Textual Few-Shot Classification For API-based Models

Anonymous authors
Paper under double-blind review

Abstract

Proprietary and closed APIs are becoming increasingly common for large language models such as GPT4 and ChatGPT, and are impacting the practical applications of natural language processing, including few-shot classification. Few-shot classification involves training a model to perform a new classification task with a handful of labeled data. This paper presents three contributions. First, we introduce a scenario where a pre-trained model is served through a gated API with compute-cost and data-privacy constraints. Second, we propose a transductive inference, a learning paradigm that has been overlooked by the NLP community. Transductive inference, unlike traditional inductive learning, leverages the statistics of unlabelled data. We also introduce a new parameter-free transductive regularizer based on the Fisher-Rao loss, which can be used on top of the gated API embeddings. This method fully utilizes unlabelled data, does not share any label with the third-party API provider and could serve as a baseline for future research. Third, we propose an improved experimental setting and compile a benchmark of eight datasets involving multiclass classification in four different languages, with up to 151 classes. We evaluate our methods using eight backbone models, along with an episodic evaluation over 1,000 episodes, which demonstrate the superiority of transductive inference over the standard inductive setting.

1 Introduction

Recent advances in Natural Language Processing (NLP) have been largely driven by the scaling paradigm (Kaplan et al., 2020; Rosenfeld et al., 2019), where larger models with increased parameters have been shown to achieve state-of-the-art results in various NLP tasks (Touvron et al., 2023; Radford et al., 2019). This approach has led to the development of foundation models such as ChatGPT (Lehman et al., 2023; Kocoń et al., 2023), GPT-4 (OpenAI, 2023), GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), and BERT (Devlin et al., 2018), which have achieved unprecedented performance in text classification (Liu et al., 2019b), language modeling, machine translation (Fan et al., 2021), and coding tasks (Chen et al., 2021a).

Despite the success of the scaling paradigm, significant challenges still exist especially when the many practical constraints of real-world scenarios have to be met: labeled data can be severely limited (*i.e.,* few-shot scenario (Song et al., 2022; Ye et al., 2021)), data privacy is critical for many industries and has become the subject of increasingly many regulatory pieces (Commission, 2020; 2016), compute costs need to be optimized (Strubell et al., 2019). Furthermore, these challenges are made even more complex as stronger foundation models are now available only through APIs (*e.g.*, OpenAI's GPT-3, GPT-4 or ChatGPT, Anthropic's Claude or Google's PaLM (Chowdhery et al., 2022)) which has led to some of their parameters being concealed, presenting new challenges for model adaptation (Solaiman, 2023). This paper is centered on the fundamental task of few-shot text classification, specifically focusing on cloud-based/API access. Specifically, we formulate three requirements for API-based few-shot learning (see Fig. 1):

(R1) Black-box scenario. We focus on learning from models that are opaquely deployed in production to the end-user, who only has access to the end-point of the encoder, *i.e.*, the resulting text embedding produced by the final layer of the network.

(R2) Low resources / computation time. AI systems are often required to make rapid predictions at high frequencies in various real-world applications. Therefore, any few-shot classifier used in such

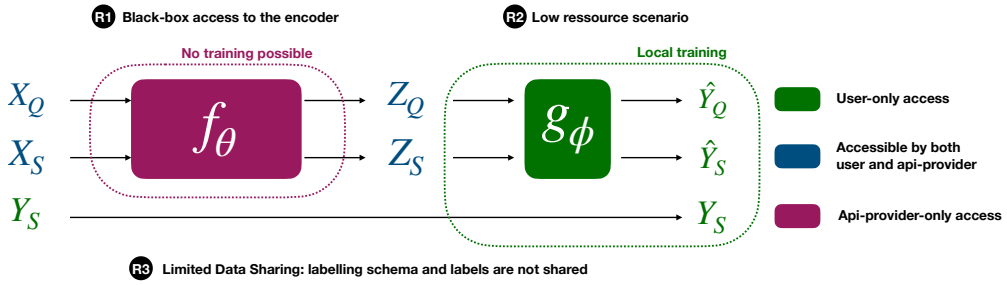

Figure 1: API-based few-shot learning scenario. The black-box API providing embeddings from the pre-trained encoder f_θ. The black-box scenario discard existing inductive approaches and in-context learning methods due to inaccessible of model's parameters (**(R1)**) and privacy concerns (**(R3)**). This scenario, allows to tune a classification head g_ϕ (using induction or transduction) at low computational cost **(R2)**, while retaining all support labels locally.

scenarios should have a low training and inference time, as well as require minimal computational resources.

(R3) Limited Data Sharing. When utilizing API models, data sharing becomes a major concern. In the current landscape, providers are increasingly offering less transparent procedures for training their networks. As a result, users prefer sharing as little information as possible, such as labeling schema and annotated data, to safeguard their data privacy.

While numerous previous studies have addressed the popular *few-shot* classification setting, to our knowledge no existing line of work adequately satisfies the three API requirements described above. In particular, prompt-based FSL (Schick & Schütze, 2020a) and parameter-efficient fine-tuning FSL (Houlsby et al., 2019) both require access to the model's gradients, while in-context learning scales poorly with the task's size (*e.g* number of shots, number of classes) (Chen et al., 2021b; Min et al., 2021; 2022; Brown et al., 2020) and requires full data sharing. Instead, in this work, we focus on methods that can operate within API-based constraints.

Under R1, R2, and R3 requirements, the standard inductive learning (Liu et al., 2022) may be quite limiting. To mitigate the labeled data scarcity while retaining API compliance, we revisit transduction (Vapnik, 1999) in the context of textual few-shot classification. Specifically, in the context of few-shot learning, transductive few-shot learning (Liu et al., 2019a) advocates leveraging unlabeled test samples of a task as an additional source of information on the underlying task's data distribution in order to better define decision boundaries. Such additional source essentially comes for free in many *offline* applications, including sentiment analysis for customer feedback, legal document classification, or text-based medical diagnosis.

Our findings corroborate recent findings in computer vision (Liu et al., 2019a; Ziko et al., 2020; Lichtenstein et al., 2020; Boudiaf et al., 2020; Hu et al., 2021b), that substantial gains can be obtained from using transduction over induction, opening new avenue of research for the NLP community. However, the transductive gain comes at the cost of introducing additional hyperparameters, and carefully tuning them. Motivated by Occam's razor principle, we propose a novel hyperparameter-free transductive regularizer based on Fisher-Rao distances and demonstrate the strongest predictive performances across various benchmarks and models while keeping hyper-parameter tuning minimal. We believe that this parameter-free transductive regularizer can serve as a baseline for future research.

Contributions

In this paper, we make several contributions to the field of textual few-shot learning. Precisely, our contributions are threefold:

• **A new textual few-shot scenario:** We present a new scenario for few-shot learning using textual API-based models that accurately captures real-world constraints. Our novel scenario opens up new research avenues and opportunities to address the challenges associated with few-shot learning using API-based models, paving the way for improved performance and practical applications in the field.

