Prompt-based learning (i.e., prompting) is an emerging paradigm for exploiting knowledge learned by a pretrained language model. In this paper, we propose Automatic Multi-Label Prompting (AMuLaP), a simple yet effective method to automatically select label mappings for few-shot text classification with prompting. Our method exploits one-to-many label mappings and a statistics-based algorithm to select label mappings given a prompt template. Our experiments demonstrate that AMuLaP achieves competitive performance on the GLUE benchmark without human effort or external resource.

1 Introduction

Since the release of GPT-3 (Brown et al., 2020), several studies have focused on exploiting pretrained language models with only a few training examples (Brown et al., 2020; Gao et al., 2021; Shin et al., 2020). These works demonstrate the potential of using natural language prompts to encourage the model to recall similar patterns in its training corpus and thus make accurate predictions. This setting of few-shot learning is closer to how humans learn to solve a task, often without many examples as in a traditional deep learning paradigm. The use of prompts can strengthen the explicit connection between input and output, helping the model exploit the knowledge learned from pretraining in a better way. Furthermore, recent works (Schick and Schütze, 2021a,b; Gao et al., 2021) show that prompts can also help the model generalize better in fine-tuning.

Prompt-based learning (i.e., prompting) aims to use a template to convert the original input into a prompt-based input with some unfilled masked tokens, and then use the pretrained language model to fill these masked tokens, and finally the tokens filled into these slots are mapped to the corresponding labels as the final output. In prompting, the design of prompts often plays an important role. Many attempts have been made in this emerging direction of prompt engineering (Shin et al., 2020; Gao et al., 2021). Meanwhile, finding a good mapping from the original task labels to tokens (i.e., label engineering) is also critical to few-shot performance, as found in Schick et al. (2020); Gao et al. (2021). However, manually assigning the label mapping requires human expertise with trial and error. One may argue that the same effort can be used to label more supervised data for a conventional deep learning pipeline. Thus, an efficient automatic label mapping method is desirable.

In this paper, we aim to design a method that can automatically find a good label mapping to save human effort from label engineering. We propose Automatic Multi-Label Prompting (AMuLaP), a simple yet effective method to tackle the label selection problem for few-shot classification. AMuLaP is a parameter-free statistical technique that can identify the label patterns from a few-shot training set given a prompt template. AMuLaP exploits multiple labels to suppress the noise and inherently extend the training set for prompt-based fine-tuning. Compared with a hand-crafted label mapping and previous works on automatic label mapping (Schick et al., 2020; Gao et al., 2021), AMuLaP achieves competitive performance despite being simpler and does not require access to the weights of the backbone model, or an external text-infilling model. We conduct extensive experiments and demonstrate the effectiveness of our method under multiple settings. Moreover, we attempt to scale AMuLaP with different sizes of the training set and find AMuLaP to work surprisingly well even with one or two shots. To understand the few-shot performance on different datasets, we investigate the relation between accuracy and interclass token distribution divergence, shedding light...
on the explainability of few-shot text classification.

2 Related Work

Discrete Prompts The release of GPT-3 (Brown et al., 2020) has led to interest in prompting, a new way to leverage pretrained language models (PLM). Brown et al. (2020) proposes an intuitive in-context learning paradigm by concatenating a few input and output examples and feeding them to the language model and let the model autoregressively generate answers for new examples. Recent works (Petroni et al., 2019; Davison et al., 2019; Jiang et al., 2020) design prompts to probe the factual and commonsense knowledge encoded within a PLM. Recent works (Schick and Schütze, 2021a,b; Gao et al., 2021) demonstrate that even smaller PLMs have similar few-shot learning capacity. Le Scao and Rush (2021) analyzes the effect of prompting and concludes that a single prompt may be worth 100 training examples in fine-tuning.

Instead of manually designing prompts (i.e., prompt engineering), some recent studies also explore automatic prompt generation. PETAL (Schick et al., 2020) augments Pattern Exploiting Training (PET, Schick and Schütze, 2021a,b) with automatically identified label words; Gao et al. (2021) searches the vocabulary for label words by fine-tuning the model on the candidates generated by a text-infilling model and using an external generation model for data augmentation of prompt templates; AutoPrompt (Shin et al., 2020) uses a gradient-based search to determine both prompts and label words. However, these methods require parameter updates with gradient descent, which is infeasible without access to the model weights (e.g., GPT-3). PET and its variants also require a large unlabeled set and need to be fine-tuned multiple times. AutoPrompt uses discretization techniques to approximately map a continuous vector back to tokens in the vocabulary (i.e., “vocabulary”). These searched prompts and labels are often uninterpretable by humans. Guo et al. (2021) introduces Q-Learning to optimize the soft prompt. Different from these prior studies, our proposed AMuLaP is a simple and interpretable method for few-shot prompting that can work well with and without access to model weights.

