83±15.81 | 80.48±17.85 | 63.99±18.21 | 87.69±1.78 |
| | F1 | 47.23 | 12.75±6.00 | 20.69±14.78 | 32.54±15.02 | 45.91±4.15 | 47.86±4.91 | 43.98±7.61 | 47.12±0.36 |
| ethos | Accuracy | 25.29 | 28.05±22.34 | 35.63±26.32 | 51.03±19.26 | 56.09±20.46 | 68.97±18.90 | 54.25±25.87 | 69.89±20.20 |
| national_origin | F1 | 25.25 | 23.49±16.69 | 28.11±14.46 | 40.95±14.75 | 43.90±10.49 | 45.34±3.12 | 41.16±14.06 | 48.08±9.99 |
| ethos | Accuracy | 32.18 | 22.07±0.51 | 43.68±26.06 | 48.05±19.22 | 65.75±9.56 | 78.16±0.00 | 55.17±21.12 | 78.16±0.00 |
| race | F1 | 31.86 | 18.25±0.72 | 34.13±17.43 | 41.94±14.12 | 49.77±5.11 | 43.87±0.00 | 43.53±11.74 | 43.87±0.00 |
| ethos | Accuracy | 29.89 | 19.54±1.63 | 28.97±10.76 | 30.57±13.74 | 69.43±13.35 | 80.00±2.52 | 51.05±24.01 | 77.01±3.90 |
| religion | F1 | 29.74 | 17.18±1.30 | 26.89±11.02 | 28.79±13.64 | 46.93±3.16 | 46.52±5.17 | 37.60±11.80 | 47.83±3.62 |
| tweet_eval | Accuracy | 42.70 | 44.02±6.56 | 47.76±5.18 | 53.08±5.71 | 55.76±4.35 | 59.72±2.77 | 54.74±2.32 | 54.42±4.48 |
| hate | F1 | 29.92 | 40.54±4.18 | 43.40±3.39 | 42.43±4.22 | 48.60±8.50 | 49.67±9.04 | 42.19±7.08 | 46.74±6.60 |
| tweet_eval | Accuracy | 25.00 | 20.00±2.58 | 21.92±1.05 | 22.31±2.19 | 28.55±10.88 | 45.38±17.50 | 20.00±2.58 | 43.85±16.06 |
| stance_atheism | F1 | 17.82 | 15.57±5.70 | 13.37±2.60 | 14.86±2.85 | 20.43±11.06 | 29.66±11.31 | 13.41±2.69 | 25.86±6.71 |
| tweet_eval | Accuracy | 49.51 | 44.92±4.44 | 47.18±4.62 | 50.33±2.90 | 50.03±1.54 | 50.92±2.20 | 50.36±0.99 | 51.11±1.33 |
| feminist | F1 | 33.11 | 36.08±3.17 | 39.66±6.01 | 40.11±8.56 | 38.63±5.83 | 42.22±8.49 | 37.51±4.04 | 38.17±6.85 |

Table 5: Full experiment results on GPT-J.