• **A novel transductive baseline.** Our paper proposes a transductive few-shot learning algorithm that utilizes a novel parameter-free Fisher-Rao based loss. By leveraging only the network's embedding **(R1)**, our approach enables fast and efficient predictions **(R2)** without the need to share the labeling schema or the labels of few-shot examples making it compliant with **(R3)**. This innovative method marks a significant step forward in the field of few-shot learning, offering improved performance and practicality for real-world applications.

• **A truly improved experimental setting.** Previous studies on textual few-shot classification (Schick & Schütze, 2022; 2020b; Mahabadi et al., 2022; Tam et al., 2021; Gao et al., 2020) have predominantly assessed their algorithms on classification tasks with a restricted number of labels (typically less than five). We take a step forward and create a benchmark that is more representative of real-world scenarios. Our benchmark relies on a total of eight datasets, covering multiclass classification tasks with up to 151 classes, across four different languages. Moreover, we further enhanced the evaluation process by not only considering 10 classifiers trained with 10 different seeds (Logan IV et al., 2021; Mahabadi et al., 2022), but also by relying on episodic evaluation on 1,000 episodes (Hospedales et al., 2021). Our results clearly demonstrate the superiority of transductive methods.

2 Related Work

2.1 Few-shot learning in Natural Language Processing

Numerous studies have tackled the task of few-shot learning in Natural Language Processing (NLP) by utilizing pre-trained language models (Devlin et al., 2018; Liu et al., 2019b; Radford et al., 2019; Yang et al., 2019). Theses methods can be classified into three major categories: prompt-based, parameter-efficient tuning and in-context learning.

Prompt-based Few-shot Learning: Prompt-based few-shot learning involves the use of natural language prompts or templates to guide the model to perform a specific task (Ding et al., 2021; Liu et al., 2023). For example, the seminal work (Schick & Schütze, 2020a) proposed a model called PET, which uses a pre-defined set of prompts to perform various NLP tasks as text classification. They also impose a choice of a verbalizer which highly impact the classification performances (Cui et al., 2022; Hu et al., 2021a). However, recent studies have questioned the benefits of prompt-based learning due to the high variability in performance caused by the choice of prompt (Liu et al., 2022). To address this issue, researchers have proposed prompt tuning which involves a few learnable parameters in addition to the prompt (Lester et al., 2021). Nevertheless, these approaches face limitations when learning from API: (i) encoder access for gradient computation is infeasible (as in **R1**), (ii) prompting requires to send data and label which raises privacy concerns (as in **R3**), and (iii) labeling new points is time-consuming (see in **R3**) and expensive due to the need to send all shots for each input token[1].

Parameter-efficient fine-tuning. These methods, such as adapters (Houlsby et al., 2019; Pfeiffer et al., 2020), keep most of the model's parameters fixed during training and only update small feed-forward networks that are inserted within the larger model architecture. A recent example is T-FEW (Liu et al., 2022), which adds learned vectors that rescale the network's internal activations. Additionally, it requires a set of manually created prompts for each dataset making it hard to use in practice. Relying on parameter-efficient fine-tuning methods with an API is not possible due to the need to compute gradients of the encoder (as per **R1**) and the requirement to send both the labeling schema and the labels, which violates **R3**.

In Context Learning. In-context learning models are a unique type of model that utilizes input-to-output training examples as prompts to make predictions, without any parameter updates Wei et al. (2022). These models, such as GPT-3 and ChatGPT, rely solely on the provided examples to generate predictions, without any additional training. However, a significant drawback of this approach is that the user must supply the

[1]The cost of API queries is determined by the number of input tokens that are transmitted.

input, label examples, and task description, which is both slow (Liu et al., 2022) (**R2**) and raises data privacy concerns (as highlighted in **R3**). Additionally, the inability to reuse text embeddings for new tasks or with new labels without querying the model's API limits practicality and scalability, making reusable encoding unfeasible for in-context learning models[2].

Meta-learning. Meta-learning approaches have for quite long stood as the *de-facto* paradigm for few-shot learning (Snell et al. (2017); Rusu et al. (2019); Sung et al. (2018b); Lee et al. (2019); Raghu et al. (2019); Sun et al. (2019a)). In meta-learning, the objective is to provide the model with the intrinsic ability to learn in a data-efficient manner. For instance, MAML (Finn et al. (2017b); Antoniou et al. (2018)), arguably the most popular meta-learning method, tries to train a model such that it can be fine-tuned end-to-end using only a few supervised samples while retaining high generalization ability. Unlike the three previous lines of work, meta-learning methods operate by modifying the pre-training procedure and therefore assume access to both the training data and the model, which wholly breaks both **R1** and **R3**.

2.2 Inductive vs transductive few-shot learning

Learning an inductive classifier on embeddings generated by an API-based model, as proposed by (Snell et al., 2017), is a common baseline for performing few-shot learning. This approach is prevalent in NLP, where a parametric model is trained on data to infer general rules that are applied to label new, unseen data (known as inductive learning (Vapnik, 1999)). However, in few-shot learning scenarios with limited labeled data, this approach can be highly ambiguous and lead to poor generalization.

Transduction offers an attractive alternative to inductive learning (Sain, 1996). Unlike inductive learning, which infers general rules from training data, transduction involves finding rules that work specifically for the unlabeled test data. By utilizing more data, such as unlabeled test instances, and aiming for a more localized rule rather than a general one, transductive learning has shown promise and practical benefits in computer vision (Boudiaf et al., 2020; 2021; Ziko et al., 2020). Transductive methods yield substantially better performance than their inductive counterparts by leveraging the statistics of the query set (Dhillon et al., 2019). However, this approach has not yet been explored in the context of textual data.

3 API based Few-shot Learning

3.1 Problem Statement

Let Ω be the considered vocabulary, we denote Ω^* its Kleene closure. The Kleene closure corresponds to sequences of arbitrary size written with tokens in Ω, *i.e.*, $\Omega^* = \bigcup_{i=0}^{\infty} \Omega^i$. Given an input space $\mathcal{X}$ with $\mathcal{X} \subseteq \Omega^*$ and a latent space $\mathcal{Z}$, we consider a pre-trained backbone model $f_\theta : \mathcal{X} \to \mathcal{Z} = \mathcal{R}^d$, where $\theta \in \Theta$ represents the parameters of the encoder and d is the embedding dimension size. In the API-based setting, we assume that we are unable to access the exact structure of f_θ as mentioned in **R1**. However, we do have access to the last embedding of the encoder which is available for our use (see **R1**).