Continuous Prompts In parallel with text-based discrete prompts, there is also a line of work focused on tuning only a fraction of parameters of an LM with the help of continuous prompts (i.e., soft prompts). Zhong et al. (2021) and Qin and Eisner (2021) propose continuous prompts for knowledge probing by tuning some trainable vectors in the input sequence while fixing the rest of the input. Li and Liang (2021) applies a similar method for natural language generation and achieves comparable performance to fine-tuning while updating only 0.1% of model parameters. Lester et al. (2021) reveals that prompt tuning is more competitive when scaled up and can achieve identical performance to conventional fine-tuning when the model is large enough. Notably, different from discrete prompting, these works often use all training data to update model weights. Different from these works, AMuLaP is a discrete prompting method that has better interpretability and works well in the few-shot setting.

3 Prompting for Few-Shot Classification

We follow the setup in LM-BFF (Gao et al., 2021) for few-shot text classification. Given a pretrained language model $L$, a task $D$ and its defined label space $Y$, we have $n$ training examples per class for the training set $D_{train}$. As pointed out in Perez et al. (2021), using the full development set may be misleading to claim a few-shot setting. Thus, we use a few-shot development set with the same size as the training set (i.e., $|D_{train}| = |D_{dev}|$), to be consistent with Gao et al. (2021) and constitute a “true few-shot” setting (Perez et al., 2021).

For an input example $x$ (a single sentence or a sentence pair), we first use a task-specific template $T$ to convert it to $x'$, a token sequence with a $[\text{MASK}]$ token. We then map the original label space to a set of selected words from the vocabulary, denoted as $M : \mathcal{Y} \rightarrow \mathcal{V}'$. Some examples of $T$ and $M$ are shown in Table 1. Note that since we focus on automatically finding the label mapping $M$, we use the manual templates $T$ from Gao et al. (2021) throughout this paper. Since $L$ is trained to complete the $[\text{MASK}]$ token in an input sequence, we can directly make zero-shot prediction of the probability of class $y \in \mathcal{Y}'$ by the masked language modeling:

$$p(y|x) = p([\text{MASK}] = M(y) | x').$$

Alternately, one can further fine-tune $L$ with supervised pairs $\{x', M(y)\}$ to achieve even better performance.
4 Automatic Multi-Label Prompting

4.1 Exploiting Multiple Labels

Selecting one label word can be insufficient for some complicated tasks, as mentioned in Schick et al. (2020). We also argue that selecting only one label (especially automatically) may bring noise. This can be resolved by introducing multiple label words. Schick et al. (2020) use multiple label combinations for PET (Schick and Schütze, 2021a) and ensemble them afterwards. We instead use a straightforward sum to consider multiple label words when making predictions. This design has a similar advantage of exploiting multiple labels without training and ensembling multiple models.

Instead of a one-to-one mapping from the original label space \( \mathcal{Y} \) to \( \mathcal{V} \), we map each \( y \in \mathcal{Y} \) to its label word set \( S(y) \). We denote the mapping as \( \mathcal{M} : \mathcal{Y} \rightarrow \mathcal{S}(y) \). For class \( y \in \mathcal{Y} \), the predicted probability is calculated as:

\[
p(y|x) = \sum_{v \in \mathcal{S}(y)} p([\text{MASK}] = v | x)
\]

Then, we can simply make predictions by selecting the label with the largest likelihood.

Similarly, if we need to fine-tune \( \mathcal{L} \) with supervised pairs, instead of optimizing the cross-entropy loss between the gold label and a single token, we optimize the loss between the sum of the output probabilities of \( \mathcal{S}(y) \) and the gold label with a cross-entropy loss:

\[
l = - \sum_{x \in \mathcal{D}_{\text{train}}} \sum_{y \in \mathcal{Y}} [\mathbb{1}[y = \hat{y}] \cdot \log p(y|x)]
\]

where \( \hat{y} \) is the ground truth label for the input \( x \) and \( p(y|x) \) is defined in Equation 2.

