Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?

Sewon Min1,2 Xinxi Lyu1 Ari Holtzman1 Mikael Artetxe2 Mike Lewis2 Hannaneh Hajishirzi1,3 Luke Zettlemoyer1,2
1University of Washington 2Meta AI 3Allen Institute for AI
{sewon,alrope,ahai,hannaneh,lsz}@cs.washington.edu {artetxe,mikelewis}@meta.com

Abstract

Large language models (LMs) are able to perform a new task via inference alone by conditioning on a few input-label pairs (demonstrations) and making predictions for new inputs. However, there has been little understanding of how the model learns and which aspects of the demonstrations contribute to end task performance. In this paper, we show that ground truth demonstrations are in fact not required—randomly replacing labels in the demonstrations barely hurts performance on a range of classification and multi-choice tasks, consistently over 12 different models including GPT-3. Instead, we find that other aspects of the demonstrations are the key drivers of end task performance, including the fact that they provide a few examples of (1) the label space, (2) the distribution of the input text, and (3) the overall format of the sequence. Together, our analysis provides a new way of understanding how and why in-context learning works, while opening up new questions about how much can be learned from large language models through inference alone.

1 Introduction

Large language models (LMs) have shown impressive performance on downstream tasks by simply conditioning on a few input-label pairs (demonstrations); this type of inference has been referred to as in-context learning (Brown et al., 2020). Despite in-context learning consistently outperforming zero-shot inference on a wide range of tasks (Zhao et al., 2021), there is little understanding of how it works and which aspects of the demonstrations contribute to end task performance.

In this paper, we show that ground truth demonstrations are in fact not required for effective in-context learning (Section 4). Specifically, replacing the labels in demonstrations with random labels barely hurts performance in a range of classification and multi-choice tasks (Figure 1). The result is consistent over 12 different models including the GPT-3 family (Radford et al., 2019; Min et al., 2021b; Wang and Komatsuzaki, 2021; Artetxe et al., 2021; Brown et al., 2020). This strongly suggests, counter-intuitively, that the model does not rely on the input-label mapping in the demonstrations to perform the task.

Further analysis investigates which parts of demonstrations actually do contribute to the performance. We identify possible aspects of demonstrations (e.g., the label space and the distribution of the input text) and evaluate a series of variants of the demonstrations to quantify the impact of each (Section 5). We find that: (1) the label space and the distribution of the input text specified by the demonstrations are both key to in-context learning (regardless of whether the labels are correct for individual inputs); (2) specifying the overall format is also crucial, e.g., when the label space is unknown, using random English words as labels is significantly better than using no labels; and

Figure 1: Results in classification (top) and multi-choice tasks (bottom), using three LMs with varying size. Reported on six datasets on which GPT-3 is evaluated; the channel method is used. See Section 4 for the full results. In-context learning performance drops only marginally when labels in the demonstrations are replaced by random labels.
Circulation revenue has increased by 5% in Finland. Panostaja did not disclose the purchase price. Paying off the national debt will be extremely painful. The acquisition will have an immediate positive impact.

Demonstrations

<table>
<thead>
<tr>
<th>Test Input</th>
<th>Prediction</th>
<th>LM</th>
<th>Meta-trained</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulation revenue has increased by 5%</td>
<td>Positive</td>
<td>✅</td>
<td>✗</td>
</tr>
<tr>
<td>Panostaja did not disclose the purchase</td>
<td>Neutral</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>Paying off the national debt will be</td>
<td>Negative</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>extremely painful.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>The acquisition will have an immediate</td>
<td>Positive</td>
<td>✅</td>
<td>✖</td>
</tr>
</tbody>
</table>

Table 1: A list of LMs used in the experiments: GPT-2 (Radford et al., 2019), MetaICL (Min et al., 2021b), GPT-J (Wang and Komatsuzaki, 2021), fairseq LMs (Artetxe et al., 2021) and GPT-3 (Brown et al., 2020). ‘Public’ indicates whether the model weights are public; ‘Meta-trained’ indicates whether the model is meta-trained with an in-context learning objective. We use dense models in Artetxe et al. (2021) and refer them as fairseq LMs for convenience. We use the Davinci API (the base version, not the instruct version) and assume it to be 175B, following Gao et al. (2021) and Artetxe et al. (2021).

(3) meta-training with an in-context learning objective (Min et al., 2021b) magnifies these effects—the models almost exclusively exploit simpler aspects of the demonstrations like the format rather than the input-label mapping.

In summary, our analysis provides a new way of understanding the role of the demonstrations in in-context learning. We empirically show that the model (1) counter-intuitively does not rely on the ground truth input-label mapping provided in the demonstrations as much as we thought (Section 4), and (2) nonetheless still benefits from knowing the label space and the distribution of inputs specified by the demonstrations (Section 5). We also include a discussion of broader implications, e.g., what we can say about the model learning at test time, and avenues for future work (Section 6).