The objective of few-shot classification is to learn a classifier from limited labeled data and generalize to new, unseen tasks or classes. To accomplish this, randomly sampled few-shot tasks are created from a test dataset $\mathcal{D}_{test} := \{(x_i, y_i)\}_{i=1}^{N_{test}}$ that has a set of unseen classes $\mathcal{Y}_{test}$. Each task involves a few labeled examples from K different classes chosen at random among $\mathcal{Y}_{test}$. These labeled examples constitute the support set $S = \{x_i, y_i\}_{i \in \mathcal{I}_S}$, with a size of $|S| = N_S \times K$. Additionally, each task has an unlabeled query set $Q = \{x_i\}_{i \in \mathcal{I}_Q}$ composed of $|Q| = N_Q \times K$ unseen examples from each of the K classes. Pre-trained models use few-shot techniques and the labeled support sets to adapt to the tasks at hand and are evaluated based on their performances on the unlabeled query sets.

Remark Setting the values of N and K in textual few-shot learning is not standardized, as discussed in Sec. 3.1. Therefore, in all of our experiments, we have relied on setting $(N, K) \in \{5, 10\}^2$.

[2]Furthermore, as the number of considered classes increases, the fixed size of the transformer limits the number of possible shots that can be fed to the model. Previous studies have often neglected this limitation by focusing on a few number of labels.

3.2 Proposed Methods and Transductive approaches

NLP few-shot classifiers rely only on inductive inference, while computer vision has shown significant performance improvements using transductive inference for few-shot learning. Transductive inference succeeds in few-shot learning because it jointly classifies all unlabeled query samples of a single task, leading to more efficient and accurate classification compared to inductive methods that classify one sample at a time. Let us begin by introducing some basic notation and definitions before introducing our new transductive loss based on the Fisher-Rao distance.

In the API-based few-shot classification setting, our goal is to train a classification head $g_\phi : \mathcal{Z} \to \mathbb{R}^K$ that maps the feature representations to the posterior distribution space for making predictions. To simplify the equations for the rest of the paper, we use the following notations for the posterior predictions of each $i \in \mathcal{I}_S \cup \mathcal{I}_Q$ and for the class marginals within Q:

$$p_{ik} = g_\phi(f_\theta(x_i))_k = \mathbb{P}(Y = k|X = x_i; \theta, \phi) \text{ and } \widehat{p}_k = \frac{1}{|Q|} \sum_{x_i \in Q} p_{ik} = \mathbb{P}(Y_Q = k; \theta, \phi)$$

where X and Y are the random variables associated with the raw features and labels, respectively, and where Y_Q means restriction of the random variable Y to set Q.

For training the classification head in the transductive setting, prior research aims at finding ϕ such that $\phi = \arg\min \text{CE} - \lambda \times R_Q$[3], with $\text{CE} := -\frac{1}{|S|} \sum_{i \in S} \sum_{k=1}^{K} y_{ik} \log(p_{ik})$ being the cross-entropy supervision on the support set (in which y_{ik} is the k^{th} coordinate of the one-hot encoded label vector associated to sample i) and R_Q being a transductive loss on the query set Q.

Note that this transductive regularization has been proposed in the literature based on the InfoMax principle (Cardoso, 1997; Linsker, 1988) and the inductive loss can be found by setting $\lambda = 0$. In what follows, we review the regularizers introduced in previous work.

Entropic Minimization (H) An effective regularizer for transductive few-shot learning can be derived from the field of semi-supervised learning, drawing inspiration from the approach introduced in (Grandvalet & Bengio, 2004). This regularizer, proposed in (Dhillon et al., 2019), utilizes the conditional Shannon Entropy (Cover, 1999) of forecast results from query samples during testing to enhance model generalization. Formally:

$$R_Q^H = \frac{1}{|Q|} \sum_{i \in Q} \sum_{k=1}^{K} p_{ik} \log(p_{ik}). \tag{1}$$

Mutual Information Maximization (I) A promising alternative to the entropic minimization for addressing the challenges of transductive few-shot learning is to adopt the Info-max principle. (Boudiaf et al., 2020) extended this idea, introduced in (Hu et al., 2017), and propose as regularizer a surrogate of the mutual-information $R_Q^I(\alpha)$:

$$R_Q^I(\alpha) := -\sum_{k=1}^{K} \hat{p}_k \log \hat{p}_k + \alpha \frac{1}{|Q|} \sum_{i \in Q} \sum_{k=1}^{K} p_{ik} \log(p_{ik}). \tag{2}$$

Limitation of existing strategies: Despite its effectiveness, the previous method has a few limitations that should be taken into account. One of these limitations is the need to fine-tune the weight of different entropies using the hyperparameter α. This parameter tuning process can be time-consuming and may require extensive experimentation to achieve optimal results. Additionally, recent studies have shown that relying solely on the first Entropic term, which corresponds to the Entropic minimization scenario in Equation 1, can lead to suboptimal performance in few-shot learning.

[3] λ is set to 1 in all the experiements.

3.3 A Fisher-Rao Based Regularizer

In the few-shot learning scenario, minimizing parameter tuning is crucial. Motivated by this, in this section we introduce a new parameter-free transductive regularizer which fits into the InfoMax framework. Additionally, our loss inherits the attractive properties of the recently introduced Fisher-Rao distance between soft-predictions $\mathbf{q} := (q_1, ;q_K)$ and $\mathbf{p} := (p_1, ;p_K)$, which is given by (Picot et al., 2023):

$$d_{\mathrm{FR}}(\mathbf{q}, \mathbf{p}) := 2 \arccos \left(\sum_{k=1}^{K} \sqrt{q_k \times p_k} \right). \tag{3}$$

The proposed transductive regularizer denoted by R_Q^{FR}, for each single few-shot task, can be described as measuring the Fisher-Rao distance between pairs of query samples:

$$R_Q^{\mathrm{FR}} := \frac{1}{|Q|} \sum_{i \in Q} - \log \sum_{j \in Q} \sum_{k=1}^{K} \sqrt{p_{ik} \times p_{jk}} = \frac{1}{|Q|} \sum_{i \in Q} - \log \sum_{j \in Q} \cos \left(\frac{d_{\mathrm{FR}}(\mathbf{p}_i, \mathbf{p}_j)}{2} \right), \tag{4}$$

where $d_{\mathrm{FR}}(\mathbf{p}_i, \mathbf{p}_j)$ is the Fisher-Rao distance between pairs of soft-predictions $(\mathbf{p}_i, \mathbf{p}_j)$. Furthermore, it is shown that expression (4) yields a surrogate of the Mutual Information as shown by the following proposition. This result to the best of our knowledge is new, as far as we can tell.

Proposition 1 *(Fisher-Rao as a surrogate to maximize Mutual Information) Let $(\mathbf{q}_i)_{i \in Q}$ be a collection of soft-predictions corresponding to the query samples. Then, it holds that:*

$$R_Q^{FR} + \log |Q| \leq R_Q^I(1) \leq R_Q^I(\alpha), \quad \forall\ 0 \leq \alpha \leq 1. \tag{5}$$

Proof: Further details are relegated to Ap. A.

Advantage of R_Q^{FR} over $R_Q^I(\alpha)$: Similarly to $R_Q^I(\alpha)$, R_Q^{FR} can be exploited to maximize the Mutual Information. However, R_Q^{FR} is parameter free and thus, it does not require to tune α.