4.2 Automatic Label Selection

Finding a good label mapping \( \mathcal{M} \) is non-trivial, especially when \( \mathcal{M}' \) maps an original label to a set of label words instead of one. Selecting a good label mapping often requires significant human effort, including domain knowledge and trial-and-error. Previously, Schick and Schütze (2021a,b) both use hand-crafted label mappings while Schick et al. (2020) explores automatic label mapping searching but it still requires manual pre-filtering and significantly underperforms the manual mapping. Gao et al. (2021) exploits a large pretrained text infilling model (T5, Raffel et al., 2020) to fill in the label words and then determine the final mapping by fine-tuning on all of them and selecting the best one with \( \mathcal{D}_{\text{dev}} \). We introduce a new selection algorithm for label mapping that achieves competitive results compared to previous efforts.

We aim to achieve two goals: (1) Selecting the most likely label mapping based on the training set. For example, in a sentiment classification task, we would like to see positive words in the label set of the “positive” class while negative words in the label set of the “negative” class. A simple solution is to select the \( k \) most likely tokens predicted for

| Task       | Template            | Class            | Manual (2021) | Selected Labels by AMuLaP |
|------------|---------------------|------------------|---------------|---------------------------|
| MNLI       | \(<S_1> \text{? [MASK]} , <S_2>\) | entailment, neutral, contradiction | Yes | Yes, Indeed, Also, Currently |
|            |                     |                  | Maybe         | Historically, Suddenly, Apparently, And |
|            |                     |                  | No            | No, However, Instead, Unfortunately |
| SST-2      | \(<S_1> \text{It was [MASK]} \) | positive, negative | great         | great, perfect, fun, brilliant |
|            |                     |                  | terrible      | terrible, awful, disappointing, not |
| QNLI       | \(<S_1> \text{? [MASK]} , <S_2>\) | entailment, not_entailment | Yes        | Yes, Historically, Overall, Indeed |
|            |                     |                  | No            | Well, First, However, Unfortunately |
| RTE        | \(<S_1> \text{? [MASK]} , <S_2>\) | entailment, not_entailment | Yes        | Yes, Today, Specifically, Additionally |
|            |                     |                  | No            | However, Ironically, Also, Indeed |
| MRPC       | \(<S_1> \text{[MASK]} , <S_2>\) | equivalent, not_equivalent | Yes        | \(</s>\), Currently, Additionally, Today |
|            |                     |                  | No            | However, Meanwhile, Overall, Finally |
| QQP        | \(<S_1> \text{[MASK]} , <S_2>\) | equivalent, not_equivalent | Yes        | Or, So, Specifically, Actually |
|            |                     |                  | No            | Also, And, Finally, Well |
| CoLA       | \(<S_1> \text{This is [MASK]} \) | grammatical, not_grammatical | correct     | why, true, her, amazing |
|            |                     |                  | incorrect     | it, ridiculous, interesting, sad |

Table 1: The manual and automatically selected labels by AMuLaP. The templates used for prompting are from Gao et al. (2021).
the [MASK] token in the training examples of each class \( y \). However, in practice, we would find common words in more than one label set. For example, if we simply take the 10 most likely tokens for the SST-2 dataset (Socher et al., 2013), we would find “good” in both positive and negative label sets, although it is ranked second place in the positive set and ninth in the negative set. Thus, we want to make sure that (2) each token only belongs to at most one label set where it has the highest probability. To ensure this, we have to iterate over the vocabulary and check that for every token. Then, we can truncate the candidate sets of each class and select the \( k \) most likely tokens from each set. The time complexity of this algorithm is \( O(k \cdot |\mathcal{V}| \cdot |\mathcal{Y}|) \).

Formally, we select \( \mathcal{M}(y) : \mathcal{Y} \rightarrow \hat{\mathcal{S}}(y) \) by the following steps:

1. For each \( y_i \in \mathcal{Y} \), we iterate through all training samples \( x_j \in D_{\text{train}} \) whose ground truth label \( \hat{y}_j = y_i \). We use \( \mathcal{L} \) to predict the token probability of the [MASK] token and take the average of the predicted probabilities of the \( n \) examples to be \( z_i \).