2 Related Work

Large language models have been key to strong performance in a wide range of downstream tasks (Devlin et al., 2019; Radford et al., 2019; Liu et al., 2019; Raffel et al., 2020; Lewis et al., 2020). While finetuning has been a popular approach to transfer to new tasks (Devlin et al., 2019), it is often impractical to finetune a very large model (e.g. ≥10B parameters). Brown et al. (2020) propose in-context learning as an alternative way to learn a new task. As depicted in Figure 2, the LM learns a new task via inference alone by conditioning on a concatenation of the training data as demonstrations, without any gradient updates.

In-context learning has been the focus of significant study since its introduction. Prior work proposes better ways of formulating the problem (Zhao et al., 2021; Holtzman et al., 2021; Min et al., 2021a), better ways of choosing labeled examples for the demonstrations (Liu et al., 2021; Lu et al., 2021; Rubin et al., 2021), meta-training with an explicit in-context learning objective (Chen et al., 2021; Min et al., 2021b), and learning to follow instructions as a variant of in-context learning (Mishra et al., 2021b; Efrat and Levy, 2020; Wei et al., 2022a; Sanh et al., 2022). At the same time, some work reports brittleness and oversensitivity for in-context learning (Lu et al., 2021; Zhao et al., 2021; Mishra et al., 2021a).

Relatively less work has been done to understand why in-context learning works. Xie et al. (2022) provide theoretical analysis that in-context learning can be formalized as Bayesian inference that uses the demonstrations to recover latent concepts. Razeghi et al. (2022) show that in-context learning performance is highly correlated with term frequencies in the pretraining data. To the best of our knowledge, this paper is the first that provides an empirical analysis that investigates why in-context learning achieves performance gains over zero-shot inference. We find that the ground truth input-label mapping in the demonstrations has only a marginal effect, and measure the impact of finer-grained aspects of the demonstrations.

3 Experimental Setup

We describe the experimental setup used in our analysis (Section 4 and 5).

Models. We experiment with 12 models in total. We include 6 language models (Table 1), all of which are decoder-only, dense LMs. We use each LM with two inference methods, direct and channel, following Min et al. (2021a). The sizes of LMs vary from 774M to 175B. We include the

Figure 2: An overview of in-context learning. The demonstrations consist of \( k \) input-label pairs from the training data (\( k = 3 \) in the figure).
largest dense LM (GPT-3) and the largest publicly released dense LM (fairseq 13B) at the time of conducting experiments. We also include MetaICL, which is initialized from GPT-2 Large and then meta-trained on a collection of supervised datasets with an in-context learning objective, and ensure that our evaluation datasets do not overlap with those used at meta-training time.

**Evaluation Data.** We evaluate on 26 datasets, including sentiment analysis, paraphrase detection, natural language inference, hate speech detection, question answering, and sentence completion (full list and references provided in Appendix A).\(^1\) All datasets are classification and multi-choice tasks.

We use these datasets because they (1) are true low-resource datasets with less than 10K training examples, (2) include well-studied benchmarks from GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019a), and (3) cover diverse domains including science, social media, finance, and more.

**Other Details.** We use \(k = 16\) examples as demonstrations by default for all experiments in the paper, unless otherwise specified. Examples are sampled at uniform from the training data. We choose a set of \(k\) training examples using 5 different random seeds and run experiments 5 times. For fairseq 13B and GPT-3, due to limited resources, we experiment with a subset of 6 datasets\(^2\) and 3 random seeds. We report Macro-F1\(^3\) for classification tasks and Accuracy for multi-choice tasks. We compute per-dataset average over seeds, and then report macro-average over datasets. We use the minimal templates in forming an input sequence from an example. We refer to Appendix B for more details. All experiments are reproducible from github.com/Alrope123/rethinking-demonstrations.

\(4\) **Ground Truth Matters Little**

\(4.1\) **Gold labels vs. random labels**

To see the impact of correctly-paired inputs and labels in the demonstrations—which we call the ground truth input-label mapping—we compare the following three methods.\(^4\)

**No demonstrations** is a typical zero-shot method that does not use any labeled data. A prediction is made via \(\text{argmax}_{y \in C} P(y|x)\), where \(x\) is the test input and \(C\) is a small discrete set of possible labels.

**Demonstrations w/ gold labels** are used in a typical in-context learning method with \(k\) labeled examples \((x_1, y_1)\ldots(x_k, y_k)\). A concatenation of \(k\) input-label pairs is used to make a prediction via \(\text{argmax}_{y \in C} P(y|x, y_1\ldots y_k, x)\).

\(^2\)Three classification and three multi-choice: MRPC, RTE, Tweet_eval-hate, OpenbookQA, CommonsenseQA, COPA.

\(^3\)Known to be better for imbalanced classes.

\(^4\)Without loss of generality, all methods in Section 4 and 5 are described based on the direct method, but can be trivially converted to the channel method by flipping \(x\) and \(y\).
Demonstrations w/ random labels are formed with random labels, instead of gold labels from the labeled data. Each $x_i$ (1 ≤ i ≤ k) is paired with $\tilde{y}_i$, which is randomly sampled at uniform from C. A concatenation of $(x_1, \tilde{y}_1) \ldots (x_k, \tilde{y}_k)$ is then used to make a prediction via $\arg\max_{y\in C} P(y|x_1, \tilde{y}_1 \ldots x_k, \tilde{y}_k, x)$.

