A Practical Guide to Fine-tuning Language Models with Limited Data

Márton Szép[1,2*]   **Daniel Rueckert**[2,3,4]   **Rüdiger von Eisenhart-Rothe**[1]   **Florian Hinterwimmer**[1,2]

[1]Department of Orthopaedics and Sports Orthopaedics, TUM University Hospital, Germany
[2]Chair for AI in Healthcare and Medicine, Technical University of Munich (TUM)
and TUM University Hospital, Munich, Germany
[3]Munich Center for Machine Learning (MCML), Munich, Germany
[4]Department of Computing, Imperial College London, UK

Abstract

Employing pre-trained Large Language Models (LLMs) has become the de facto standard in Natural Language Processing (NLP) despite their extensive data requirements. Motivated by the recent surge in research focused on training LLMs with limited data, particularly in low-resource domains and languages, this paper surveys recent transfer learning approaches to optimize model performance in downstream tasks where data is scarce. We first address initial and continued pre-training strategies to better leverage prior knowledge in unseen domains and languages. We then examine how to maximize the utility of limited data during fine-tuning and few-shot learning. The final section takes a task-specific perspective, reviewing models and methods suited for different levels of data scarcity. Our goal is to provide practitioners with practical guidelines for overcoming the challenges posed by constrained data while also highlighting promising directions for future research.

1 Introduction

Pre-trained Language Models (PLMs) are transforming the field of NLP, showing outstanding capabilities to learn and model the underlying distributions of natural language data from complex and diverse domains (Han et al., 2021). Yet, their training demands extensive data and computational resources, which can be prohibitive in many real-world scenarios (Bai et al., 2024), particularly for languages other than English and for specialized domains such as, i.a., medicine (Crema et al., 2023; Van Veen et al., 2023), chemistry (Jablonka et al., 2024), law (Noguti et al., 2023), finance (Zhao et al., 2021), engineering (Beltagy et al.,

2019). The predominant approach to deal with this common issue relies on the transfer learning paradigm, which involves a self-supervised pre-training phase on vast amounts of general- or mixed-domain data, followed by (possibly multiple) domain adaptation and fine-tuning or few-shot learning steps on domain- and task-specific data. Notably, the second stage of this process is also data-hungry. Data scarcity can lead to overfitting, poor generalization, and suboptimal performance. Fine-tuning PLMs with limited data requires careful selection of pre-training strategies, domain adaptation, and efficient parameter optimization to achieve optimal performance by leveraging the model's pre-existing knowledge effectively to avoid catastrophic forgetting (Kirkpatrick et al., 2017; Ramasesh et al., 2021). This paper addresses the challenge of training LLMs with limited data, particularly for low-resource languages and specialized domains, by exploring recent advances in transfer learning (Table 1). We conduct a systematic review, starting with more than 2500 papers collected from Scopus, Web of Science, Google Scholar and ACL Anthology. It is intended for both researchers and practitioners in NLP, providing an overview of the current state-of-the-art methods and practical guidelines for optimizing model performance in data-scarce scenarios. We examine the process of adapting LLMs to specific tasks and domains with limited data, focusing on **(1)** selecting appropriate (continued) pre-training methods to leverage prior knowledge effectively in the low-resource scenario at hand (§ 3); **(2)** maximizing the utility of limited data during fine-tuning (§ 4) and few-shot learning (§ 5); **(3)** discussing the assumptions, benefits and limitations of various transfer learning strategies highlighting open challenges for researchers; and **(4)** task-specific perspectives on overcoming different levels of data scarcity as guidance for practitioners (§ 6).

---

* Corresponding author: `marton.szep@tum.de`

	Method	Requirements	Advantages	Limitations	Use Cases
Cont. Pre-train.	Cross-lingual alignment (§ 3.1)	Monolingual corpora	Low-resource monolingual data often available	Computationally intensive	Leverage high-resource representations for low-resource languages
	Domain adaptation (§ 3.1)	In-domain corpora	Some domain- and task-specific text data is always available	Computationally intensive	Leverage general-domain patterns for specialized domains
Fine-tuning	Parameter-efficient training (§ 4.1)	Low labeled data	Reduced resource usage, faster fine-tuning	Relies on the quality of pre-trained models	Applications with limited data and computation
	Embedding learning (§ 4.2)	Parallel or non-parallel data	Better transferability and semantic relationships	Static embeddings lack context	Discrepant pre-trained and target domain or language
	Contrastive learning (§ 4.3)	Paired data	Useful representations, data-efficiency	Sensitive to data pair quality and quantity	Cross-lingual alignment Classification tasks
	Adversarial learning (§ 4.3)	Unpaired data	Learns generalized features	Adversarial training is prone to be unstable	Cross-domain and cross-lingual transfer
	Semi-supervised learning (§ 4.4)	Labeled & unlabeled data	Utilizes unlabeled data, cost-effective	Sensitive to pseudo-label quality or noisy data	Classification and seq2seq low-resource tasks
	Unsupervised learning (§ 4.4)	Unpaired data	No need for labeled data, scalable	Often worse performance than supervised	Scenarios with no labeled data
	Active learning (§ 4.4)	Labeled & unlabeled data	Maximizes utility of labeled data	Requires iterative training and data selection	Classification and seq2seq low-resource tasks
Few-shot learning	In-context learning (§ 5.1)	Few labeled examples	Intuitive, flexible; No gradient-based training needed	Handcrafting; Sensitive to prompt quality; High inference cost	Any low-resource NLP task can be formulated as text generation
	Pattern-exploiting training (§ 5.2)	Few labeled examples	Better alignment of fine-tuning and pre-training task	Handcrafting; Slow autoregressive decoding	Few-shot natural language understanding tasks
	Multi-task learning (§ 5.3)	Large, diverse set of related task data	Strong generalization to unseen tasks	High computational and data requirements	General-purpose generative models for zero-shot tasks
	Meta-Learning (§ 5.4)	Few labeled examples in related tasks	Adapts quickly to new tasks	Computationally intensive, potentially complex to implement	Few-shot classification tasks

Table 1: Overview of the methods discussed in this paper.

2 Related work

Recent works have explored various strategies to improve the performance of PLMs in data-scarce scenarios, focusing on specific tasks, languages, or domains (Yazar et al., 2023; Krasadakis et al., 2024; Laparra et al., 2021). We provide a comprehensive overview of transfer learning methods for fine-tuning PLMs with limited data, covering various strategies and use cases. Data augmentation is, arguably, one of the most fundamental techniques to deal with data scarcity (Feng et al., 2021). However, data augmentation alone might be insufficient or impractical for specialized domains in low-resource languages due to limited diversity and quality degradation (Chen et al., 2023a; Stylianou et al., 2023). Thus, this paper focuses on the complementary model-centric approaches that make better use of available data. Treviso et al. (2023) surveys efficient NLP methods using a broader definition of efficiency in terms of resources, including data, time, storage, and energy. Hedderich et al. (2021) focus on low-resource, supervised NLP, including distant supervision and pre-training encoder models. In contrast, our work employs a more practical perspective, giving a structured overview of the data-efficient adaptation of pre-trained models in a computationally affordable manner, also considering task-specific aspects.

3 Pre-training

Pre-training serves as the initial, foundational training phase for LLMs, enabling them to develop

robust general and domain-specific language understanding in a self-supervised manner. This is one of the key success factors for LLMs to cope with a wide range of downstream tasks even with limited labeled data (Radford et al., 2018).

The first step is to choose a suitable model architecture (see § 6), which directly entails some options for pre-training objectives. This choice needs to align with the downstream task for better transferability. Decoder models are best suited for text generation tasks and use *causal language modeling* (CLM) objective (Brown et al., 2020), which involves predicting the next token in a sequence given the previous ones. In contrast, encoder models specialize in classification tasks and employ some form of *masked language modeling* (MLM) objective (Devlin et al., 2019), where a random subset of tokens is masked, and the model is trained to predict the original tokens. This objective is particularly useful for learning bidirectional representations. An improved variant, *replaced token detection* (RTD), shows better convergence by using a small generator network to replace input tokens (Clark et al., 2020). Conversely, encoder-decoder models excel in text transformation tasks and extend masking to sequences of tokens via masked or denoising sequence-to-sequence (S2S) pre-training (Song et al., 2019; Lewis et al., 2020).