2. Initialize an empty mapping \( \hat{\mathcal{M}} : \mathcal{Y} \rightarrow \hat{\mathcal{S}}(y) \).

3. For each \( v \in \mathcal{V} \) where \( \mathcal{V} \) is the vocabulary of the model \( \mathcal{L} \), we retrieve \( v \)’s probability value \( z_i^v \) from \( z_i \) of each class.

4. We assign \( v \) to the most likely candidate token set of the \( m \)-th class \( \hat{\mathcal{S}}(y_m) \) where \( m = \arg\max_i z_i^v \).

5. For \( i \in |\mathcal{Y}| \), we choose the top-\( k \) tokens from \( \hat{\mathcal{S}}(y_i) \) with the largest probability \( z_i^v \) and obtain the truncated mapping \( \mathcal{M} : \mathcal{Y} \rightarrow \mathcal{S}(y) \).

5 Experiments

5.1 Experimental Setting

Datasets We evaluate seven classification tasks of the GLUE benchmark (Wang et al., 2019). Specifically, we test on Microsoft Research Paraphrase Matching (MRPC) (Dolan and Brockett, 2005), Quora Question Pairs (QQP)\(^2\) for Paraphrase Similarity Matching; Stanford Sentiment Treebank (SST-2) (Socher et al., 2013) for Sentiment Classification; Multi-Genre Natural Language Inference Matched (MNLI-mm) (Williams et al., 2018), Question Natural Language Inference (QNLI) (Rajpurkar et al., 2016) and Recognizing Textual Entailment (RTE) (Wang et al., 2019) for the Natural Language Inference (NLI) task; The Corpus of Linguistic Acceptability (CoLA) (Warstadt et al., 2019) for Linguistic Acceptability. We use the manual templates in Gao et al. (2021), as listed in Table 1. The metrics for each dataset are indicated in Table 2.

Baselines We compare our method to various baselines:

- **Majority**: always predict the majority class in the test set.
- **GPT-3-style in-context learning** (Brown et al., 2020): present a few examples to the language model and make it directly predict the next token as the prediction.
- **Manual prompts**: we use the human-designed prompts in Gao et al. (2021).
- **PETAL-CE** (Schick et al., 2020): the variant of PETAL using the cross-entropy metric.
- **PETAL-LR** (Schick et al., 2020): the variant of PETAL using the likelihood ratio metric.
- **AutoL-T5** (Gao et al., 2021): the automatic label searching method with an external text-infilling model, T5-3B (Raffel et al., 2020).

Task Setup We closely follow the setup in Gao et al. (2021). We sample \( n \) training examples and \( n \) development examples per class. We set \( k = 16 \) throughout all experiments. We use RoBERTa-large (Liu et al., 2019) as the backbone LM \( \mathcal{L} \). For each reported result, we measure average performance across 5 different randomly sampled \( D_{\text{train}} \) and \( D_{\text{dev}} \) splits. Following Gao et al. (2021), the original development split of each dataset is used as the test set in our experiments. We also report the standard deviation for each result. To fairly compare with different baselines, we consider the following three settings:

- **Setting 1**: We only use \( D_{\text{train}} \) alone for both label selection and tuning \( k \). The parameters of \( \mathcal{L} \) are not updated. \( D_{\text{dev}} \) is not used. This setting is for fair comparison with in-context learning.
Table 2: Experimental results under three settings with RoBERTa-large as \( \mathcal{L} \). We report the average of 5 runs along with their standard deviation in the parentheses.