Results are reported in Figure 3. First, using the demonstrations with gold labels significantly improves the performance over no demonstrations, as it has been consistently found in much of prior work (Brown et al., 2020; Zhao et al., 2021; Liu et al., 2021). We then find that replacing gold labels with random labels only marginally hurts performance. The trend is consistent over nearly all models: models see performance drop in the range of 0–5% absolute. There is less impact in replacing labels in multi-choice tasks (1.7% on average) than in classification tasks (2.6% absolute).

This result indicates that the ground truth input-label pairs are not necessary to achieve performance gains. This is counter-intuitive, given that correctly paired training data is critical in typical supervised training—it informs the model of the expected input-label correspondence required to perform the downstream task. Nonetheless, the models do achieve non-trivial performance on the downstream tasks. This strongly suggests that the models are capable of recovering the expected input-label correspondence for the task; however, it is not directly from the pairings in the demonstrations.

It is also worth noting that there is particularly little performance drop in MetaICL: 0.1–0.9% absolute. This suggests that meta-training with an explicit in-context learning objective actually encourages the model to essentially ignore the input-label mapping and exploit other components of the demonstrations (more discussion in Section 5.4).

In Appendix C.2, we provide additional results showing that (1) selecting random labels from a true distribution of labels (instead of a uniform distribution) reduces the gap even further, and (2) the trends may depend on the dataset, although the overall trend is consistent over most datasets.

4.2 Ablations

For additional ablations, we experiment with 5 classification and 4 multi-choice datasets. Does the number of correct labels matter? To further examine the impact of correctness of labels in the demonstrations, we conduct an ablation study by varying the number of correct labels in the demonstrations. We evaluate “Demonstrations w/ α% correct labels” (0 ≤ α ≤ 100) which consist of $k \times a/100$ correct pairs and $k \times (1 - a/100)$ incorrect pairs (see Algorithm 1 in Appendix B). Here, α = 100 is the same as typical in-context learning, i.e., demonstrations w/ gold labels.

Results are reported in Figure 4. Model performance is fairly insensitive to the number of correct labels in the demonstrations. In fact, always using incorrect labels significantly outperforms no-
Circulation revenue has increased by 5% in Finland. Positive

The acquisition will have an immediate positive impact. ?

Figure 6: Results with minimal templates and manual templates. ‘+T’ indicates that manual templates are used. Channel and Direct used for classification and multi-choice, respectively.

Demonstrations Distribution of inputs Label space

<table>
<thead>
<tr>
<th>Demonstrations</th>
<th>Distribution of inputs</th>
<th>Label space</th>
</tr>
</thead>
<tbody>
<tr>
<td>Circulation revenue has increased by 5% in Finland</td>
<td>in</td>
<td>Positive</td>
</tr>
<tr>
<td>Panostaja did not disclose the purchase price</td>
<td>in</td>
<td>Neutral</td>
</tr>
<tr>
<td>Paying off the national debt will be extremely painful</td>
<td>in</td>
<td>Negative</td>
</tr>
</tbody>
</table>

Figure 7: Four different aspects in the demonstrations: the input-label mapping, the distribution of the input text, the label space, and the use of input-label pairing as the format of the demonstrations.

Is the result consistent with varying k? We study the impact of the number of input-label pairs (k) in the demonstrations. Results are reported in Figure 5. First, using the demonstrations significantly outperforms the no demonstrations method even with small k (k = 4), and performance drop from using gold labels to using random labels is consistently small across varying k, in the range of 0.8–1.6%.

Interestingly, model performance does not increase much as k increases when k ≥ 8, both with gold labels and with random labels. This is in contrast with typical supervised training where model performance rapidly increases as k increases, especially when k is small. We hypothesize that larger labeled data is beneficial mainly for supervising the input-label correspondence, and other components of the data like the example inputs, example labels and the data format are easier to recover from the small data, which is potentially a reason for minimal performance gains from larger k (more discussion in Section 5).

Is the result consistent with better templates? While we use minimal templates by default, we also explore manual templates, i.e., templates that are manually written in a dataset-specific manner, taken from prior work (details in Appendix B). Figure 6 shows that the trend—replacing gold labels with random labels barely hurting performance—holds with manual templates. It is worth noting that using manual templates does not always outperform using minimal templates.

5 Why does In-Context Learning work?

Section 4 shows that the ground truth input-label mapping in the demonstrations has little impact to performance gains from in-context learning. This section further examines what other aspects of the demonstrations lead to good performance of in-context learning.

We identify four aspects of the demonstrations \((x_1, y_1) \ldots (x_k, y_k)\) that potentially provide learning signal (depicted in Figure 7).

1. The input-label mapping, i.e., whether each input \(x_i\) is paired with a correct label \(y_i\).
2. The distribution of the input text, i.e., the underlying distribution that \(x_1 \ldots x_k\) are from.
3. The label space, i.e., the space covered by \(y_1 \ldots y_k\).
4. The format—specifically, the use of input-label pairing as the format.