3.4 Additional Few-shot Inductive Baseline

In addition to the transductive methods of Sec. 3.2, we will explore two additional inductive methods for few-shot classification: prototypical networks and linear probing.

Prototypical Networks (PT) Prototypical Networks learn a metric space where the distance between two points corresponds to their degree of similarity. During inference, the distance between the query example and each class prototype is computed, and the predicted label is the class with the closest prototype. Prototypical networks have been widely used in NLP and are considered as a strong baseline (Snell et al., 2017; Sun et al., 2019b; Gao et al., 2019).

Linear Probing (CE) Fine-tuning a linear head on top of a pretrained model is a popular approach to learn a classifier for various classification tasks and was originally propose in (Devlin et al., 2018).

4 An Enhanced Experimental Setting

4.1 Datasets

Benchmarking the performance of few-shot learning methods on diverse set of datasets is critical to evaluate their generalization capabilities in a robust manner as well as their potential on real-world applications. Previous work on few-shot learning (Karimi Mahabadi et al., 2022; Perez et al., 2021) mainly focuses on datasets with a reduced number of classes (*i.e.*, $K < 5$). Motivated by practical considerations we choose to build a new benchmark composed of datasets with a larger number of classes.

Dataset	Classes (K)
Tweet Eval (Tweet)	20
Go Emotion (Emotion)	25
Amazon Review (Amazon)	30
Banking (B77)	77
Clinc	151

Table 1: Statistics of the considered datasets.

Specifically, we choose Go Emotion (Demszky et al., 2020), Tweet Eval (Barbieri et al., 2020), Clinc (Larson et al., 2019), Banking (Casanueva et al., 2020) and the Multilingual Amazon Reviews Corpus (Keung et al., 2020). These datasets cover a wide range of text classification scenarios and are of various difficulty[4]. A summary of the datasets used can be found in Tab. 1.

4.2 Model Choice

The selection of an appropriate backbone model is a critical factor in achieving high performance in few-shot NLP tasks. To ensure the validity and robustness of our findings, we have included a diverse range of transformer-based backbone models in our study, including:

- Three different sizes of RoBERTa based models (Liu et al., 2019b). Similar to BERT, RoBERTa is pretrained using the closed task (Taylor, 1953). We consider two different sizes of the RoBERTa model, namely RoBERTa (B) with 124M parameters and RoBERTa (L) with 355M parameters and DistilRoBERTa, a lighter version of RoBERTa trained through a distillation process (Hinton et al., 2015), for a total of 82M parameters.

- Three sentence-transformers encoder (Reimers & Gurevych, 2019). Following the recommendation of (Muennighoff et al., 2022), we consider MPNET-base (Song et al., 2020) (109M parameters), MiniLM (33M parameters) (Wang et al., 2020), and Albert Small V2 (11M parameters) (Lan et al., 2019).

- Multilingual models. To address realistic scenarios, we do not restrict our study to the English language. We rely on three sizes of XLM-RoBERTa (Conneau et al., 2020; 2019): base (B) with 124M, large with 355M (L) and XL (XL) with 3.5B of parameters.

- GPT-3 model: to mimic the typical setting of API-based models, we also conduct experiments on GPT-3 (Brown et al., 2020), only accessible through OpenAI's API.

Preliminary Experiment. In our experiments, the backbone models are of utmost importance. Our objective in this preliminary experiment is to assess the efficacy of these models when fine-tuning **only** the model head across a variety of datasets. Through this evaluation, we aim to gain insight into their generalization abilities and any dataset-specific factors that may influence their performance. This information will be utilized to analyze the performance of different models in the few-shot scenario, as described in Sec. 5. We present the results of this experiment in Tab. 2, noting that all classes were considered, which differs from the episodic training approach detailed in Sec. 5.

Model	Params	Emotion	Twitter	Clinic	Banking	Amazon			
		en	en	en	en	en	fr	es	de
Albert Small V2 (XS)	11M	25.2	18.3	67.0	88.1	33.5	X	X	X
MiniLM (S)	33M	30.2	19.3	67.1	92.3	39.5	X	X	X
MPNET-base (B)	109M	30.2	22.5	67.4	94.3	41.3	X	X	X
DistilRoBERTa (S)	82M	23.3	26.0	68.5	90.9	40.0	X	X	X
RoBERTa (B)	124M	21.0	25.5	66.7	91.4	39.2	X	X	X
RoBERTa (L)	355M	15.0	23.0	64.5	90.0	38.1	X	X	X
XLM-RoBERTa (B)	278M	21.0	22.1	66.5	87.0	40.1	19.2	17.5	18.3
XLM-RoBERTa (L)	559M	14.0	18.0	64.5	86.2	38.2	17.5	15.6	18.1
XLM-RoBERTa (XL)	3.48B	25.4	19.0	68.9	95.0	41.0	18.9	17.9	22.0
GPT-3.5	175B	38.9	35.3	70.4	98.7	48.4	30.4	34.0	33.5

Table 2: Preliminary experiment results. Accuracy of the different backbone trained on each training set.

[4]These datasets are available in Dataset (Lhoest et al., 2021)

4.3 Evaluation Framework

Prior research in textual few-shot learning typically involves sampling a low number of tasks, typically less than 10, of each dataset. In contrast, we utilize an episodic learning framework that generates a large number of N-ways K-shots tasks. This framework has gained popularity through inductive meta-learning approaches, such as those proposed by (Finn et al., 2017a; Snell et al., 2017; Vinyals et al., 2016; Sung et al., 2018a; Mishra et al., 2017; Rusu et al., 2019; Oreshkin et al., 2018), as it mimics the few-shot environment during evaluation and improves model robustness and generalization. In this context, episodic training implies that a different model is initialized for each generated few-shot task, and all tasks are compiled independently in parallel. This approach allows to compute more reliable performance statistics by evaluating the generalization capabilities of each method on a more diverse set of tasks. To account for the model's generalization ability, we average the results for each dataset over 1000 episodes, with the N considered classes varying in every episode. For each experiment, we consider the F1 Score.

5 Experiments

5.1 Overall Results

Global results: To evaluate the effectiveness of various few-shot methods, we conducted a comprehensive analysis of their classification performance across all datasets, all backbones, and all considered N-way/K-shot scenarios. Results are reported in Tab. 3.
An interesting observation is that transductive approaches I and FR outperform their inductive counterparts (CE and PT). Notably, we found that vanilla entropy minimization, which solely relies on H, consistently underperforms in all considered scenarios. Our analysis revealed that FR surpasses traditional fine-tuning based on cross-entropy by a margin of 3.7%.

K-shots	10		5	
N-ways	10	5	10	5
FR	**52.09**	**61.99**	**48.71**	**56.55**
I	50.07	59.17	46.42	55.74
H	15.07	27.39	15.33	25.84
CE	48.31	56.87	45.27	53.94
PT	47.29	56.05	44.32	53.20

Table 3: Aggregated performance over K,N, the different datasets and considered backbone.