3.1 Continued pre-training

Training a model from scratch often poses an insurmountable challenge in terms of data and computational resources. Fortunately, there is a steadily growing number of pre-trained models available for various tasks, languages, and domains[1]. Continued pre-training is a powerful technique that involves a limited number of training steps with the pre-training objective and in-domain or downstream task data without labels (Gururangan et al., 2020). The goal is to bridge the gap between the pre-training data and the target domain and language, leading to better performance on downstream tasks (Imani et al., 2023). In case of a considerable discrepancy between the original and target modalities, it allows us to extend the model vocabulary with previously unseen terms and phrases (Gnehm et al., 2022). Notably, subsequent pre-training enables sample-efficient reuse of labeled fine-tuning data in a self-

---

[1] https://huggingface.co/models

supervised fashion. In general, this kind of model adaptation is susceptible to catastrophic forgetting (Kalajdzievski, 2024), requiring a careful balance between training length, data size, and model size, benefiting from additional regularization (§ 4). Notable techniques include learning rate warm-up (McCloskey and Cohen, 1989) and experience replay (de Masson d' Autume et al., 2019).

For decoder models, the relation between model size and the amount of data for single-epoch pre-training has been extensively explored, showing enormous data requirements with approximately equal scaling of model size and data (Kaplan et al., 2020; Hoffmann et al., 2022). Yet, how these relations transfer to continued pre-training is still a challenging open question. In the data-scarce regime, Muennighoff et al. (2023a) find that training decoders up to 16 epochs can still achieve meaningful gains. Gnehm et al. (2022) confirm this for encoder models, showing that continued pre-training on a small in-domain corpus can benefit from longer training times. Besides continued pre-training of the entire model, an efficient alternative is to train only a subset of the model parameters (§ 4.1). Such methods allow a better balancing of model capacity and data size and aid in preserving prior knowledge while adapting to new tasks and domains (Liu et al., 2022; Jukić and Snajder, 2023).

Cross-lingual alignment. Aligning models across languages enhances their cross-lingual capabilities and enables better performance in multilingual settings. The simplest approach trains on unpaired monolingual corpora from different languages and already equips the models with remarkable multi- and cross-lingual transfer abilities (Conneau and Lample, 2019; Pires et al., 2019; Muennighoff et al., 2023b). Particularly, typologically divergent, distant languages and those with a small presence in the pre-training corpora benefit from subsequent self-supervised training, potentially with an extended vocabulary for better cross-lingual performance (Lauscher et al., 2020; Blevins and Zettlemoyer, 2022). Continued MLM pre-training with uniform sampling across high- and low-resource languages boosts performance on low-resource languages while preserving it on high-resource ones (Imani et al., 2023). Interestingly, Alabi et al. (2022) show the effect of the curse of multilinguality (Conneau et al., 2020) when adapting encoder

models with a few hundred million parameters with the MLM objective: models adapted for each language separately score better than training a single multilingual model. Having even a small amount of parallel data (e.g., translations) opens up a wide range of possibilities for learning better cross-lingual representations. The *translation language modeling* (TLM) objective predicts masked tokens in concatenated sentence pairs, learning to understand context across languages (Conneau and Lample, 2019). Another alternative masks tokens in the source language sentence and require the model to predict the matching sentence in the target language (Wang et al., 2023a).

Domain adaptation. Adapting pre-trained models to specialized domains ensures they can effectively handle domain-specific tasks, especially when the language and vocabulary of the specialized domain are considerably different from the pre-training language (Gururangan et al., 2020). There is a large number of domain-adapted language models for many fields, i.a., biomedicine (Luo et al., 2022; Bressem et al., 2024), legal (Noguti et al., 2023), finance (Zhao et al., 2021), and science (Beltagy et al., 2019). Domain adaptation works well with relatively small amounts of well-chosen data (Gnehm et al., 2022). It can be more effective than combining limited domain-specific data with general data and training from scratch (Türkmen et al., 2023). It also shows better generalization to unseen data compared to only fine-tuning on in-domain data (Bai et al., 2021). In low-resource settings without enough unlabeled data, using similar in-domain corpora for continued pre-training can improve performance (Jantscher et al., 2023), but the quality and relevance of the data can be just as important as its quantity (Mahapatra et al., 2022). Lu et al. (2023) show that the RTD objective focusing on domain-specific vocabulary is better than random masking. Incorporating an adversarial domain discriminator can further enhance domain adaptation for both encoder and decoder models (Du et al., 2020; Bellegarda, 2023). In the case of parameter-efficient methods (§ 4.1), pre-training the newly introduced parameters further improves performance (Liu et al., 2022; Jukić and Snajder, 2023). For cross-domain settings, combining the original pre-training data with in-domain corpora further boosts robustness (Diao et al., 2023).

4 Fine-tuning

Fine-tuning PLMs in low-resource scenarios poses challenges like overfitting and unstable optimization due to the limited amount of data compared to the huge model capacity, which inhibits generalization to unseen examples.

Catastrophic forgetting mitigation. To effectively fine-tune deeper transformer models with small data, proper optimization and regularization techniques are essential (Kalajdzievski, 2024). Zhang et al. (2021) explores the effect of training length and the varying number of re-initialized higher layers on BERT fine-tuning. Xu et al. (2021a) propose an input-based weight-scaling strategy to stabilize training and speed up the convergence of deep models with low data. They also empirically show that large batch size has a negative impact on generalization with small datasets. Layer-wise learning rate decay (LLRD) applies higher learning rates deeper into the network with the goal of preserving more general information from the pre-trained network and learning task-specific information in the last layers (Howard and Ruder, 2018). Regularization strategies mitigate catastrophic forgetting by encouraging parameters to remain close to their pre-trained values. Pre-trained weight decay applies a constraint on all parameters (Wiese et al., 2017), while Mixout (Lee et al., 2019) stochastically replaces some of the weights of the model with the pre-trained weights. SMART (Jiang et al., 2020a) incorporates smoothness-inducing adversarial regularization for robustness to small perturbations and Bregman proximal point optimization to avoid aggressive parameter updates outside a trust region. Besides optimization and regularization, incorporating external knowledge can also be very effective. Phang et al. (2019) propose fine-tuning first on a larger, intermediate task before moving on to the target task with limited data.

4.1 Parameter-efficient training

Fine-tuning the entire set of parameters (millions or billions) in PLMs is sample-inefficient and can be unstable in low-resource settings (Dodge et al., 2020). Parameter-efficient fine-tuning (PEFT) methods update only a reduced set of weights, avoiding high computational costs. They are particularly useful in data-scarce scenarios because they mitigate the risk of catastrophic forgetting associated with full re-training. Moreover, some

methods even managed to match the performance of full fine-tuning with only a small fraction of the parameters (Hu et al., 2021; Lester et al., 2021; Liu et al., 2022; Jukić and Snajder, 2023).

Masking-based methods do not add additional parameters but train only a subset of the model weights (specific layers, parameter types, etc.) while keeping the rest fixed. Early work simply trains (additional) last layers on top of encoder models (Devlin et al., 2019). BitFit (Ben Zaken et al., 2022) tunes the bias terms of the model, making it highly efficient. Another line of work focuses on subnetwork optimization based on gradient information, either by choosing a fixed subnetwork before training (Xu et al., 2021b; Ansell et al., 2022), or selecting it adaptively at train time using a multi-stage optimization strategy (Zhang et al., 2022; Yu et al., 2023). Masking-based methods are efficient and simple to implement; however, they tend to underperform compared to other PEFT methods that add new trainable parameters (Liu et al., 2022; Mao et al., 2022).

Adapters are trainable lightweight feedforward modules that are injected between the transformer layers, while the rest of the model is fixed (Houlsby et al., 2019; Pfeiffer et al., 2020). Sequential adapters act as a bottleneck in the model, requiring wider layers and more parameters compared to other PEFT methods to maintain performance (Hu et al., 2023b). Compacter (Karimi Mahabadi et al., 2021) alleviates this issue by utilizing low-rank matrices and parameter sharing. Adapters reduce training time and memory consumption but have a negative impact on inference time due to the additional model depth (Rücklé et al., 2021). Parallel adapters (He et al., 2021) and Ladder Side-Tuning (Sung et al., 2022) mitigate this by incorporating learnable modules in parallel to the backbone model. Multiple adapters can also be flexibly combined to handle complex tasks in a modular fashion (Pfeiffer et al., 2021; Wang et al., 2022; Chronopoulou et al., 2023).

Prefix-tuning builds upon the idea of in-context learning, but instead of finding discrete tokens, it optimizes continuous embeddings to serve as a task-specific context for the model. Specifically, the method prepends learned token vectors to the input key and values of multi-head attention layers in each transformer block, acting as virtual tokens to attend to (Li and Liang, 2021). IA^3 instead introduces learned vectors to scale the keys and values in attention mechanisms and feed-forward networks, showing improved performance (comparable to adapters) with an order of magnitude more parameters (Liu et al., 2022). Prompt-tuning (Lester et al., 2021) further reduces the number of parameters by limiting the prefix to the input embeddings. This method performs competitively with very large model sizes (billions of parameters) and shows slower convergence (Mao et al., 2022). Overall, prefix-tuning offers a flexible and efficient way to adapt models in cross-domain and cross-lingual low-resource scenarios (Tu et al., 2024; Goswami et al., 2023; Zhao et al., 2022) with considerably less parameters than adapters or reparametrization methods.