As Section 4 does for the input-label mapping, we design a series of variants of the demonstrations that quantify the impact of each aspect in isolation (Section 5.1–5.3). We then additionally discuss the trend of the models meta-trained with an in-context learning objective (Section 5.4). For all experiments, models are evaluated on five classification...
5.1 Impact of the distribution of the input text

We experiment with OOD demonstrations which include out-of-distribution (OOD) text instead of the inputs from unlabeled training data. Specifically, a set of $k$ sentences $\{x_{i,\text{rand}}\}_{i=1}^k$ are randomly sampled from an external corpus, and replace $x_1...x_k$ in the demonstrations. This variant assesses the impact of the distribution of the input text, while keeping the label space and the format of the demonstrations.

**Results.** Figure 8 shows that using out-of-distribution inputs instead of the inputs from the training data significantly drops the performance when Channel MetaICL, Direct GPT-J or Channel GPT-J are used, both in classification and multi-choice, by 3–16% in absolute. In the case of Direct GPT-J in multi-choice, it is even significantly worse than no demonstrations. Direct MetaICL is an exception, which we think is the effect of meta-training (discussion in Section 5.4).

This suggests that in-distribution inputs in the demonstrations substantially contribute to performance gains. This is likely because conditioning on the in-distribution text makes the task closer to language modeling, since the LM always conditioned on the in-distribution text during training.

5.2 Impact of the label space

We also experiment with demonstrations w/ random English words that use random English words as labels for all $k$ pairs. Specifically, we sample a random subset of English words $C_{\text{rand}}$ where $|C_{\text{rand}}| = |C|$, and randomly pair $\tilde{y}_i \in C_{\text{rand}}$ with $x_i$. This variant assesses the impact of the label space, while keeping the distribution of the input text and the format of the demonstrations.

**Results.** Based on Figure 9, direct models and channel models exhibit different patterns. With direct models, the performance gap between using random labels within the label space and using random English words is significant, ranging between 5–16% absolute. This indicates that conditioning on the label space significantly contributes to performance gains. This is true even for multi-choice tasks where there is no fixed set of labels—we hypothesize that multi-choice tasks still do have a particular distribution of the choices (e.g., objects like “Bolts” or “Screws” in the OpenBookQA dataset) that the model uses.

On the other hand, removing the output space does not lead to significant drop in the channel models: there is 0–2% drop in absolute, or sometimes even an increase. We hypothesize that this is because the channel models only condition on the labels, and thus are not benefiting from knowing the label space. This is in contrast to direct models which must generate the correct labels.

5.3 Impact of input-label pairing

Section 5.1 and 5.2 focus on variants which keep the format of the demonstrations as much as possible. This section explores variants that change the format. While there are many aspects of the format, we make minimal modifications to remove the pairings of inputs to labels. Specifically, we evaluate demonstrations with no labels where the LM is conditioned on the concatenation of $x_1...x_k$, and
**Demonstrations with labels only** where the LM is conditioned on the concatenation of $y_1...y_k$. These ablations provide the no-format counterparts of the ‘demonstrations with random English words’ and ‘demonstrations with OOD inputs’, respectively.

**Results.** Based on Figure 10, removing the format is close to or worse than no demonstrations, indicating the importance of the format. This is likely because conditioning on a sequence of input-label pairs triggers the model to mimic the overall format and complete the new example as expected when the test input is given.

More interestingly, keeping the format plays a significant role in retaining a large portion of performance gains by only using the inputs or only using the labels. For instance, with Direct MetaICL, it is possible to retain 95% and 82% of improvements from in-context learning (demonstrations with gold labels) by simply sampling random sentences from a corpus and randomly pairing them with the label set ($\square$ in Figure 10) in classification and multi-choice, respectively. Similarly, with the channel models, it is possible to retain 82%, 87%, 86% and 75% of improvements from in-context learning by simply pairing each input from the unlabeled training data with a random English word ($\square$ in Figure 10) in MetaICL classification, GPT-J classification, MetaICL multi-choice and GPT-J multi-choice, respectively. For all of these cases, removing inputs instead of using OOD inputs, or removing labels instead of using random English words is significantly worse, indicating that keeping the format of the input-label pairs is key.

**5.4 Impact of meta-training**

Different from other models, MetaICL is trained with an in-context learning objective, in line with recent work that uses multi-task training on a
large collection of supervised datasets (called meta-training) for generalization to new tasks (Aghajanyan et al., 2021; Khashabi et al., 2020; Wei et al., 2022a; Sanh et al., 2022). We aim to better understand the role of this meta-training in relation with our findings by closely examining the result of MetaICL. In particular, we observe that the patterns we see so far are significantly more evident with MetaICL than with other models. For instance, the ground truth input-label mapping matters even less, and keeping the format of the demonstrations matters even more. There is nearly zero influence of the input-label mapping and the input distribution in Direct MetaICL, and the input-label mapping and the output space in Channel MetaICL.