Mono-lingual experiment: In order to thoroughly analyze the performance of each method, we conducted a per-dataset study, beginning with a focus on the mono-lingual datasets. Fig. 2 reveals that the global trends observed in Tab. 3 remain consistent across datasets of varying difficulty levels. Notably, we observed consistent improvements achieved by transductive regularizers (such as I or FR) over CE. However, the relative improvement is highly dependent on the specific dataset being evaluated. Specifically, FR achieves +6.5% F1-score on Banking, but only a shy +1.5% on Tweet. A strong baseline generally suggests highly discriminative features for the task, and therefore a strong upside in leveraging additional unlabeled features, and vice versa. Therefore, we hypothesize that the potential gains to be obtained through transduction correlate with the baseline's performance.

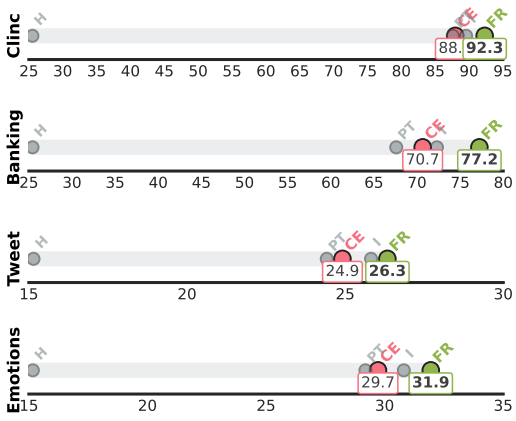

Figure 2: Performance of different pretrained encoder on the monolingual datasets.

Additional results can be found on Sec. B.2 multilingual experiments (*i.e.*, on es, de, fr) which exhibit the same behavior.

5.2 Study Under Different Data-Regime

In this experiment, we investigated the performance of different loss functions under varying conditions of 'ways' and 'shots'. As shown in Fig. 3, we observed that increasing the number of classes ('ways') led to a decrease in F1 while increasing the number of examples per class ('shots') led to an improvement in F1.

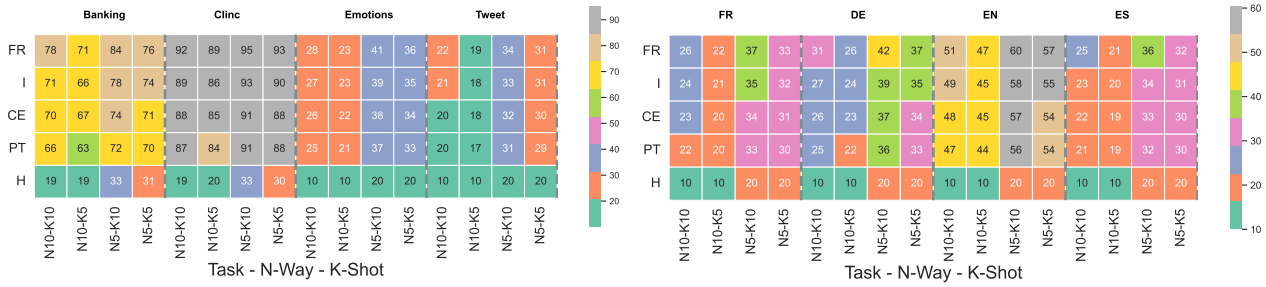

Figure 3: The effect of different ways and shots on test performance. Monolingual experiments are shown on the left, and multilingual experiments on the right.

This can be explained by the fact that having more data enables the classifier to better discern the unique characteristics of each class.

Interestingly, the relationship between the number of shots and classification F1 may not be the same for all classes or all loss functions. Fig. 3 shows that different loss functions (e.g. FR on banking) benefited greatly from adding a few shots, while others did not show as much improvement. However, this variability is dependent on the specific dataset and language being used, as different classes may have different levels of complexity and variability, and some may be inherently easier or harder to classify than others.

5.3 Ablation Study On Backbones

In this experiment, we examined how different loss functions perform when increasing the number of parameters in various models. The results, presented in Fig. 4, show the average performance across the experiments and are organized by loss function. We observed an *inverse scaling law* for both the RoBERTa and XLM-RoBERTa family of models, where increasing the number of parameters led to a decrease in performance for the losses tested. However, within the same family, we observe that the superiority of FR remains consistent. An interesting finding from Fig. 4 is that the transductive regularization technique using FR outperforms other methods on GPT-3.5. This highlights the effectiveness of FR in improving the performance of the model and suggests that transductive regularization may be a promising approach for optimizing language models.

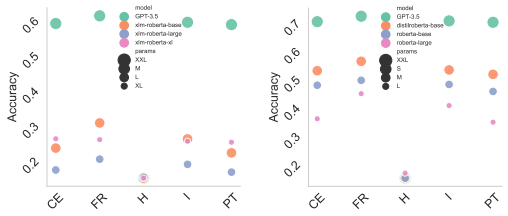

Figure 4: Impact of model size.

5.4 Practical Considerations

In this experiment, we adopt a practical standpoint and aim to evaluate the effectiveness of an API model, specifically GPT-3.5. In Sec. 5.4, we report the training speed of one episode on a MAC with CPU. Overall, we observed that the transductive loss is slower as it necessitates the computation of the loss on the query set, whereas PT is faster as it does not involve any optimization. Furthermore, we note that FR is comparable in speed to I. To provide a better understanding of these results, we can compare our method with existing approaches (in the light of **R2**). For instance, PET (Schick & Schütze, 2020a) entails a training time of 20 minutes on A100, while ADAPET (Tam et al., 2021) necessitates 10 minutes on the same hardware.

Loss	CPU Time
CE	0.45s
FR	0.83s
H	0.75s
I	0.83s
PT	0.01s

Table 4: Training time for 1 episode on a M1-CPU.

6 Conclusions

This paper presents a novel few-shot learning framework that utilizes API models while meeting critical constraints of real world applications (i.e., **R1**, **R2**, **R3**). This approach is particularly appealing as it shifts the computational requirements (**R2**), eliminating the need for heavy computations for the user. This opens up new possibilities, such as training classifiers on-the-fly in web browsers without sharing labels of the data (**R3**). Furthermore, the use of an API setting is highly advantageous as it significantly reduces the cost of embedding. To provide a better understanding, embedding over 400k sequences cost as low as 7 dollars. In this scenario, our research highlights the potential of transductive losses, which have previously been disregarded by the NLP community. A candidate loss is the Fisher-Rao distance which is parameter-free and could serve as a simple baseline in the future.

Broader Impact Statement
We are optimistic that our research will have a positive impact on society. Nonetheless, it is essential to acknowledge the limitations of API-based few-shot classification models despite their promising results in various tasks. Firstly, the performance of the introduced methods is heavily dependent on the quality of available API models. If the API models do not provide sufficient information or lack diversity, the introduced methods may struggle to accurately classify input texts. Secondly, the black-box nature of the backbone limits the interpretability of API-based few-shot classification methods, which may hinder their adoption. Ultimately, the aim of this work is to establish a baseline for future research on transductive inference. As a result, not all existing transductive methods are compared in this study.