| Settings | \( \mathcal{L} \) | \( \mathcal{D}_{\text{train}} \) | \( \mathcal{D}_{\text{dev}} \) | \( \mathcal{D}_{\text{test}} \) |
|----------|-----------------|-----------------|-----------------|-----------------|
| Setting 1: \( \mathcal{D}_{\text{train}} \), only | | | | |
| In-context learning (2020) | 52.0 (0.7) | 53.4 (0.6) | 84.8 (1.3) | 353 |
| AMuLaP (ours) | 50.8 (2.1) | 52.3 (1.8) | 86.9 (1.6) | 350 |
| Setting 2: \( \mathcal{D}_{\text{train}} + \mathcal{D}_{\text{dev}} \): No parameter update. | | | | |
| PETAL-CE (2020) | 48.8 (2.6) | 49.7 (2.3) | 75.6 (7.2) | 349 |
| PETAL-LR (2020) | 38.6 (2.0) | 38.4 (2.1) | 85.3 (3.3) | 347 |
| AutoL-T5 (2021) | 41.6 (5.4) | 42.3 (6.2) | 84.3 (3.3) | 346 |
| AutoL (Schick et al., 2020) | 50.8 (2.1) | 52.2 (1.9) | 87.0 (1.5) | 345 |
| T5 + AMuLaP (ours) | 52.9 (3.0) | 54.2 (2.7) | 90.1 (0.4) | 343 |
| Setting 3: \( \mathcal{D}_{\text{train}} + \mathcal{D}_{\text{dev}} + \mathcal{D}_{\text{test}} \): Prompt-based fine-tuning. | | | | |
| Fine-tuning | 45.8 (6.4) | 47.8 (6.8) | 81.4 (3.8) | 341 |
| Manual Label FT (2021) | 68.3 (2.3) | 70.5 (1.9) | 92.7 (0.9) | 340 |
| PETAL-CE FT (2021) | 57.5 (3.2) | 57.7 (2.6) | 92.6 (1.0) | 339 |
| PETAL-LR FT (2020) | 64.0 (6.5) | 65.9 (6.4) | 92.9 (1.7) | 338 |
| AutoL-T5 FT (2021) | 64.8 (4.7) | 67.5 (4.3) | 93.5 (0.5) | 337 |
| AutoL-FT (2020) | 70.6 (2.7) | 72.5 (2.4) | 93.2 (0.7) | 336 |
| T5 + AMuLaP FT (ours) | 68.5 (2.2) | 71.1 (2.3) | 93.4 (1.0) | 335 |

**5.2 Experimental Results**

We demonstrate experimental results under three settings in Table 2. Under Setting 1, AMuLaP outperforms GPT-3-style in-context learning by 4.5 in terms of the average score and outperforms zero-shot inference with manually designed labels by 2.4. Under Setting 2, compared to variants of PETAL (Schick et al., 2020), AMuLaP has an advantage of 5.8 and 8.5 in terms of the average score over CE and LR, respectively. Notably, AMuLaP even outperforms AutoL-T5 by 1.3 without using any external model or data. Additionally, we attempt to replace the predicted token distribution of AMuLaP with the validation score of T5-filled labels (Gao et al., 2021). With the help of an external model T5, AMuLaP outperforms AutoL-T5 by a considerable margin of 3.8 in terms of the average score, verifying the versatility of our multi-label mechanism and label selection algorithm. Under Setting 3, AMuLaP FT outperforms all baselines including AutoL-T5. However, we do not observe significant improvements when combining AMuLaP with T5. Generally speaking, methods with pa-

---

3The validation scores of T5-found labels are obtained on \( \mathcal{D}_{\text{dev}} \), as described in Gao et al. (2021). No external data used.
As shown in Table 3, we list the 10 most likely label mappings output by PETAL (Schick et al., 2020), AutoL-T5 (Gao et al., 2021) and our AMuLaP. Suitable labels annotated by the human annotator are underlined.

Table 3: Most likely label mapping for the SST-2 dataset obtained by PETAL (Schick et al., 2020), AutoL-T5 (Gao et al., 2021) and our AMuLaP. Suitable labels annotated by the human annotator are underlined.

| Class | PETAL-CE (Schick et al., 2020) | PETAL-LR (Schick et al., 2020) |
|-------|--------------------------------|---------------------------------|
| **positive** | amazing, great, brilliant, perfect, fun, wonderful, beautiful, fantastic, awesome, not | superb, fearless, acclaimed, addictive, visionary, immersive, irresistible, timely, unforgettable, gripping |
| **negative** | not, awful, fun, funny, terrible, great, amazing, hilarious, awesome, good | annoying, insulting, meaningless, lame, shitty, humiliating, childish, stupid, embarrassing, irritating |

| Class | AutoL-T5 (Gao et al., 2021) | AMuLaP (ours) |
|-------|-----------------------------|---------------|
| **positive** | exquisite, perfection, effective, fabulous, intense inspiring, spectacular, sublime, astounding, thrilling | great, perfect, fun, brilliant, amazing, good, wonderful, beautiful, excellent, fantastic |
| **negative** | embarrassing, boring, frustrating, ridiculous, awkward silly, nothing, disgusting, ugly, confusing | terrible, awful, disappointing, not, horrible, obvious, funny, inevitable, bad, boring |