Reparametrization methods take inspiration from the observation that parameters of LLMs reside on a low-dimensional manifold. Intrinsic SAID (Aghajanyan et al., 2021) investigates the intrinsic dimensionality and projects the additive weight matrices into this subspace. Similarly, LoRA (Hu et al., 2021) decomposes the weight matrices into the product of two low-rank matrices, significantly reducing parameter count without compromising much performance. KronA (Edalati et al., 2022) replaces the Kronecker decomposition and shows better downstream performance. Reparametrization methods work best for large weight matrices and, thus, large models with moderate or greater amounts of data (Van Veen et al., 2023).

Hybrid methods combine multiple PEFT methods to leverage their individual strengths. UniPELT (Mao et al., 2022) uses a gating mechanism to dynamically activate adapters, prefix-tuning, and LoRA to optimize performance for a given task and data setup. Chen et al. (2022a) introduce design spaces to parametrize layer grouping, trainable parameter allocation, and PEFT strategy selection.

4.2 Embedding learning

Embedding vectors are the numerical representations of input tokens in NLP tasks. They are crucial for the success of LLMs in downstream tasks, enabling the model to capture the semantic information of the input text, including the nuances of the specific language, domain, and task (Collobert et al., 2011). However, language models are limited to a predefined vocabulary resulting from

the tokenization stage of fixed granularity. Choosing the granularity implies a trade-off in time and space complexity. Word-level granularity is more expressive but requires a larger vocabulary and more memory, while character- or byte-level granularity is more space-efficient but less expressive and produces very long sequences (Bojanowski et al., 2017; Ruder et al., 2023). As a compromise, most models use subword-level granularity, which is often a good balance between the two (Kudo and Richardson, 2018). The best granularity depends on the task and language. A fixed vocabulary is sensitive to small textual perturbations and limits generalization to new tasks and domains. To address this, Sun et al. (2023) propose training a shallow transformer to learn word representations from characters, making it robust to spelling errors and domain shifts. Recent work trains the token embeddings with frozen transformer body, to address the discrepancy between the pre-trained and target domain and language vocabulary (Artetxe et al., 2020; Hung et al., 2023). This approach is much more efficient than full fine-tuning and can act as an intermediate training step. Alternatively, Nag et al. (2023) identify words vulnerable to fragmentation based on the entropy of the token embeddings and augment the vocabulary with new embeddings for these words. Another approach to enhance cross-lingual transfer is to map the embeddings for different languages into a shared space using parallel data or seed dictionaries (Mikolov et al., 2013; Lalrempuii and Soni, 2023). Leveraging high-resource language embeddings this way can be highly beneficial for low-resource downstream tasks (Minixhofer et al., 2022; Deshpande et al., 2022; Deb et al., 2023). Other works establish an implicit mapping through alignment training objectives on paired data (Cao et al., 2019; Saadi et al., 2022).

4.3 Contrastive and adversarial learning

Contrastive and adversarial learning methods extract meaningful information from the differences and similarities across languages and domains, enhancing model alignment and adaptation.

Contrastive learning (CL) aims to learn effective representations by pulling semantically close pairs together and pushing apart unrelated samples (Chen et al., 2020b). It usually requires parallel data and can happen on multiple levels of granularity: sentence (Chi et al., 2021a) and word

(Chi et al., 2021b; Chen et al., 2023b). Alignment is significantly correlated with cross-lingual transfer across different languages and models (Gaschi et al., 2023), achieving remarkable performance in downstream tasks with labeled examples only in the source language (Hu et al., 2023a; Kowsher et al., 2023). A number of other works highlight the synergy of contrastive cross-linguality with PEFT methods like adapters (Liu et al., 2023b,a; Ahmat et al., 2023). Besides cross-lingual alignment, CL can also be beneficial at the downstream task level to make better use of the data. For text similarity tasks, it can act as an unsupervised objective by generating positive pairs with data augmentation techniques (Gao et al., 2021b; Yan et al., 2021). CL pairs can be easily created for binary classification tasks, like machine-generated text detection (Liu et al., 2023e), and for other classification tasks, like sentiment analysis or Named Entity Recognition (NER), by anchoring to the textual description of the class (Pauli et al., 2023) or reformulating as a question answering (QA) task (Chen et al., 2023c).

Adversarial learning refers to training two models simultaneously with contradicting objectives, guiding each other towards better performance (Goodfellow et al., 2014). Adversarial training can help bridge the gap between the pre-training and target domain or language without any paired data. The key mechanism is training a language or domain discriminator that needs to be deceived by the model, forcing it to learn robust representations that are domain-invariant (Du et al., 2020; Grießhaber et al., 2020) or language-agnostic (Lange et al., 2020; Huang et al., 2023c). This can also be combined with PEFT methods like adapters Ngo Trung et al. (2021) or learning prefixes that are either specific or independent of the domain Zhao et al. (2022).

4.4 Limited supervision

In low-resource scenarios, semi-supervised, unsupervised, and active learning methods can successfully leverage unlabeled data to boost model generalization and robustness.

Semi-supervised learning (SSL) leverages both labeled and unlabeled data during the training process (Chapelle et al., 2009). One common approach is self-training, where a model is trained on labeled data, and its predictions on unlabeled data

are treated as pseudo-labels for additional training mostly based on model confidence (Schick and Schütze, 2021a; Wang et al., 2023b; Lalrempuii and Soni, 2023), or entropy (Chen et al., 2020a). However, to avoid confirmation bias, proper regularization is essential (Toivanen et al., 2022). Unlike standard methods, consistency regularization promotes stable predictions on perturbed inputs (Sohn et al., 2020; Xie et al., 2020; Li et al., 2019). Another common SSL approach is co-training, where multiple modules are trained on different views of the input data and either predict pseudo-labels for each other or need to agree on the final prediction (Clark et al., 2018; Bhattacharjee et al., 2020). SSL has been particularly useful in transfer learning approaches, where language models can profit from the pre-trained knowledge and generate better pseudo-labels for the target task.

Unsupervised methods train models using only unlabeled data, making them especially useful in scenarios where labeled data is scarce or unavailable. The most prominent example and a key factor in the success of PLMs is their self-supervised pre-training objectives (§ 3), which can also be employed for the pre-training of PEFT modules like adapters (Diao et al., 2023). On the other hand, unsupervised methods during fine-tuning capitalize on some form of consistency condition. Besides consistency regularization (§ 4.4), which involves perturbing the input, the relation of different modalities can also be described using cycle-consistency (Zhu et al., 2017). Unpaired data is required to learn bidirectional relationships between potentially disparate domains (Karisani, 2022; Buehler, 2023), languages (Lample et al., 2018; Ren et al., 2019), or even text styles (Jalota et al., 2023). In practice, unsupervised objectives require a large amount of unlabeled data to learn meaningful representations of the data distributions. Therefore, they are often combined with direct supervision to improve performance.

Active learning (AL) techniques focus on selecting the most informative data points to maximize the effectiveness of limited training data. This method assumes a special scenario with unlabeled data and a constrained annotation budget, a situation that is common in real-world applications (Ren et al., 2021). High model uncertainty through metrics like confidence scores, entropy, Monte Carlo dropout, and perplexity are the most prominent data sampling criteria (Lewis and Gale, 1994; Gal and Ghahramani, 2016; Houlsby et al., 2011; Yuan et al., 2020; Muradoglu and Hulden, 2022; Jantscher et al., 2023). These can be combined with diversity-based sampling strategies – based on data distribution similarity or gradient variance – to ensure balanced data representation and mitigate outliers (Sener and Savarese, 2018; Gissin and Shalev-Shwartz, 2018; Ash et al., 2019; Ein-Dor et al., 2020; Margatina et al., 2021; Karamcheti et al., 2021; François and Gay, 2023).

Training models with AL is an iterative process where a small batch of unlabeled samples is selected for annotation based on the progressively refined model. Re-initializing the model between rounds is more stable than incrementally updating it with new data, especially for underrepresented classes (Lemmens and Daelemans, 2023). For the cold start problem, an alternative to random sampling is to use a self-supervised objective as a surrogate for uncertainty (Yuan et al., 2020). Integrating AL with PEFT techniques, such as adapters and UniPELT, has shown promising results in boosting performance for low-resource tasks (Jukić and Snajder, 2023). Overall, AL strategies allow models to learn efficiently from limited data, reducing the annotation burden and adjusting it to the task complexity.

5 Few-shot learning

In few-shot learning, the model is given limited examples of a new task and must generalize to unseen data. This section covers methods with and without fine-tuning on target data.