References

Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020.

Malik Boudiaf, Imtiaz Ziko, Jérôme Rony, José Dolz, Pablo Piantanida, and Ismail Ben Ayed. Information maximization for few-shot learning. *Advances in Neural Information Processing Systems*, 33:2445–2457, 2020.

Malik Boudiaf, Hoel Kervadec, Ziko Imtiaz Masud, Pablo Piantanida, Ismail Ben Ayed, and Jose Dolz. Few-shot segmentation without meta-learning: A good transductive inference is all you need? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13979–13988, 2021.

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

J-F Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal processing letters*, 4(4):112–114, 1997.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 2020.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021a.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. *arXiv preprint arXiv:2110.07814*, 2021b.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022.

European Commission. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), 2016. URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

European Commission. Proposal for a regulation of the european parliament and of the council on european data governance (data governance act), 11 2020. URL https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0767&from=EN. COM(2020) 767 final.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Ganqu Cui, Shengding Hu, Ning Ding, Longtao Huang, and Zhiyuan Liu. Prototypical verbalizer for prompt-based few-shot tuning. *arXiv preprint arXiv:2203.09770*, 2022.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A Dataset of Fine-Grained Emotions. In *58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Guneet S Dhillon, Pratik Chaudhari, Avinash Ravichandran, and Stefano Soatto. A baseline for few-shot image classification. *arXiv preprint arXiv:1909.02729*, 2019.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. Openprompt: An open-source framework for prompt-learning. *arXiv preprint arXiv:2111.01998*, 2021.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. Beyond english-centric multilingual machine translation. *The Journal of Machine Learning Research*, 22(1):4839–4886, 2021.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, 2017a.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017b.

Tianyu Gao, Xu Han, Zhiyuan Liu, and Maosong Sun. Hybrid attention-based prototypical networks for noisy few-shot relation classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6407–6414, 2019.

Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*, 2020.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *ArXiv*, 2015.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Shengding Hu, Ning Ding, Huadong Wang, Zhiyuan Liu, Juanzi Li, and Maosong Sun. Knowledgeable prompt-tuning: Incorporating knowledge into prompt verbalizer for text classification. *arXiv preprint arXiv:2108.02035*, 2021a.

Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, 2017.

Yuqing Hu, Vincent Gripon, and Stéphane Pateux. Leveraging the feature distribution in transfer-based few-shot learning. In *Artificial Neural Networks and Machine Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part II 30*, pp. 487–499. Springer, 2021b.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Marzieh Saeidi, Lambert Mathias, Veselin Stoyano, and Majid Yazdani. Perfect: Prompt-free and efficient few-shot learning with language models. In *Annual Meeting of the Association for Computational Linguistics*, 2022.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. The multilingual Amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. Chatgpt: Jack of all trades, master of none, 2023.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019. URL https://www.aclweb.org/anthology/D19-1131.

Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10657–10665, 2019.

Eric Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary Ziegler, Daniel Nadler, Peter Szolovits, Alistair Johnson, and Emily Alsentzer. Do we still need clinical language models?, 2023.

Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. *arXiv preprint arXiv:2109.02846*, 2021.

Moshe Lichtenstein, Prasanna Sattigeri, Rogerio Feris, Raja Giryes, and Leonid Karlinsky. Tafssl: Task-adaptive feature sub-space learning for few-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII*, pp. 522–539. Springer, 2020.

Ralph Linsker. Self-organization in a perceptual network. *Computer*, 21(3):105–117, 1988.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint arXiv:2205.05638*, 2022.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023.

Yanbin Liu, Juho Lee, Minseop Park, Saehoon Kim, Eunho Yang, Sung Ju Hwang, and Yi Yang. Learning to propagate labels: Transductive propagation network for few-shot learning. *ICLR*, 2019a.

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

Robert L Logan IV, Ivana Balažević, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. Cutting down on prompts and parameters: Simple few-shot learning with language models. *arXiv preprint arXiv:2106.13353*, 2021.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Marzieh Saeidi, Lambert Mathias, Veselin Stoyanov, and Majid Yazdani. Perfect: Prompt-free and efficient few-shot learning with language models. *arXiv preprint arXiv:2204.01172*, 2022.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and P. Abbeel. A simple neural attentive meta-learner. In *International Conference on Learning Representations*, 2017.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

OpenAI. Gpt-4 technical report, 2023.

Boris N. Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: Task dependent adaptive metric for improved few-shot learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 719–729, 2018.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070, 2021.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020.

Marine Picot, Francisco Messina, Malik Boudiaf, Fabrice Labeau, Ismail Ben Ayed, and Pablo Piantanida. Adversarial robustness via fisher-rao regularization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2698–2710, 2023. doi: 10.1109/TPAMI.2022.3174724.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *arXiv preprint arXiv:1909.09157*, 2019.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Jonathan S Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. A constructive prediction of the generalization error across scales. *arXiv preprint arXiv:1909.12673*, 2019.

Andrei Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization, 2019.

Stephan R Sain. The nature of statistical learning theory, 1996.

Timo Schick and Hinrich Schütze. Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*, 2020a.

Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. *arXiv preprint arXiv:2009.07118*, 2020b.

Timo Schick and Hinrich Schütze. True few-shot learning with prompts—a real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731, 2022.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Irene Solaiman. The gradient of generative ai release: Methods and considerations. *arXiv preprint arXiv:2302.04844*, 2023.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 2022.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp, 2019.

Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 403–412, 2019a.

Shengli Sun, Qingfeng Sun, Kevin Zhou, and Tengchao Lv. Hierarchical attention prototypical networks for few-shot text classification. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 476–485, 2019b.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip Torr, and Timothy Hospedales. Learning to compare: Relation network for few-shot learning. 2018a. doi: 10.1109/CVPR.2018.00131.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1199–1208, 2018b.

Derek Tam, Rakesh R Menon, Mohit Bansal, Shashank Srivastava, and Colin Raffel. Improving and simplifying pattern exploiting training. *arXiv preprint arXiv:2103.11955*, 2021.

Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4): 415–433, 1953.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10 (5):988–999, 1999.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, 2016.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Qinyuan Ye, Bill Yuchen Lin, and Xiang Ren. Crossfit: A few-shot learning challenge for cross-task generalization in nlp. *arXiv preprint arXiv:2104.08835*, 2021.

Imtiaz Ziko, Jose Dolz, Eric Granger, and Ismail Ben Ayed. Laplacian regularized few-shot learning. In *International conference on machine learning*, pp. 11660–11670. PMLR, 2020.

A Proof of Proposition 1

In this Appendix, we prove the inequality (Eq. 5) provided in Proposition 1. The right-hand side of (Eq. 5) follows straightforwardly from the definition of $R_Q^I(\alpha)$ and the non-negativity of the Shannon entropy. In order to prove the first inequality, we need to introduce the following intermediate result.