Based on this observation, we hypothesize that meta-training encourages the model to exclusively exploit simpler aspects of the demonstrations and to ignore others. This is based on our intuition that (1) the input-label mapping is likely harder to exploit, (2) the format is likely easier to exploit, and (3) the space of the text that the model is trained to generate is likely easier to exploit than the space of the text that the model conditions on.\footnote{That is, the direct model exploits the label space better than the input distribution, and the channel model exploits the input distribution better than the label space.}

6 Discussion & Conclusion

In this paper, we study the role of the demonstrations with respect to the success of in-context learning. We find that the ground truth input-label mapping in the demonstrations matters significantly less than one might think—replacing gold labels with random labels in the demonstrations only marginally lowers the performance. We then identify a series of aspects in the demonstrations and examine which aspect actually contributes to performance gains. Results reveal that (1) gains are mainly coming from independent specification of the input space and the label space, (2) the models can still retain up to 95\% of performance gains by using either the inputs only or the label set only if the right format is used, and (3) meta-training with random labels in the demonstrations only has the same role to LMs, and hypothesize that our findings hold for instruction-following models: the demonstrations and instructions largely have the same role to LMs, and hypothesize that our findings hold for instruction-following models: the demonstrations are for task location and the intrinsic ability to perform the task is obtained at pretraining time.\footnote{However, while Reynolds and McDonell (2021) claims that the demonstrations are thus unnecessary, we think using the demonstrations is actually the most unambiguous and the easiest way to prompt the model to perform a task.}

On one hand, this suggests that the language modeling objective has led to great zero-shot capacity, even if it is not always evident from the naive zero-shot accuracy. On the other hand, this suggests that in-context learning may not work on a task whose input-label correspondence is not already captured in the LM. This leads to the research question of how to make progress in NLP problems that in-context learning does not solve: whether we need a better way of extracting the input-label mappings that are already stored in the LM, a better variant of the LM objective that learns a wider range of task semantics, or explicit supervision through fine-tuning on the labeled data.

Connection to instruction-following models.

Prior work has found it promising to train the model that reads the natural language description of the task (called instructions) and performs a new task at inference (Mishra et al., 2021b; Efrat and Levy, 2020; Wei et al., 2022a; Sanh et al., 2022). We think the demonstrations and instructions largely have the same role to LMs, and hypothesize that our findings hold for instruction-following models: the
instructions prompt the model to recover the capacity it already has, but do not supervise the model to learn novel task semantics. This has been partially verified by Webson and Pavlick (2022) who showed that the model performance does not degrade much with irrelevant or misleading instructions. We leave more analysis on instruction-following models for future work.

**Significantly improved zero-shot performance.**

One of our key findings is that it is possible to achieve nearly $k$-shot performance without using any labeled data, by simply pairing each unlabeled input with a random label and using it as the demonstrations. This means our zero-shot baseline level is significantly higher than previously thought. Future work can further improve the zero-shot performance with relaxed assumptions in access to the unlabeled training data.

**Limitation**

**Effect of types of tasks and datasets.** This paper focuses on the tasks from established NLP benchmarks that have real natural language inputs. Synthetic tasks with more limited inputs may actually use the ground truth labels more, as observed by Rong (2021).

We report macro-level analysis by examining the average performance over multiple NLP datasets, but different datasets may behave differently. Appendix C.2 discusses this aspect, including findings that there are larger gaps between using the ground truth labels and using the random labels in some dataset-model pairs (e.g., in the most extreme case, nearly 14% absolute on the financial_phrasebank dataset with GPT-J). Since the first version of our paper, Kim et al. (2022) showed that using negated labels substantially lowers the performance in classification.$^{11}$ We believe it is important to understand to what extend the model needs the ground truth labels to successfully perform in-context learning.

**Extensions to generation.** Our experiments are limited to classification and multi-choice tasks. We hypothesize that ground truth output may not be necessary for in-context learning in the open-set tasks such as generation, but leave this to future work. Extending of our experiments to such tasks is not trivial, because it requires a variation of the output which has incorrect input-output correspondence while keeping the correct output distribution (which is important based on our analysis in Section 5).

Since the first version of our paper, Madaan and Yazdanbakhsh (2022) conducted a similar analysis with the chain of thought prompting (Wei et al., 2022b) which generates a rationale to perform complex tasks such as math problems. Madaan and Yazdanbakhsh (2022) show that, while simply using a random rationale in the demonstrations (e.g., pairing with a rationale from a different example) significantly degrades the performance, other types of counterfactual rationales (e.g., wrong equations) do not degrade the performance as much as we thought. We refer to Madaan and Yazdanbakhsh (2022) for more discussions on what aspects of the rationale matter or do not matter.

**Acknowledgements**

We thank Gabriel Ilharco, Julian Michael, Ofir Press, UW NLP members and anonymous reviewers for their comments in the paper. This research was supported by NSF IIS-2044660, ONR N00014-18-1-2826, a Sloan fellowship and gifts from AI2.

**References**


Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. ArXiv.