5.1 In-context learning

With the rise of large decoder models able to absorb a plethora of tasks implicitly from extensive text datasets, in-context learning (ICL), or prompting, has become a popular approach to leverage the generalization ability of these models (Liu et al., 2023c). After CLM pre-training, conversational models are often aligned with user intent via reinforcement learning to improve helpfulness, accuracy, and safety (Ouyang et al., 2022). This enables models to handle new tasks with only a few examples, without costly gradient-based training (Petroni et al., 2019; Radford et al., 2019; Brown et al., 2020). In addition, ICL can even outperform fine-tuning in very low-shot settings (Gao et al., 2021a; Jiang et al., 2020b; Garcia et al., 2023), but

performance can vary significantly with changes in prompt quality (Liu et al., 2023d).

Discrete prompts have the advantage of being intuitive and interpretable but mostly require handcrafting, making it inefficient for transferring to new models and tasks (Sanh et al., 2022). Suitable fill-in templates and prompt libraries can mitigate this burden (Shin et al., 2020; Bach et al., 2022; Bodonhelyi et al., 2024) and retriever modules offer an automatic way to select relevant examples, alleviating manual effort and boosting performance (Li et al., 2023a). Conveniently, any natural language understanding task can be formulated as a text generation task (Liu et al., 2023c), and there are many sophisticated prompting techniques available to improve few-shot learning performance (Xie et al., 2022; Liu et al., 2023c), like chain-of-thought (Wei et al., 2022). For low-resource languages, cross-lingual prompting offers an effective way to transfer knowledge from high-resource languages (Li et al., 2023b; Nie et al., 2023; Qin et al., 2023), with English task descriptions performing more consistently in few-shot settings (Lin et al., 2022; Chen et al., 2022b; Huang et al., 2023a). Some of the drawbacks of ICL include being most effective with very large generative models with large context windows, which in turn causes a dramatic increase of computational costs during inference (Reynolds and McDonell, 2021; Pawar et al., 2024).

5.2 Pattern-exploiting training

Pattern-exploiting training (PET), or prompt-based fine-tuning, formulates classification tasks in a cloze format (Taylor, 1953) enabling models to predict targets with an MLM objective, thus reconciling pre-training and fine-tuning (Schick and Schütze, 2021a,b). This approach is particularly suitable for few-shot classification in low-resource languages and specialized domains (Ullah et al., 2023; Song et al., 2023; Lu et al., 2023) but requires handcrafted patterns and verbalizers tailored to the task and data. Pattern templates transform the input into a cloze-style prompt, while verbalizers map the label into a target token sequence in the vocabulary. In cross-lingual scenarios, combining PET with a consistency loss helps to learn better inter-language correspondences (Qi et al., 2022). For inference, this method uses a costly autoregressive decoding scheme for verbalized targets consisting of multiple tokens. Numerous studies focus on optimizing inference efficiency by calculating the similarity with the average of the target embeddings (Hardalov et al., 2022), by using prototypical nearest neighbor decoding (Karimi Mahabadi et al., 2022), or evolutionary verbalizer search algorithms (Ling et al., 2023). Karimi Mahabadi et al. (2022) also replace the handcrafted verbalizers and patterns with learned label embeddings and task-specific adapters, respectively, improving performance.

5.3 Multi-task learning

Multitask fine-tuning has become the standard recipe for improving zero-shot task generalization for huge generative models with billions of parameters (Ruder, 2017; Aribandi et al., 2021; Wei et al., 2021). These models are instruction-tuned on a vast and diverse set of related downstream tasks, which makes them very strong baselines to compare against even when the limited data or computational resources do not allow to fine-tune them for a specific use case (Liu et al., 2019; Sanh et al., 2022; Van Veen et al., 2023). Multi-task models often use shared layers for learning robust common features while employing task-specific layers to capture the unique aspects of each task (Caruana, 1997; Ruder, 2017). Building upon this, mixture-of-experts (MoE) models route computation through small subnetworks to share representations between tasks and improve overall performance (Shazeer et al., 2017; Fedus et al., 2022; Zoph, 2022; Baniata and Kang, 2024).

5.4 Meta-learning

Inspired by human development theory, meta-learning emphasizes learning priors from past experiences that can facilitate efficient downstream adaptation. In the context of few-shot learning, metric-based approaches learn a similarity score in a latent space to compare new and seen samples. Class prototypes encode class-specific information in a metric space (Snell et al., 2017; Wen et al., 2021). Additional options to better structure the embedding space and include expert knowledge include instance- and class-specific CL combined with PET (Wu et al., 2024), or anchoring to class descriptions with triplet loss (Pauli et al., 2023). During inference, pairwise comparisons or other non-parametric learning algorithms like k-Nearest Neighbors (François and Gay, 2023) are used to predict the class of a new example. The ef-

fectiveness of these methods relies on representative class examples. Another prominent direction aims to learn a set of suitable initialization parameters that can be fine-tuned in a few optimization steps to achieve fast adaptation on a previously unseen task (Finn et al., 2017). Optimization-based techniques treat meta-learning as a two-level optimization problem, a task-specific inner loop and a task-agnostic outer loop. The updated weights from the inner loop support set are used to provide gradients on the query set to update the model parameters in the outer loop (Bansal et al., 2020; Li et al., 2022; Huang et al., 2023b; Chien et al., 2023). Supervised few-shot updates on the target task follow this. These two strategies are suitable for classification tasks with encoder models.

6 Data-efficient NLP techniques

This section overviews the most effective methods for different NLP tasks with limited data (Table 2).

Model selection. Choosing the right pre-trained model is crucial for achieving optimal performance in targeted tasks, domains, and languages. The most important factors include the model architecture, the number of parameters, and the pre-training data size, type, and quality (Alabi et al., 2020). Large generative models can be applied to all NLP tasks with appropriate task formulation (Liu et al., 2023c). However, for NLU tasks, bidirectional encoder models with only a few million parameters can match or outperform decoder models with billions of parameters (Schick and Schütze, 2021a; García-Díaz et al., 2023). At the same time, the decoder component is indispensable for text generation tasks and requires 1-2 orders of magnitude more parameters than encoder-only models (Brown et al., 2020). This is apparent from the fact that NLU tasks can be traced back to classification, which is much simpler than the task of text generation predicting the next tokens in the sequence from tens of thousands of tokens in the vocabulary (Radford et al., 2018).

For low-resource languages or specialized domains, larger models with 70B or more parameters, pre-trained on diverse multilingual or domain-specific data can provide strong zero- or few-shot in-context learning capabilities (Brown et al., 2020; Armengol-Estapé et al., 2022; Lin et al., 2022). In contrast, for fine-tuning with limited data, a generative model with 8-11B parameters can match the performance of substan-

tially larger models, while being much more efficient (Liu et al., 2022; Muennighoff et al., 2023b). For low-resource classification tasks, larger encoder models consistently outperform smaller ones (Schick and Schütze, 2021a; Karimi Mahabadi et al., 2022; Kowsher et al., 2023).

Dealing with limited data. In very low-resource scenarios, the quickest option for any NLP task is in-context learning (§ 5.1), which also does not require any gradient-based training and is often a competitive baseline (Liu et al., 2022; Chen et al., 2022b). For few-shot classification tasks, PET (§ 5.2) with adapters is very effective in leveraging pre-trained knowledge (Qi et al., 2022; Ullah et al., 2023; Karimi Mahabadi et al., 2022; Zhao et al., 2023). Below 1K annotated examples, additional annotation or using similar task datasets from a high-resource language or general domain for an intermediate fine-tuning step can greatly improve generalization to the target task (Chen et al., 2023c; Moscato et al., 2023; Laurer et al., 2024). Above this threshold, continued pre-training (§ 3.1) becomes increasingly effective, scaling with the amount of unlabeled data available (Bai et al., 2021; Mahapatra et al., 2022; Goswami et al., 2023). In contrast to the large-scale encoder pre-training with billions of tokens, continued pre-training – on task, domain, or language data – using only 100K tokens can benefit downstream tasks, especially for the introduced weights of PEFT methods (Gururangan et al., 2020; Gnehm et al., 2022; Jukić and Snajder, 2023). Besides the quantity of the pre-training data, the quality of the data is also an important factor (Mahapatra et al., 2022; Buonocore et al., 2023). With limited data, mitigating catastrophic forgetting with proper regularization is crucial (§ 4).