Table 4: Experimental results for the ablation study. We report the average of 5 runs along with their standard deviation in the parentheses.

| Setting | MNLI (acc) | MNLI-mm (acc) | SST-2 (acc) | QNLI (acc) | RTE (acc) | MRPC (F1) | QQP (F1) | CoLA (Matt.) | Avg. |
|---------|------------|---------------|-------------|------------|-----------|-----------|----------|-------------|------|
| Setting 2: $D_{base} + D_{dev}$; No parameter update. | | | | | | | | | |
| AMuLaP | 50.8 (2.1) | 52.2 (1.9) | 87.0 (1.5) | 53.5 (2.3) | 59.1 (7.4) | 56.7 (5.7) | 61.5 (1.7) | 2.6 (1.8) | 52.9 |
| w/o dedup. | 45.4 (2.7) | 46.5 (2.5) | 87.9 (1.0) | 53.8 (3.0) | 54.6 (6.0) | 66.7 (12.3) | 57.2 (2.1) | 2.5 (4.2) | 51.8 |
| k = 1 | 46.5 (2.7) | 48.4 (2.6) | 68.8 (12.0) | 51.9 (1.6) | 58.8 (12.7) | 55.0 (4.8) | 59.2 (0.0) | 5.6 (2.1) | 49.3 |
| Setting 3: $D_{base} + D_{dev}$; Prompt-based fine-tuning. | | | | | | | | | |
| AMuLaP FT | 70.6 (2.7) | 72.5 (2.4) | 93.2 (0.7) | 65.1 (5.9) | 65.9 (6.3) | 79.3 (4.0) | 69.1 (2.5) | 18.3 (9.4) | 66.8 |
| w/o dedup. | 56.9 (5.4) | 58.2 (5.2) | 92.8 (0.9) | 50.6 (0.4) | 57.1 (10.8) | 79.2 (3.6) | 55.0 (26.0) | 5.6 (7.1) | 56.9 |
| k = 1 | 67.7 (4.1) | 69.8 (3.8) | 92.6 (1.0) | 65.9 (5.2) | 63.1 (8.0) | 80.2 (3.8) | 66.7 (3.2) | 19.3 (15.5) | 65.7 |
| random $M$ | 58.8 (6.2) | 61.1 (6.2) | 92.1 (2.1) | 62.1 (7.1) | 57.0 (11.2) | 74.7 (9.2) | 60.8 (5.8) | 31.0 (13.9) | 62.2 |
| random $M$ (k = 1) | 52.6 (7.8) | 55.4 (8.3) | 89.0 (4.9) | 65.2 (4.5) | 55.2 (6.2) | 73.4 (10.6) | 60.7 (3.7) | 17.3 (14.7) | 58.6 |

Table 4: Experimental results for the ablation study. We report the average of 5 runs along with their standard deviation in the parentheses.

6.1 Case Study

As shown in Table 3, we list the 10 most likely label mappings output by PETAL (Schick et al., 2020), AutoL-T5 (Gao et al., 2021) and our AMuLaP. Suitable labels annotated by the human annotator are underlined.

6.2 Ablation Study

As shown in Table 4, we evaluate the effect of each design choice on the GLUE benchmark. For both non-finetuning and prompt-based fine-tuning settings, our deduplication algorithm can effectively improve the overall performance by 1.1 and 9.9 in terms of the GLUE average score, respectively. Notably, deduplication is especially important for prompt-based fine-tuning since if the same label maps to two classes, optimization would be difficult due to the contradiction of supervision signals. Also, our multi-label strategy is shown to be effective at improving the average GLUE scores by 3.6 and 1.1 for non-finetuning and fine-tuning settings, respectively. Moreover, a random label mapping...
leads to lower performance than a label mapping selected based on the training set except for CoLA. An interesting observation is that the random mapping even outperforms all label selection methods in Table 2 (both manual and automatic) and is close to the fine-tuning baseline. We will discuss this more in Section 6.4.