Daniel Khashabi, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, and Hannaneh Hajishirzi. 2020. UnifiedQA: Crossing format boundaries with a single qa system. In Findings of EMNLP.


Quentin Lhoest, Albert Villanova del Moral, Yacine Jermite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid,


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 LREC.


Timo Schick and Hinrich Schütze. 2021. It’s not just size that matters: Small language models are also few-shot learners. In NAACL-HLT.


Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. TACL.

Oyvind Tafjord, Peter Clark, Matt Gardner, Wen-tau Yih, and Ashish Sabharwal. 2019a. Quarel: A dataset and models for answering questions about qualitative relationships. In AAAI.

Oyvind Tafjord, Matt Gardner, Kevin Lin, and Peter Clark. 2019b. QuaRTz: An open-domain dataset of qualitative relationship questions. In EMNLP.


Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In NAACL-HLT.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2022a. Finetuned language models are zero-shot learners. In ICLR.

A Full Datasets

We include 26 datasets as follows: financial_phrasebank (Malo et al., 2014), poem_sentiment (Sheng and Uthus, 2020), medical_questions_pairs (McCreery et al., 2020), glue-mrpc (Dolan and Brockett, 2005), glue-wnli (Levesque et al., 2012), climate_fever (Diggelmann et al., 2020), glue rte (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), superglue-cb (de Marneffe et al., 2019), sick (Marelli et al., 2014), hate_speech18 (de Gibert et al., 2018), ethos-national_origin (Mollas et al., 2020), ethos-race (Mollas et al., 2020), ethos-religion (Mollas et al., 2020), tweet_eval-hate (Barbieri et al., 2020), tweet_eval-stance_atheism (Barbieri et al., 2020), tweet_eval-stance_feminist (Barbieri et al., 2020), quarel (Tafjord et al., 2019a), openbookqa (Mihaylov et al., 2018), qasc (Khot et al., 2020), commonsense_qa (Talmor et al., 2019), ai2_arc (Clark et al., 2018), codah (Chen et al., 2019), superglue-copa (Gordon et al., 2012), dream (Sun et al., 2019), quartz-with_knowledge (Tafjord et al., 2019b), quartz-no_knowledge (Tafjord et al., 2019b). The choice of datasets is made following low-resource datasets in Min et al. (2021b), with the exact same set of k-shot train data using 5 random seeds. We use the HuggingFace version of the data (Lhoest et al., 2021) and use the development data for evaluation, following Ye et al. (2021). See Table 2 for statistics.

B Experimental Details

Example template We follow Ye et al. (2021); Min et al. (2021b); Logan IV et al. (2021) in using the minimal format to transform the input to a sequence (e.g. a concatenation of multiple inputs) and using the label words from each dataset as it is. We also explore manual templates taken from prior work (Holtzman et al., 2021; Zhao et al., 2021) as reported in Section 4.2, although we find that using these templates is not consistently better than using minimal templates. We thus run main experiments with minimal templates. Example templates are provided in Table 3.

Format of the demonstrations We follow the standard of each model for formatting the demonstrations, either from exploration in prior work or the example code provided in the official tutorial. For GPT-2, we separate the input and the label, and each demonstration example with a space. For MetaICL, GPT-J and GPT-3, we separate the input and the label with a newline (\n), and each demonstration example with three newlines. For fairseq models, we use a newline to separate the input and the label as well as each demonstration example.

Details in variants of the demonstrations For “demonstrations w/ a% accurate labels” (0 ≤ a ≤ 100), we use k × a/100 correct pairs and k × (1 − a/100) incorrect pairs in a random order, as described in Algorithm 1. For “OOD demonstrations”, we use CC-News (Nagel, 2016) as an external corpus. We consider the length of the text during sampling, so that sampled sentences have similar length to the test input. For “demonstrations with random English words”, we use pypi.org/project/english-words for the set of En-