PEFT methods (§ 4.1) are particularly effective in adapting large pre-trained models to new tasks with limited data while preserving generalization capabilities (Table 2). Adapters, prefix-tuning, and in some cases, parameter masking tend to be better suited for less data, while LoRA and full fine-tuning perform better as the amount of labeled data increases (Tai et al., 2020; Li and Liang, 2021; Xu et al., 2021a; Mao et al., 2022; Buonocore et al., 2023; Wu et al., 2023; Ding et al., 2023; Jablonka et al., 2024). Prompt-tuning can also outperform full fine-tuning in low-resource settings, but it usually converges slower

Task	Model Selection	(Continued) Pre-training	Labeled Data*	Fine-tuning	Complementary Options
Text Classification	Encoder, ~400M params (e.g. RoBERTa)	75K+ tokens for PEFT params, MLM, LR warm-up + ER	< 250	PET + Adapters	Similar task data
			< 1K	NLI format with Prefix-tuning / UniPELT	Fine-tune first on high-resource NLI data
			< 10K	LoRA; LLRD	Active learning
			10K+	Full fine-tuning	Semi-supervised learning, Regularization (SMART)
Sequence Labeling	Encoder, ~400M params (e.g. ELECTRA)	25K+ tokens for PEFT params, RTD, LR warm-up + ER + Mixout	< 250	PET; QA format with CL	Multi-task learning, KGs for NER
			< 5K	Adapters, Parameter masking	Similar task datasets, Active learning
			5K+	Full fine-tuning	LLRD
Text Comparison	Encoder, ~400M params (e.g. DeBERTa)	75K+ tokens for PEFT params, MLM, consistency CL, LR warm-up	< 250	PET + Adapters / Prompt-tuning	Data augmentation
			< 1K	Adapters / UniPELT	Multilingual data with CL
			< 10K	PEFT (LoRA)	CL on NLI data
			10K+	Full fine-tuning	Bounded gradient updates
Text Generation	Decoder, ~8-13B params (e.g. Llama)	100K+ tokens for PEFT params, CLM, regularization	< 1k	In-context learning / RAG	Large-scale multitask models
			< 15K	Prefix- / Prompt-tuning	Data augmentation (e.g. Translation)
			15K+	LoRA	Semi-supervised learning
Text Transformation	Encoder-Decoder, ~8-13B params (e.g. T5 / T0)	100K+ tokens for PEFT params, S2S + T0 Multitask, LR warm-up	< 1k	In-context learning / RAG	Large-scale multitask models
			< 15K	PEFT (IA^3, Prefix-tuning)	Intermediate domain/task, Cycle-consistency
			15K+	LoRA	Semi-supervised learning

Table 2: Suggested approaches for NLP task groups with limited data. All information in this table was compiled from reviewed papers. *Abbreviations not defined within the text*: ER (Experience Replay), LLRD (Layer-wise Learning Rate Decay), LR (Learning Rate), NLI (Natural Language Inference), RAG (Retrieval-Augmented Generation). *Labeled data for Sequence Labeling is provided in sentences; for other tasks, in examples.

than prefix-tuning and works best with larger models (Goswami et al., 2023; Choi and Lee, 2023). As complementary options, CL improves representation quality for downstream tasks (Gao et al., 2021b; Yan et al., 2021) and across languages (Hu et al., 2023a; Kowsher et al., 2023), active learning maximizes data utility (Yuan et al., 2020; Lemmens and Daelemans, 2023; Jantscher et al., 2023), and semi-supervised learning further improves performance and robustness (Clark et al., 2018; Wang et al., 2023b; Shi et al., 2023).

7 Conclusion

This survey addresses the challenges of applying LMs in data-scarce scenarios. Specifically, we first give a systematic overview of methods addressing the important aspects to be considered for effective and efficient subsequent pre-training and downstream fine-tuning with scarce data, also highlighting advantages and limitations.

We categorize NLP tasks into five groups and provide a summary of suitable pre-trained models, adaptation and fine-tuning methods, and auxiliary options across different dimensions of data availability. Our findings suggest that choosing larger models and combining suitable parameter-efficient methods with appropriate regularization techniques and complementary training options can significantly improve performance in low-resource scenarios. There is still limited theoretical and experimental work on preventing catastrophic forgetting during model adaptation across fixed degrees of data scarcity. Moreover, benchmarking a broader spectrum of methods across different tasks at the intersection of specialized domains and resource-poor languages is largely neglected, also emphasizing the need for additional public datasets and standardized evaluation frameworks. We also encourage the community to investigate the combination of different methods to leverage their complementary strengths.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7319–7328, Online. Association for Computational Linguistics.

Ahtamjan Ahmat, Yating Yang, Bo Ma, Rui Dong, Kaiwen Lu, and Lei Wang. 2023. WAD-X: Improving Zero-shot Cross-lingual Transfer via Adapter-based Word Alignment. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(9):224:1–224:23.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable Sparse Fine–Tuning for Cross-Lingual Transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2021. ExT5: Towards Extreme Multi-Task Scaling for Transfer Learning. In *International Conference on Learning Representations*.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the Multilingual Capabilities of Very Large-Scale English Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2019. Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds. In *International Conference on Learning Representations*.

Cyprien de Masson d' Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic Memory in Lifelong Language Learning. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-jian Jiang, and Alexander Rush. 2022. PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 93–104, Dublin, Ireland. Association for Computational Linguistics.

Fan Bai, Alan Ritter, and Wei Xu. 2021. Pre-train or Annotate? Domain Adaptation with a Constrained Budget. In *Proceedings of the 2021 Conference on Empirical Methods in Natural*

Language Processing, pages 5002–5015, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Guangji Bai, Zheng Chai, Chen Ling, Shiyu Wang, Jiaying Lu, Nan Zhang, Tingwei Shi, Ziyang Yu, Mengdan Zhu, Yifei Zhang, Carl Yang, Yue Cheng, and Liang Zhao. 2024. Beyond Efficiency: A Systematic Survey of Resource-Efficient Large Language Models. ArXiv:2401.00625v3 [cs].

Laith H. Baniata and Sangwoo Kang. 2024. Switch-Transformer Sentiment Analysis Model for Arabic Dialects That Utilizes a Mixture of Experts Mechanism. *Mathematics*, 12(2):242. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Trapit Bansal, Rishikesh Jha, Tsendsuren Munkhdalai, and Andrew McCallum. 2020. Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 522–534, Online. Association for Computational Linguistics.

Jerome R. Bellegarda. 2023. Pareto Efficiency of Learning-Forgetting Trade-Off in Neural Language Model Adaptation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. 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)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Elad Ben Zaken, Yoav Goldberg, and Shauli Ravfogel. 2022. BitFit: Simple Parameter-efficient Fine-tuning for Transformer-based Masked Language-models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Kasturi Bhattacharjee, Miguel Ballesteros, Rishita Anubhai, Smaranda Muresan, Jie Ma, Faisal Ladhak, and Yaser Al-Onaizan. 2020. To BERT or Not to BERT: Comparing Task-specific and Task-agnostic Semi-Supervised Approaches for Sequence Tagging. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7927–7934, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2022. Language Contamination Helps Explains the Cross-lingual Capabilities of English Pretrained Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3563–3574, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anna Bodonhelyi, Efe Bozkir, Shuo Yang, Enkelejda Kasneci, and Gjergji Kasneci. 2024. User Intent Recognition and Satisfaction with Large Language Models: A User Study with ChatGPT. ArXiv:2402.02136v1.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146. Place: Cambridge, MA Publisher: MIT Press.

Keno K. Bressem, Jens-Michalis Papaioannou, Paul Grundmann, Florian Borchert, Lisa C. Adams, Leonhard Liu, Felix Busch, Lina Xu, Jan P. Loyen, Stefan M. Niehues, Moritz Augustin, Lennart Grosser, Marcus R. Makowski, Hugo J. W. L. Aerts, and Alexander Löser. 2024. medBERT.de: A comprehensive German BERT model for the medical domain. *Expert Systems with Applications*, 237:121598.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. ArXiv:2005.14165v4 [cs].

Markus J. Buehler. 2023. Unsupervised cross–domain translation via deep learning and adversarial attention neural networks and application to music-inspired protein designs. *Patterns*, 4(3):100692.

Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431.

Steven Cao, Nikita Kitaev, and Dan Klein. 2019. Multilingual Alignment of Contextual Word Representations. In *International Conference on Learning Representations*.

Rich Caruana. 1997. Multitask Learning. *Machine Learning*, 28(1):41–75.

O. Chapelle, B. Scholkopf, and A. Zien, Eds. 2009. Semi-Supervised Learning. *IEEE Transactions on Neural Networks*, 20(3):542–542.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023a. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211. Place: Cambridge, MA Publisher: MIT Press.

Jiaao Chen, Zichao Yang, and Diyi Yang. 2020a. MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2147–2157, Online. Association for Computational Linguistics.

Jiaao Chen, Aston Zhang, Xingjian Shi, Mu Li, Alex Smola, and Diyi Yang. 2022a. Parameter-Efficient Fine-Tuning Design Spaces. In *International Conference on Learning Representations*.

Nuo Chen, Linjun Shou, Tengtao Song, Ming Gong, Jian Pei, Jianhui Chang, Daxin Jiang, and Jia Li. 2023b. Structural Contrastive Pretraining for Cross-Lingual Comprehension. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2042–2057, Toronto, Canada. Association for Computational Linguistics.