For any arbitrary random variable X and countable random variable Y, and any real number β, let

$$I_\beta(X;Y) := -\mathbb{E}_{X^\star Y} \log \mathbb{E}_X \left[\frac{P(Y|X)}{P(Y|X^\star)} \right]^\beta,$$

where the random variable $X^\star$ follows the same distribution than X. Notice that it is obvious that $I_1(X;Y) = I(X;Y)$, where $I(X;Y)$ is Shannon Mutual Information.

Lemma 1 *For any arbitrary random variable X and countable random variable Y, we have*

$$I(X;Y) \geq I_\beta(X;Y), \quad \textit{for } 0 \leq \beta \leq 1.$$

Proof of the lemma: We must show that the different of $I(X;Y) - I_\beta(X;Y)$ is nonnegative. To this end, we write this difference as:

$$\begin{aligned}
I(X;Y) - I_\beta(X;Y) &= -\mathbb{E}_{X^\star Y} \log \frac{P^{1-\beta}(Y|X^\star)\mathbb{E}_X P(Y|X)}{\mathbb{E}_X P^\beta(Y|X)} && (6)\\
&\geq -\log \mathbb{E}_{X^\star Y} \frac{P^{1-\beta}(Y|X^\star)\mathbb{E}_X P(Y|X)}{\mathbb{E}_X P^\beta(Y|X)} && (7)\\
&= -\log \sum_{y\in\mathcal{Y}} \mathbb{E}_{X^\star} P(y|X^\star) \frac{P^{1-\beta}(y|X^\star)\mathbb{E}_X P(y|X)}{\mathbb{E}_X P^\beta(y|X)} && (8)\\
&= -\log \sum_{y\in\mathcal{Y}} \frac{\mathbb{E}_{X^\star} P^\beta(y|X^\star)\mathbb{E}_X P(y|X)}{\mathbb{E}_X P^\beta(y|X)} && (9)\\
&= -\log \sum_{y\in\mathcal{Y}} \mathbb{E}_X P(y|X) && (10)\\
&= 0, && (11)
\end{aligned}$$

where the first inequality follows by applying Jensen's inequality to the function $t \mapsto -\log(t)$.

Proof of Proposition 1: From Lemma 1, using Jensen's inequality, we have

$$\begin{aligned}
I(X;Y) &= -\mathbb{E}_{X^\star Y} \log \mathbb{E}_X \left[\frac{P(Y|X)}{P(Y|X^\star)} \right], && (12)\\
&\geq -\mathbb{E}_{X^\star Y} \log \mathbb{E}_X \left[\frac{P(Y|X)}{P(Y|X^\star)} \right]^\beta && (13)\\
&\geq -\mathbb{E}_{X^\star} \log \mathbb{E}_X \mathbb{E}_{Y|X^\star} \left[\frac{P(Y|X)}{P(Y|X^\star)} \right]^\beta && (14)\\
&= -\mathbb{E}_{X^\star} \log \mathbb{E}_X \sum_{y\in\mathcal{Y}} P^\beta(Y|X) P^{1-\beta}(Y|X^\star), && (15)
\end{aligned}$$

where inequality (13) follows by applying Lemma 1 and inequality (14) follows by exploiting the convexity of the function $t \mapsto -\log(t)$ for any $0 \leq \beta \leq 1$. Finally, it is not difficult to check from the definition of the Fisher-Rao distance given by expression (3) that

$$\cos\left(\frac{d_{\mathrm{FR}}(P(y|X=x), P(y|X=x^\star))}{2} \right) = \sum_{y\in\mathcal{Y}} \sqrt{P(y|X=x)P(y|X=x^\star)}. \quad (16)$$

Using the identity given by (16) in expression (15) setting $\beta = 1/2$, we obtain the desired inequality

$$I(X;Y) \geq -\mathbb{E}_{X^\star} \log \mathbb{E}_X \cos\left(\frac{d_{\mathrm{FR}}(P(y|X), P(y|X^\star))}{2}\right). \tag{17}$$

The inequality (5) immediately follows by replacing the distribution of the random variable X with the empirical distribution on the query and $P(y|x)$ with the soft-prediction corresponding to the feature x, which concludes the proof of the proposition.

B Additional Experimental Results

B.1 A Dive Into GPT-3.5 results

GPT-3.5 appears to be the backbone providing the most informative a priori embeddings in Tab. 2 and could be considered as the prime model for API-based Few-shot learning, showcasing the current requirements in this area. It is thus a typical candidate for application uses that must meet the following criteria **(R1)** - **(R3)**. Therefore, we put a special emphasis on its related results.

Fig. 5 (top) details the GPT-3.5 results of the experiments conducted on the mono-lingual datasets. These plots highlight the consistency of the tendencies emerged in Tab. 2, Tab. 3 and Fig. 2, namely: the superiority of transductive approaches (FR and I) over inductive ones (CE and PT), the underperformance of the entropic-minimization-based strategy (H), and the higher amount of information conveyed by GPT-3.5 learned embeddings over other backbones, resulting in higher F1 scores on all datasets.

These phenomena still occur in the multi-lingual setting, as illustrated in Fig. 5 (bottom), stressing the superiority of transductive (and especially FR) over other approaches for presumably universal tasks, beyond english-centered ones, and without the need of using language-specific engineering as for prompting-based strategies.

Note that for both of these settings, the entropic-minimization-based strategy (H) seems to be capped at a 15% F1 score, thus with no improvement over other backbones embeddings, and independently of the dataset difficulty.

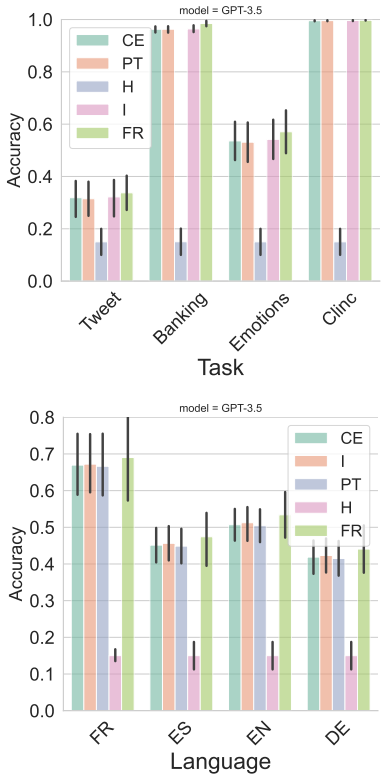

Figure 5: The different losses when training a on GPT3.5 embeddings.

B.2 Multilingual Experiment

To provide an exhaustive analysis, we report the same experiment that is made in Sec. B.2 for multi-lingual model on Amazon. The observations made in Sec. B.1 are not specific to GPT-3.5 backbone and extend to the other multi-lingual encoders (that is XLM-RoBERTa-based ones). While both latin languages (French and Spanish) share almost identical results, with a trend very similar to the one of English language (an F1 gain of around 4% for FR over CE), the results on German language exhibit an F1 increased by more than 6% when switching from inductive CE to transductive FR, flirting with performances obtained on English tasks.