<table>
<thead>
<tr>
<th>Dataset category: Sentiment analysis</th>
<th># Train</th>
<th># Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>financial_phrasebank</td>
<td>1,811</td>
<td>453</td>
</tr>
<tr>
<td>poem_sentiment</td>
<td>892</td>
<td>105</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset category: Paraphrase detection</th>
<th># Train</th>
<th># Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>medical_questions_pairs</td>
<td>2,438</td>
<td>610</td>
</tr>
<tr>
<td>glue-mrpc</td>
<td>3,668</td>
<td>408</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset category: Natural language inference</th>
<th># Train</th>
<th># Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>glue-wnli</td>
<td>635</td>
<td>71</td>
</tr>
<tr>
<td>climate_fever</td>
<td>1,228</td>
<td>307</td>
</tr>
<tr>
<td>glue rte</td>
<td>2,490</td>
<td>277</td>
</tr>
<tr>
<td>superglue-cb</td>
<td>250</td>
<td>56</td>
</tr>
<tr>
<td>sick</td>
<td>4,439</td>
<td>495</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset category: Hate speech detection</th>
<th># Train</th>
<th># Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>hate_speech18</td>
<td>8,562</td>
<td>2,141</td>
</tr>
<tr>
<td>ethos-national_origin</td>
<td>346</td>
<td>87</td>
</tr>
<tr>
<td>ethos-race</td>
<td>346</td>
<td>87</td>
</tr>
<tr>
<td>ethos-religion</td>
<td>346</td>
<td>87</td>
</tr>
<tr>
<td>tweet_eval-hate</td>
<td>8,993</td>
<td>999</td>
</tr>
<tr>
<td>tweet_eval-stance_atheism</td>
<td>461</td>
<td>52</td>
</tr>
<tr>
<td>tweet_eval-stance_feminist</td>
<td>597</td>
<td>67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset category: Question answering</th>
<th># Train</th>
<th># Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>quarel</td>
<td>1,941</td>
<td>278</td>
</tr>
<tr>
<td>openbookqa</td>
<td>4,957</td>
<td>500</td>
</tr>
<tr>
<td>qasc</td>
<td>8,134</td>
<td>926</td>
</tr>
<tr>
<td>commonsense_qa</td>
<td>9,741</td>
<td>1,221</td>
</tr>
<tr>
<td>ai2_arc</td>
<td>1,119</td>
<td>299</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset category: Sentence completion</th>
<th># Train</th>
<th># Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>codah</td>
<td>1665</td>
<td>556</td>
</tr>
<tr>
<td>superglue-copa</td>
<td>400</td>
<td>100</td>
</tr>
<tr>
<td>dream</td>
<td>6116</td>
<td>2040</td>
</tr>
<tr>
<td>quartz-with_knowledge</td>
<td>2696</td>
<td>384</td>
</tr>
<tr>
<td>quartz-no_knowledge</td>
<td>2696</td>
<td>384</td>
</tr>
</tbody>
</table>

Table 2: 26 datasets used for experiments, classified into 6 task categories. # Train and # Test indicate the number of training and test examples of the dataset. Note that # train is based on the original training dataset but we use k random samples for k-shot evaluation.
Forming the demonstrations with an accuracy of $a\%$.

\begin{algorithm}
\caption{Algorithm 1}
\begin{algorithmic}[1]
\Procedure{FormDemons}{$\{(x_i, y_i)\}_{i=1}^a$}
\State $D \leftarrow []$ \Comment{demonstration to be formed}
\State $n \leftarrow k \times a/100$ \Comment{number of correct pairs}
\State $\mathcal{G} \leftarrow \text{Sample(}\text{Range}(1,k), n)$
\For{$i \in \text{Range}(1,k)$}
\If{$i \in \mathcal{G}$} \Comment{add correct pair}
\State $D.\text{append}((x_i, y_i))$
\Else \Comment{add incorrect pair}
\State $D.\text{append}((x_i, \text{Sample}(\mathcal{C} - y_i)))$
\EndIf
\EndFor
\State \Return $D$
\EndProcedure
\end{algorithmic}
\end{algorithm}

English words, which consists of 61,569 words.

Table 4 provides a list of example demonstrations for each method used in Section 5.

C More Experimental Results

C.1 Gold labels vs. random labels

Figure 11 shares the same interface as Figure 3, but all models are evaluated on 3 classification and 3 multi-choice datasets and are thus comparable to each other.

C.2 Random labels from true distribution of labels & Task breakdown

In Section 4, random labels are sampled from the label space from a uniform distribution. We experiment with another variant of demonstrations in the classification tasks, where labels are randomly sampled from the true distribution of labels on the training data. This may have large impact if labels are far from uniform on the training data. Results indicate that performance drop from using gold labels is further reduced compared to using uniformly random labels: with Channel MetaICL, the gap is reduced from 1.9% to 1.3% absolute, and with Channel GPT-J, the gap is reduced from 5.0% to 3.5% absolute.

Figure 12 shows performance gap between using gold labels and using random labels per dataset. We find that the trend that the gap is smaller than previously thought is consistent across most datasets. Nonetheless, there are a few outlier datasets where performance gap is non-negligible, such as financial_phrasebank and a few hate speech detection datasets. Future work may investigate on which tasks the model makes more use of the correctly paired training data.

C.3 More variants of the demonstrations

We explored demonstrations with a constant label where all labels in the demonstrations are replaced with a constant text, “answer”. Specifically, a prediction is made via $\arg\max_{y \in \mathcal{C}} P(y|x_1, \text{answer}...x_k, \text{answer}, x)$. This can be viewed as another way to remove the impact of the label space while keeping the impact of the distribution of the input text. However, results are consistently worse than the results of demonstrations with random English labels. We think this is because constant labels actually change the format of the demonstrations, since they can be viewed as part of a separator between different demonstration examples.