Peng Chen, Jian Wang, Hongfei Lin, Di Zhao, and Zhihao Yang. 2023c. Few-shot biomedical named entity recognition via knowledge-guided instance generation and prompt contrastive learning. *Bioinformatics*, 39(8):btad496.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1597–1607. PMLR. ISSN: 2640-3498.

Yuxuan Chen, David Harbecke, and Leonhard Hennig. 2022b. Multilingual Relation Classification via Efficient and Effective Prompting. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1059–1075, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving Pretrained Cross-Lingual Language Models via Self-Labeled Word Alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430, Online. Association for Computational Linguistics.

Jen-Tzung Chien, Ming-Yen Chen, and Jing-Hao Xue. 2023. Learning Meta Soft Prompt for Few-Shot Language Models. In *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 57–62. ISSN: 2640-0103.

YunSeok Choi and Jee-Hyong Lee. 2023. Code-Prompt: Task-Agnostic Prefix Tuning for Program and Language Generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5282–5297, Toronto, Canada. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, Alexander Fraser, and Jesse Dodge. 2023. AdapterSoup: Weight Averaging to Improve Generalization of Pretrained Language Models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2054–2063, Dubrovnik, Croatia. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. 2018. Semi-Supervised Sequence Modeling with Cross-View Training. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1914–1925, Brussels, Belgium. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, 12(null):2493–2537.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual Language Model Pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Claudio Crema, Tommaso Mario Buonocore, Silvia Fostinelli, Enea Parimbelli, Federico Verde, Cira Fundarò, Marina Manera, Matteo Cotta Ramusino, Marco Capelli, Alfredo Costa, Giuliano Binetti, Riccardo Bellazzi, and Alberto Redolfi. 2023. Advancing Italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application. *Journal of Biomedical Informatics*, 148:104557.

Ujan Deb, Ridayesh Parab, and Preethi Jyothi. 2023. Zero-shot Cross-lingual Transfer With Learned Projections Using Unlabeled Target-Language Data. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 449–457, Toronto, Canada. Association for Computational Linguistics.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shizhe Diao, Tianyang Xu, Ruijia Xu, Jiawei Wang, and Tong Zhang. 2023. Mixture-of-Domain-Adapters: Decoupling and Injecting Domain Knowledge to Pre-trained Language Models' Memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5113–5129, Toronto, Canada. Association for Computational Linguistics.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, Jing Yi, Weilin Zhao, Xiaozhi Wang, Zhiyuan Liu,

Hai-Tao Zheng, Jianfei Chen, Yang Liu, Jie Tang, Juanzi Li, and Maosong Sun. 2023. Parameter-efficient fine-tuning of large-scale pre–trained language models. *Nature Machine Intelligence*, 5(3):220–235. Publisher: Nature Publishing Group.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-Tuning Pretrained Language Models: Weight Initializations, Data Orders, and Early Stopping. ArXiv:2002.06305v1 [cs].

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and Domain-Aware BERT for Cross-Domain Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

Ali Edalati, Marzieh Tahaei, Ivan Kobyzev, Vahid Partovi Nia, James J. Clark, and Mehdi Rezagholizadeh. 2022. KronA: Parameter Efficient Tuning with Kronecker Adapter. ArXiv:2212.10650v1 [cs].

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A Survey of Data Augmentation Approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135. PMLR. ISSN: 2640-3498.

Matthieu François and Paul Gay. 2023. Active Learning with few shot learning for crisis management. In *Proceedings of the 20th International Conference on Content-based Multimedia Indexing*, CBMI '23, pages 233–237, New York, NY, USA. Association for Computing Machinery.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1050–1059. PMLR. ISSN: 1938-7228.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021a. Making Pre-trained Language Models Better Few-shot Learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The Unreasonable Effectiveness of Few-shot Learning for Machine Translation. In *Proceedings of the 40th International Conference on Machine Learning*, pages 10867–10878. PMLR. ISSN: 2640-3498.

José Antonio García-Díaz, Ronghao Pan, and Rafael Valencia-García. 2023. Leveraging Zero and Few-Shot Learning for Enhanced Model Generality in Hate Speech Detection in Spanish and English. *Mathematics*, 11(24):5004. Number: 24 Publisher: Multidisciplinary Digital Publishing Institute.

Felix Gaschi, Patricio Cerda, Parisa Rastin, and Yannick Toussaint. 2023. Exploring the Relationship between Alignment and Cross-lingual Transfer in Multilingual Transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3020–3042, Toronto, Canada. Association for Computational Linguistics.

Daniel Gissin and Shai Shalev-Shwartz. 2018. Discriminative Active Learning.

Ann-Sophie Gnehm, Eva Bühlmann, and Simon Clematide. 2022. Evaluation of Transfer Learning and Domain Adaptation for Analyzing German-Speaking Job Advertisements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3892–3901, Marseille, France. European Language Resources Association.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Koustava Goswami, Lukas Lange, Jun Araki, and Heike Adel. 2023. SwitchPrompt: Learning Domain-Specific Gated Soft Prompts for Classification in Low-Resource Domains. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2689–2695, Dubrovnik, Croatia. Association for Computational Linguistics.

Daniel Grießhaber, Ngoc Thang Vu, and Johannes Maucher. 2020. Low-resource text classification using domain-adversarial learning. *Computer Speech & Language*, 62:101056.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre–trained models: Past, present and future. *AI Open*, 2:225–250.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-Shot Cross-Lingual Stance Detection with Sentiment-Based Pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10729–10737. Number: 10.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a Unified View of Parameter-Efficient Transfer Learning. In *International Conference on Learning Representations*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training Compute-Optimal Large Language Models. In *Advances in Neural Information Processing Systems*, volume 36, pages 30016–30030.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR. ISSN: 2640-3498.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian Active Learning for Classification and Preference Learning. ArXiv:1112.5745v1.

Jeremy Howard and Sebastian Ruder. 2018. Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Xiyang Hu, Xinchi Chen, Peng Qi, Deguang Kong, Kunlun Liu, William Yang Wang, and Zhiheng Huang. 2023a. Language Agnostic Multilingual Information Retrieval with Contrastive Learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9133–9146, Toronto, Canada. Association for Computational Linguistics.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023b. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine–Tuning of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023a. Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Yinqiu Huang, Min Gao, Jia Wang, Junwei Yin, Kai Shu, Qilin Fan, and Junhao Wen. 2023b. Meta-prompt based learning for low-resource false information detection. *Information Processing & Management*, 60(3):103279.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023c. Improving Cross-lingual Information Retrieval on Low-Resource Languages via Optimal Transport Distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, pages 1048–1056, New York, NY, USA. Association for Computing Machinery.

Chia-Chien Hung, Lukas Lange, and Jannik Strötgen. 2023. TADA: Efficient Task-Agnostic Domain Adaptation for Transformers. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 487–503, Toronto, Canada. Association for Computational Linguistics.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Kevin Maik Jablonka, Philippe Schwaller, Andres Ortega-Guerrero, and Berend Smit. 2024. Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2):161–169. Publisher: Nature Publishing Group.

Rricha Jalota, Koel Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2023. Translating away Translationese without Parallel Data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7086–7100, Singapore. Association for Computational Linguistics.

Michael Jantscher, Felix Gunzer, Roman Kern, Eva Hassler, Sebastian Tschauner, and Gernot Reishofer. 2023. Information extraction from German radiological reports for general clinical text and language understanding. *Scientific Reports*, 13(1):2353. Publisher: Nature Publishing Group.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao.

2020a. SMART: Robust and Efficient Fine–Tuning for Pre-trained Natural Language Models through Principled Regularized Optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020b. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438. Place: Cambridge, MA Publisher: MIT Press.

Josip Jukić and Jan Snajder. 2023. Parameter–Efficient Language Model Tuning with Active Learning in Low-Resource Settings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5061–5074, Singapore. Association for Computational Linguistics.

Damjan Kalajdzievski. 2024. Scaling Laws for Forgetting When Fine-Tuning Large Language Models. ArXiv:2401.05605v1 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models. ArXiv:2001.08361v1 [cs, stat].

Siddharth Karamcheti, Ranjay Krishna, Li Fei-Fei, and Christopher Manning. 2021. Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7265–7281, Online. Association for Computational Linguistics.

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient Low-Rank Hypercomplex Adapter Layers. In *Advances in Neural Information Processing Systems*, volume 34, pages 1022–1035. Curran Associates, Inc.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Lambert Mathias, Marzieh

Saeidi, Veselin Stoyanov, and Majid Yazdani. 2022. Prompt-free and Efficient Few-shot Learning with Language Models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3638–3652, Dublin, Ireland. Association for Computational Linguistics.