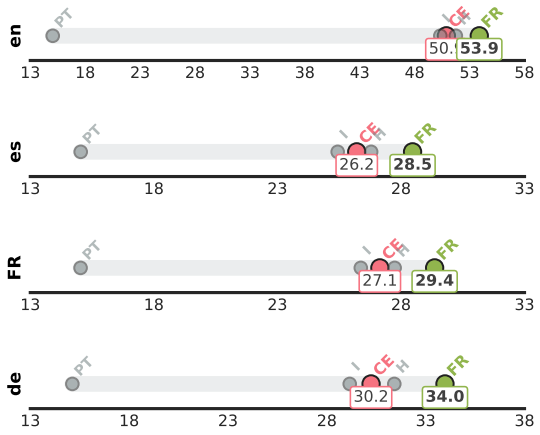

Figure 6: Performance of the different losses on multilingual datasets.

B.3 Importance of Model Backbones on Monolingual Experiment

In this section, we report the results of our experiment aggregated per backbone. The goal is to understand how the different losses behave on the different backbone. The results are presented in Fig. 8. While the trends observed in the previous charts are retrieved for the majority of backbones, some of these models are exceptions. For example, while transductive methods perform generally better than inductive methods, the CE-based method seems to perform slightly better than I for XLM-RoBERTa-xl. Additionally, while FR is the most effective method for the majority of backbones, it is surpassed by I for the all-distilroberta-v1 model. Furthermore, the inverse-scaling-law details are found for the RoBERTa(B/L) and XLM-RoBERTa (B/L) models per dataset. In general, it is interesting to note that although model performance is constrained by dataset difficulty, the performance order of each method is consistent across all 4 datasets for each considered backbone.

B.4 Importance of Model Backbones on Multilingual Experiment

In this experiment, we report the performance of different losses on the Amazon dataset by averaging the results over the number of shots, ways for the different losses. The results are presented in Fig. 10. Our observations indicate that the transductive regularization, both for I and FR, consistently improves the results for different models, including base and large models, as well as GPT-3.5. Similar to the findings reported in the main paper, we observe an inverse scaling law, with XLM-RoBERTa-base outperforming the larger versions.

B.4.1 Results Per Language

In this experiment, we report the performance of different losses on the Amazon dataset by averaging the results over the number of shots, ways, and model backbones. The results are presented in Tab. 5. Our observations indicate that the transductive regularization improves the results for two languages over the inductive baseline (i.e., CE). Additionally, we note that the observed improvements for FR are more consistent. This further demonstrates that the transductive loss can be useful in few-shot NLP.

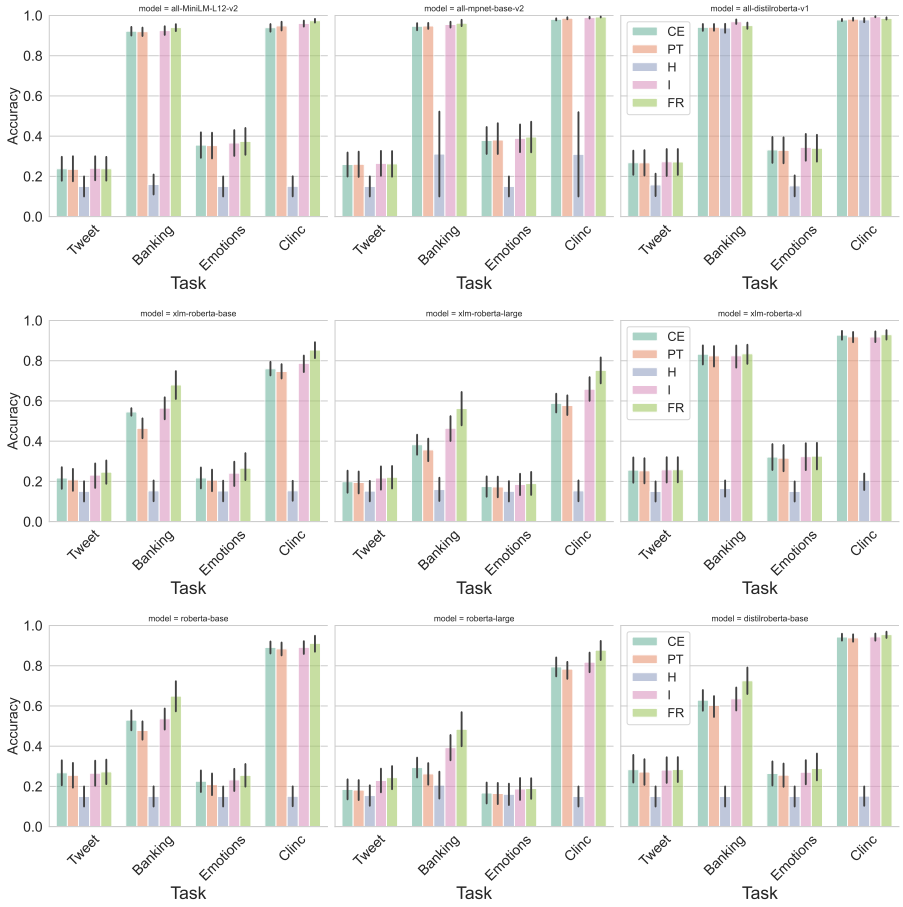

Figure 8: Performance of different pretrained encoder on the monolingual datasets.

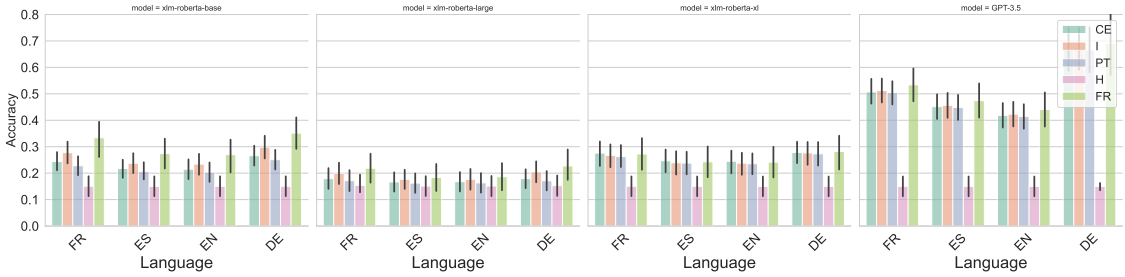

Figure 10: Performance of different pretrained backbone on multilingual Amazon.

	fr	de	en	es
FR	**29.36**	**33.98**	**53.89**	**28.47**
I	27.74	31.41	51.75	26.79
H	15.04	15.13	15.04	15.04
CE	27.15	30.24	50.89	26.21
PT	26.37	29.16	50.34	25.44

Table 5: Global Results for multilingual Amazon