6.3 Scaling Few-Shot Learning

Le Scao and Rush (2021) explore the scaling law of PET (Schick and Schütze, 2021a) when using more examples for training. Similarly, in this section, we aim to test how AMuLaP scales to different training set sizes \( n \). Figure 1 illustrates how standard fine-tuning and our AMuLaP with non-finetuning and fine-tuning compare as \( n \) increases. For MNLI and SST-2 task, AMuLaP outperforms standard fine-tuning when we use no more than 16 training examples for non-finetuning and fine-tuning setting. When using more than 16 training examples, AMuLaP under fine-tuning setting still outperforms standard fine-tuning. For an easier task like SST-2, although only 32 training examples are used, the performance of our AMuLaP with non-finetuning and fine-tuning is close to saturation and can be comparable to standard fine-tuning on the entire dataset. For a harder task like MNLI, although the performance of AMuLaP under non-finetuning setting gradually becomes saturated as \( n \) increases, AMuLaP under fine-tuning settings continues to improve as \( n \) increases and continues to outperform the standard fine-tuning. For MRPC, although the performance of our AMuLaP and standard fine-tuning fluctuate as \( n \) increases, in general, AMuLaP with fine-tuning can still achieve comparable performance to standard fine-tuning. In addition, the results demonstrate the effectiveness of AMuLaP especially for extreme few-shot settings. With only one example, AMuLaP achieves decent performance while standard fine-tuning is close to random.

6.4 Understanding Few-Shot Performance

As shown in Table 1, AMuLaP seems to be able to find good labels for some datasets while failing on others. Intuitively, this phenomenon should correspond to classification performance. To quantitatively understand the relation between the quality of found labels and the final performance, we design a meta-experiment. For every two-class dataset (all GLUE datasets except three-class MNLI), we calculate the JS divergence between the average predicted token probabilities \( z_0 \) and \( z_1 \). This metric measures how different the language model \( L \) considers the examples from two classes. This can be regarded as the “confidence” of \( L \) to distinguish between the two classes. If the model can easily distinguish examples from one class with the other, we would expect the divergence to be large, and vice versa.

We illustrate the relation between inter-class JS divergence and the performance of AMuLaP on...
As we analyze, increasing \( k \) (end-of-sequence marker) label found by AMuLaP, (2020) argues that one single label sometimes can- not observe significant improvement when continuing increasing \( k \) with labels selected by AMuLaP. As we analyze, increasing \( k \) harms the overall quality of selected labels and thus overrides the benefit of a larger \( k \). In general, we do not observe a clear law for choosing the best \( k \) for AMuLaP. As mentioned before, \( k \) can influence both the overall quality of labels (in both ways) and the training procedure (for fine-tuning). Thus, for the optimal performance, we find it essential to search \( k \) with a development set.

**Limitations and Future Directions** In this paper, we only focus on the selection of the label mapping with a fixed prompt template. There is more to explore when considering the prompt template at the same time. Similar to our paper, previous works (Schick et al., 2020; Gao et al., 2021) separately search for a prompt template \( \tau \) and the label mapping \( M \). However, these two variables are closely related and greedily search for the best template \( \tau \) then the best mapping under \( \tau \) may be suboptimal. Jointly searching for \( \tau \) and \( M \) could be a promising direction for future research.

More broadly, we would like to point out some limitation and contradictions within current few-shot prompting techniques. There is a natural contradiction between performance and access to the model weights. Brown et al. (2020) highlights few-shot prompting as a way to mitigate their de- cision to not release the model weights. However, as shown in our Table 2, with the same backbone model \( \mathcal{L} \), GPT-3-style in-context learning and other methods that do not access the model weights generally underperform those with access to the model weights by a large margin. Also, in-context learning cannot handle more training examples due to the maximum length limit of the model while AMuLaP without fine-tuning gets saturated quickly, as shown in Figure 1.

In addition, complicated prompting techniques are not practically useful for real-world scenarios. For most techniques, the required effort for finding good templates and label mappings, and sometimes training models outweighs the cost of simply labeling more training examples. As shown in Figure 2, 64 examples per class are enough to bring the performance of standard fine-tuning to the same level of prompting. Although recent works on automatic selection of prompts and label mappings are making meaningful contribution to the practica- bility of few-shot learning, we believe more work should be done to simplify the learning procedure and eliminate human effort while achieving good performance.
Transformers: State-of-the-art natural language processing. In *EMNLP (Demos)*, pages 38–45. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: learning vs. learning to recall. In *NAACL-HLT*, pages 5017–5033. Association for Computational Linguistics.