Payam Karisani. 2022. Multiple-Source Domain Adaptation via Coordinated Domain Encoders and Paired Classifiers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(7):7087–7095. Number: 7.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526. Publisher: Proceedings of the National Academy of Sciences.

Md Kowsher, Md. Shohanur Islam Sobuj, Nusrat Jahan Prottasha, Mohammad Shamsul Arefin, and Yasuhiko Morimoto. 2023. Contrastive Learning for Universal Zero-Shot NLI with Cross-Lingual Sentence Embeddings. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 239–252, Singapore. Association for Computational Linguistics.

Panteleimon Krasadakis, Evangelos Sakkopoulos, and Vassilios S. Verykios. 2024. A Survey on Challenges and Advances in Natural Language Processing with a Focus on Legal Informatics and Low-Resource Languages. *Electronics*, 13(3):648. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Candy Lalrempuii and Badal Soni. 2023. Investigating Unsupervised Neural Machine Translation for Low-resource Language Pair English-Mizo via Lexically Enhanced Pre-trained Language Models. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(8):209:1–209:18.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-Based & Neural Unsupervised Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Lukas Lange, Anastasiia Iurshina, Heike Adel, and Jannik Strötgen. 2020. Adversarial Alignment of Multilingual Models for Extracting Temporal Expressions from Text. In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 103–109, Online. Association for Computational Linguistics.

Egoitz Laparra, Aurelie Mascio, Sumithra Velupillai, and Timothy Miller. 2021. A Review of Recent Work in Transfer Learning and Domain Adaptation for Natural Language Processing of Electronic Health Records. *Yearbook of Medical Informatics*, 30(1):239–244.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI. *Political Analysis*, 32(1):84–100.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. 2019. Mixout: Effective Regularization to Finetune Large-scale Pretrained Language Models. In *International Conference on Learning Representations*.

Jens Lemmens and Walter Daelemans. 2023. Combining Active Learning and Task Adaptation with BERT for Cost-Effective Annotation of Social Media Datasets. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 237–250, Toronto, Canada. Association for Computational Linguistics.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

David D. Lewis and William A. Gale. 1994. A Sequential Algorithm for Training Text Classifiers. In *SIGIR '94*, pages 3–12, London. Springer.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chunyou Li, Mingtong Liu, Hongxiao Zhang, Yufeng Chen, Jinan Xu, and Ming Zhou. 2023a. MT2: Towards a Multi-Task Machine Translation Model with Translation-Specific In-Context Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8616–8627, Singapore. Association for Computational Linguistics.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2022. Few-Shot Named Entity Recognition via Meta-Learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4245–4256. Conference Name: IEEE Transactions on Knowledge and Data Engineering.

Junnan Li, Richard Socher, and Steven C. H. Hoi. 2019. DivideMix: Learning with Noisy Labels

as Semi-supervised Learning. In *International Conference on Learning Representations*.

Xiang Lisa Li and Percy Liang. 2021. Prefix–Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Xiaoqian Li, Ercong Nie, and Sheng Liang. 2023b. Crosslingual Retrieval Augmented In--context Learning for Bangla. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 136–151, Singapore. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot Learning with Multilingual Generative Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tongtao Ling, Lei Chen, Yutao Lai, and Hai-Lin Liu. 2023. Evolutionary Verbalizer Search for Prompt-Based Few Shot Text Classification. In *Knowledge Science, Engineering and Management*, pages 279–290, Cham. Springer Nature Switzerland.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A. Raffel. 2022. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Meizhen Liu, Xu Guo, He Jiakai, Jianye Chen, Fengyu Zhou, and Siu Hui. 2023a. InteMATs: Integrating Granularity-Specific Multilingual Adapters for Cross-Lingual Transfer. In

Findings of the Association for Computational Linguistics: EMNLP 2023, pages 5035–5049, Singapore. Association for Computational Linguistics.

Meizhen Liu, Jiakai He, Xu Guo, Jianye Chen, Siu Cheung Hui, and Fengyu Zhou. 2023b. GranCATs: Cross-Lingual Enhancement through Granularity-Specific Contrastive Adapters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, pages 1461–1471, New York, NY, USA. Association for Computing Machinery.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023c. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):1–35.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023d. GPT understands, too. *AI Open*.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics.

Xiaoming Liu, Zhaohan Zhang, Yichen Wang, Hang Pu, Yu Lan, and Chao Shen. 2023e. CoCo: Coherence-Enhanced Machine-Generated Text Detection Under Low Resource With Contrastive Learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16167–16188, Singapore. Association for Computational Linguistics.

Keming Lu, Peter Potash, Xihui Lin, Yuwen Sun, Zihan Qian, Zheng Yuan, Tristan Naumann, Tianxi Cai, and Junwei Lu. 2023. Prompt Discriminative Language Models for Domain Adaptation. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 247–258, Toronto, Canada. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan

Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.

Aniruddha Mahapatra, Sharmila Reddy Nangi, Aparna Garimella, and Anandhavelu N. 2022. Entity Extraction in Low Resource Domains with Selective Pre-training of Large Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 942–951, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuning Mao, Lambert Mathias, Rui Hou, Amjad Almahairi, Hao Ma, Jiawei Han, Scott Yih, and Madian Khabsa. 2022. UniPELT: A Unified Framework for Parameter-Efficient Language Model Tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6253–6264, Dublin, Ireland. Association for Computational Linguistics.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active Learning by Acquiring Contrastive Examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In Gordon H. Bower, editor, *Psychology of Learning and Motivation*, volume 24, pages 109–165. Academic Press.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting Similarities among Languages for Machine Translation. ArXiv:1309.4168v1 [cs].

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

Vincenzo Moscato, Giuseppe Napolano, Marco Postiglione, and Giancarlo Sperlì. 2023. Multi-task learning for few-shot biomedical relation extraction. *Artif. Intell. Rev.*, 56(11):13743–13763.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A. Raffel. 2023a. Scaling Data-Constrained Language Models. *Advances in Neural Information Processing Systems*, 36:50358–50376.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023b. Crosslingual Generalization through Multitask Finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. How to choose data for morphological inflection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7294–7303, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2023. Entropy-guided Vocabulary Augmentation of Multilingual Language Models for Low-resource Tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8619–8629, Toronto, Canada. Association for Computational Linguistics.

Nghia Ngo Trung, Duy Phung, and Thien Huu Nguyen. 2021. Unsupervised Domain Adaptation for Event Detection using Domain-specific Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP*

2021, pages 4015–4025, Online. Association for Computational Linguistics.

Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. 2023. Cross-Lingual Retrieval Augmented Prompt for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8320–8340, Toronto, Canada. Association for Computational Linguistics.

Mariana Noguti, Eduardo Vellasques, and Luiz S. Oliveira. 2023. A Small Claims Court for the NLP: Judging Legal Text Classification Strategies With Small Datasets. In *2023 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1840–1845. ISSN: 2577-1655.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Amalie Pauli, Leon Derczynski, and Ira Assent. 2023. Anchoring Fine-tuning of Sentence Transformer with Semantic Label Information for Efficient Truly Few-shot Classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11254–11264, Singapore. Association for Computational Linguistics.

Saurav Pawar, S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Aman Chadha, and Amitava Das. 2024. The What, Why, and How of Context Length Extension Techniques in Large Language Models – A Detailed Survey. ArXiv:2401.07872v1.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language Models as Knowledge Bases? 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)*, pages 2463–2473,

Hong Kong, China. Association for Computational Linguistics.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-Destructive Task Composition for Transfer Learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R. Bowman. 2019. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. ArXiv:1811.01088v2 [cs].

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is Multilingual BERT? ArXiv:1906.01502v1 [cs].

Kunxun Qi, Hai Wan, Jianfeng Du, and Haolan Chen. 2022. Enhancing Cross-lingual Natural Language Inference by Prompt-learning from Cross-lingual Templates. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1910–1923, Dublin, Ireland. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual Prompting: Improving Zero-shot Chain-of-Thought Reasoning across Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre–Training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vinay Venkatesh Ramasesh, Aitor Lewkowycz, and Ethan Dyer. 2021. Effect of scale on catastrophic forgetting in neural networks. In *International Conference on Learning Representations*.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B. Gupta, Xiaojiang Chen, and Xin Wang. 2021. A Survey of Deep Active Learning. *ACM Comput. Surv.*, 54(9):180:1–180:40.

Shuo Ren, Zhirui Zhang, Shujie Liu, Ming Zhou, and Shuai Ma. 2019. Unsupervised Neural Machine Translation with SMT as Posterior Regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):241–248. Number: 01.

Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, pages 1–7, New York, NY, USA. Association for Computing Machinery.

Sebastian Ruder. 2017. An Overview of Multi-Task Learning in Deep Neural Networks. ArXiv:1706.05098v1 [cs].