We also explored demonstrations with the test input where all inputs in the demonstrations are replaced with the test input, each paired with a ran-
dom label. Specifically, a prediction is made via
\[ \arg\max_{y \in C} P(y|x, \tilde{y}_1 ... x, \tilde{y}_k, x), \]
where \( \tilde{y}_i \) (1 \( \leq \) \( i \) \( \leq \) \( k \)) is randomly sampled at uniform from \( C \). This variant is seemingly a reasonable choice given that it satisfies the condition that the inputs in the demonstrations come from the same distribution as the test input (since they are identical), and using random labels is as good as using gold labels. Nonetheless, we find that this variant is significantly worse than most other methods with demonstrations. We think this is because using the constant input for all demonstration example significantly changes the format of the sequence, since the input can be viewed as part of a separator between different demonstration examples.
Figure 12: Performance gap from using the demonstrations with gold labels to using the demonstrations with random labels. Datasets are sorted in descending order. The top two figures use random labels that are sampled at uniform, with Channel MetaICL and Channel GPT-J, respectively. The bottom two figures use random labels that are sampled from a true distribution of labels on the training data, with Channel MetaICL and Channel GPT-J, respectively.
### Dataset Type Example

#### MRPC

| Minimal | sentence 1: Cisco pared spending to compensate for sluggish sales. [SEP] sentence 2: In response to sluggish sales, Cisco pared spending. | \{equivalent,not_equivalent\} |
| Manual | Cisco pared spending to compensate for sluggish sales. The question is: In response to sluggish sales, Cisco pared spending. True or False? | \{True, False\} |

#### RTE

| Minimal | sentence 1: The girl was found in Drummondville. [SEP] sentence 2: Drummondville contains the girl. | \{entailment,not_entailment\} |
| Manual | The girl was found in Drummondville. The question is: Drummondville contains the girl. True or False? | \{True, False\} |

#### Tweet_eval-hate

| Minimal | The Truth about #Immigration | \{hate,non-hate\} |
| Manual | Tweet: The Truth about #Immigration | \{against,favor\} |

#### SICK

| Minimal | sentence 1: A man is screaming. [SEP] sentence 2: A man is scared. | \{contradiction,entailment,neutral\} |
| Manual | A man is screaming. The question is: A man is scared. True or False? | \{False, True, Not sure\} |

#### poem-sentiment

| Minimal | willis sneered: \{negative, no_impact, positive\} |
| Manual | willis sneered: The sentiment is: \{negative, no_impact, positive\} |

#### OpenbookQA

| Minimal | What creates a valley? | \{feet, rock, water, sand\} |
| Manual | The question is: What creates a valley? | \{feet, rock, water, sand\} |

#### CommonsenseQA

| Minimal | What blocks sunshine? | \{summer, park, desktop, sea, moon\} |
| Manual | The question is: What blocks sunshine? | \{summer, park, desktop, sea, moon\} |

#### COPA

| Minimal | Effect: I coughed. | \{Cause: I inhaled smoke, Cause: I lowered my voice\} |
| Manual | I coughed because | \{I inhaled smoke, I lowered my voice\} |

#### ARC

| Minimal | Which biome has the most vegetation? | \{desert, forest, grassland, tundra\} |
| Manual | The question is: Which biome has the most vegetation? | \{desert, forest, grassland, tundra\} |

#### Table 3: Example demonstrations when using methods in Section 5. The \textit{financial_phrasebank} dataset with $\mathcal{C} = \{\text{“positive”, “neutral”, “negative”}\}$ is used. \textcolor{red}{Red} text indicates the text is sampled from an external corpus; \textcolor{blue}{blue} text indicates the labels are randomly sampled from the label set; \textcolor{purple}{purple} text indicates a random English word.

#### Table 4: Example demonstrations when using methods in Section 5. The \textit{financial_phrasebank} dataset with $\mathcal{C} = \{\text{“positive”, “neutral”, “negative”}\}$ is used. \textcolor{red}{Red} text indicates the text is sampled from an external corpus; \textcolor{blue}{blue} text indicates the labels are randomly sampled from the label set; \textcolor{purple}{purple} text indicates a random English word.

| Demos w/ gold labels | (Format ✓ Input distribution ✓ Label space ✓ Input-label mapping ✓) | Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \{positive\} Panostaja did not disclose the purchase price. \{neutral\} |
| Demos w/ random labels | (Format ✓ Input distribution ✓ Label space ✓ Input-label mapping ✓) | Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \{neutral\} Panostaja did not disclose the purchase price. \{negative\} |
| OOD Demos w/ random labels | (Format ✓ Input distribution ✓ Label space ✓ Input-label mapping ✓) | Colour-printed lithograph. Very good condition. Image size: 15 x 23 1/2 inches. \{neutral\} Many accompanying marketing claims of cannabis products are often well-meaning. \{negative\} |
| Demos w/ random English words | (Format ✓ Input distribution ✓ Label space ✓ Input-label mapping ✓) | Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. \{unanimity\} Panostaja did not disclose the purchase price. \{wave\} |
| Demos w/o labels | (Format ✓ Input distribution ✓ Label space ✓ Input-label mapping ✓) | Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. Panostaja did not disclose the purchase price. |
| Demos labels only | (Format ✓ Input distribution ✓ Label space ✓ Input-label mapping ✓) | Circulation revenue has increased by 5% in Finland and 4% in Sweden in 2008. Panostaja did not disclose the purchase price. |

11064