Sebastian Ruder, Jonathan Clark, Alexander Gutkin, Mihir Kale, Min Ma, Massimo Nicosia, Shruti Rijhwani, Parker Riley, Jean-Michel Sarr, Xinyi Wang, John Wieting, Nitish Gupta, Anna Katanova, Christo Kirov, Dana Dickinson, Brian Roark, Bidisha Samanta, Connie Tao, David Adelani, Vera Axelrod, Isaac Caswell, Colin Cherry, Dan Garrette, Reeve Ingle, Melvin Johnson, Dmitry Panteleev, and Partha Talukdar. 2023. XTREME-UP: A User-Centric Scarce-Data Benchmark for Under-Represented Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1856–1884, Singapore. Association for Computational Linguistics.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the Efficiency of Adapters in Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hossain Shaikh Saadi, Viktor Hangya, Tobias Eder, and Alexander Fraser. 2022. Comparative Analysis of Cross-lingual Contextualized Word Embeddings. In *Proceedings of the 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 64–75, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *International Conference on Learning Representations*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Ozan Sener and Silvio Savarese. 2018. Active Learning for Convolutional Neural Networks: A Core-Set Approach. In *International Conference on Learning Representations*.

Noam Shazeer, *Azalia Mirhoseini, *Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*.

Zhengxiang Shi, Francesco Tonolini, Nikolaos Aletras, Emine Yilmaz, Gabriella Kazai, and Yunlong Jiao. 2023. Rethinking Semi-supervised Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5614–5634, Toronto, Canada. Association for Computational Linguistics.

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

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. *Advances in Neural Information Processing Systems*, 33:596–608.

Chengyu Song, Fei Cai, Mengru Wang, Jianming Zheng, and Taihua Shao. 2023. TaxonPrompt: Taxonomy-aware curriculum prompt learning for few-shot event classification. *Knowledge-Based Systems*, 264:110290.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5926–5936. PMLR. ISSN: 2640-3498.

Nikolaos Stylianou, Despoina Chatzakou, Theodora Tsikrika, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2023. Domain-Aligned Data Augmentation for Low-Resource and Imbalanced Text Classification. In *Advances in Information Retrieval*, pages 172–187, Cham. Springer Nature Switzerland.

Li Sun, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio, and Cha Zhang. 2023. From Characters to Words: Hierarchical Pre–trained Language Model for Open-vocabulary Language Understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3605–3620, Toronto, Canada. Association for Computational Linguistics.

Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. LST: Ladder Side-Tuning for Parameter and Memory Efficient Transfer Learning. *Advances in Neural Information Processing Systems*, 35:12991–13005.

Wen Tai, H. T. Kung, Xin Dong, Marcus Comiter, and Chang-Fu Kuo. 2020. exBERT: Extending Pre-trained Models with Domain-specific Vocabulary Under Constrained Training Resources. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1433–1439, Online. Association for Computational Linguistics.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433. Publisher: SAGE Publications.

Ida Toivanen, Jari Lindroos, Venla Räsänen, and Sakari Taipale. 2022. Dealing with a small amount of data – developing Finnish sentiment analysis. In *2022 9th International Conference on Behavioural and Social Computing (BESC)*, pages 1–7.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H. Martins, André

F. T. Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. 2023. Efficient Methods for Natural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 11:826–860. Place: Cambridge, MA Publisher: MIT Press.

Lifu Tu, Jin Qu, Semih Yavuz, Shafiq Joty, Wenhao Liu, Caiming Xiong, and Yingbo Zhou. 2024. Efficiently Aligned Cross-Lingual Transfer Learning for Conversational Tasks using Prompt-Tuning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1278–1294, St. Julian's, Malta. Association for Computational Linguistics.

Hazal Türkmen, Oğuz Dikenelli, Cenk Eraslan, Mehmet Cem Çallı, and Süha Süreyya Özbek. 2023. BioBERTurk: Exploring Turkish Biomedical Language Model Development Strategies in Low-Resource Setting. *Journal of Healthcare Informatics Research*, 7(4):433–446.

Faizad Ullah, Ubaid Azam, Ali Faheem, Faisal Kamiran, and Asim Karim. 2023. Comparing Prompt-Based and Standard Fine-Tuning for Urdu Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6747–6754, Singapore. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, Akshay Chaudhari, and John Pauly. 2023. RadAdapt: Radiology Report Summarization via Lightweight Domain Adaptation of Large Language Models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460, Toronto, Canada. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Towards Unifying Multi-Lingual and Cross-Lingual Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15127–15143, Toronto, Canada. Association for Computational Linguistics.

Minzheng Wang, Jia Cao, Qingchao Kong, and Yin Luo. 2023b. Boosting Domain-Specific Question Answering Through Weakly Supervised Self-Training. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6.

Yaqing Wang, Sahaj Agarwal, Subhabrata Mukherjee, Xiaodong Liu, Jing Gao, Ahmed Hassan Awadallah, and Jianfeng Gao. 2022. AdaMix: Mixture-of-Adaptations for Parameter-efficient Model Tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5744–5760, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned Language Models are Zero-Shot Learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Wen Wen, Yongbin Liu, Chunping Ouyang, Qiang Lin, and Tonglee Chung. 2021. Enhanced prototypical network for few-shot relation extraction. *Information Processing & Management*, 58(4):102596.

Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural Domain Adaptation for Biomedical Question Answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 281–289, Vancouver, Canada. Association for Computational Linguistics.

Hui Wu, Yuting He, Yidong Chen, Yu Bai, and Xiaodong Shi. 2024. Improving few-shot relation extraction through semantics-guided learning. *Neural Networks*, 169:453–461.

Ting-Wei Wu, Changsheng Zhao, Ernie Chang, Yangyang Shi, Pierce Chuang, Vikas Chandra, and Biing Juang. 2023. Towards Zero-Shot Multilingual Transfer for Code-Switched Responses. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7551–7563, Toronto, Canada. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised Data Augmentation for Consistency Training. In *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An Explanation of In-context Learning as Implicit Bayesian Inference. In *International Conference on Learning Representations*.

Peng Xu, Dhruv Kumar, Wei Yang, Wenjie Zi, Keyi Tang, Chenyang Huang, Jackie Chi Kit Cheung, Simon J.D. Prince, and Yanshuai Cao. 2021a. Optimizing Deeper Transformers on Small Datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2089–2102, Online. Association for Computational Linguistics.

Runxin Xu, Fuli Luo, Zhiyuan Zhang, Chuanqi Tan, Baobao Chang, Songfang Huang, and Fei Huang. 2021b. Raise a Child in Large Language Model: Towards Effective and Generalizable Fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9514–9528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuanmeng Yan, Rumei Li, Sirui Wang, Fuzheng Zhang, Wei Wu, and Weiran Xu. 2021. ConSERT: A Contrastive Framework for Self-Supervised Sentence Representation Transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5065–5075, Online. Association for Computational Linguistics.

Bılge Kağan Yazar, Durmuş Özkan Şahin, and Erdal Kiliç. 2023. Low-Resource Neural Machine Translation: A Systematic Literature Review. *IEEE Access*, 11:131775–131813. Conference Name: IEEE Access.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning Bias in Language Models by Partitioning Gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada. Association for Computational Linguistics.

Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start Active Learning through Self-supervised Language Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7935–7948, Online. Association for Computational Linguistics.

Haojie Zhang, Ge Li, Jia Li, Zhongjin Zhang, Yuqi Zhu, and Zhi Jin. 2022. Fine-Tuning Pre-Trained Language Models Effectively by Optimizing Subnetworks Adaptively. *Advances in Neural Information Processing Systems*, 35:21442–21454.

Tianyi Zhang, Felix Wu, Arzoo Katiyar, Kilian Q. Weinberger, and Yoav Artzi. 2021. Revisiting Few-sample BERT Fine-tuning. In *International Conference on Learning Representations*.

Haodong Zhao, Ruifang He, Mengnan Xiao, and Jing Xu. 2023. Infusing Hierarchical Guidance into Prompt Tuning: A Parameter-Efficient Framework for Multi-level Implicit Discourse Relation Recognition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6477–6492, Toronto, Canada. Association for Computational Linguistics.

Lingyun Zhao, Lin Li, Xinhao Zheng, and Jianwei Zhang. 2021. A BERT based Sentiment Analysis and Key Entity Detection Approach for Online Financial Texts. In *2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 1233–1238.

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Ruotong Geng, Huixing Jiang, Wei Wu, and Weiran Xu. 2022. ADPL: Adversarial Prompt-based Domain Adaptation for Dialogue Summarization with Knowledge Disentanglement. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 245–255, New York, NY, USA. Association for Computing Machinery.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-To-Image Translation Using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232.

Barret Zoph. 2022. Designing Effective Sparse Expert Models. In *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*, pages 1044–1044.