How to Handle Different Types of Out-of-Distribution Scenarios in Computational Argumentation? A Comprehensive and Fine-Grained Field Study

Andreas Waldis[*1,2], **Yufang Hou**[1,3], **Iryna Gurevych**[1]
[1]Ubiquitous Knowledge Processing Lab (UKP Lab)
Department of Computer Science and Hessian Center for AI (hessian.AI)
Technical University of Darmstadt
[2]Information Systems Research Lab, Lucerne University of Applied Sciences and Arts
[3]IBM Research Europe - Ireland
www.ukp.tu-darmstadt.de www.hslu.ch

Abstract

The advent of pre-trained Language Models (LMs) has markedly advanced natural language processing, but their efficacy in out-of-distribution (OOD) scenarios remains a significant challenge (Hupkes et al., 2023). The field of computational argumentation (CA), modeling human argumentation processes, is notably impacted by these challenges because complex annotation schemes and high annotation costs naturally lead to resources barely covering the multiplicity of available text sources and topics. Due to this data scarcity, generalization to data from uncovered covariant distributions is a common challenge for CA tasks like stance detection or argument classification. This work systematically assesses LMs' capabilities for such OOD scenarios. While previous work targets specific OOD types like topic shifts (Stab et al., 2018) or OOD uniformly (Yuan et al., 2023), we address three prevalent OOD scenarios in CA: *topic shift, domain shift,* and *language shift*. Our findings challenge the previously asserted general superiority of in-context learning (ICL) for OOD. We find that the efficacy of such learning paradigms varies with the type of OOD. Specifically, while ICL excels for domain shifts with heavy label divergences between train and test data, prompt-based fine-tuning surpasses for shifts when semantic differences prevail, like topic shifts. Navigating the heterogeneity of OOD scenarios in CA, our work empirically underscores the potential of base-sized LMs to overcome these challenges. [1]

1 Introduction

Argumentation as a communication tool for human reasoning has engaged researchers over millennia (Aristotle and Kennedy, ca. 350 B.C.E., translated 2007; Toulmin, 1960; Van Eemeren et al., 2019) and has become an important research area in natural language processing under the umbrella of *computational argumentation* (Lippi and Torroni, 2016; Lauscher et al., 2022). Specifically, computational argumentation (CA) models human argumentative processes and leads to complex tasks such as stance detection (Mohammad et al., 2016) and argument quality evaluation (Toledo et al., 2019). However, developing resources for such CA tasks requires significant annotation efforts (Habernal and Gurevych, 2017; Schiller et al., 2022), which often inadequately capture the wide range of heterogeneity in available text sources and topics. This situation makes OOD scenarios, especially those involving significant *covariant distribution shifts*, a common challenge for CA tasks since LMs are anticipated to generalize across such shifts in current and future applications (Slonim et al., 2021). These shifts occur when input data distribution changes between the training and testing phrases and can be viewed as a specific aspect of out-of-distribution (OOD) scenarios (Zhang et al., 2020).

This work focuses on three types of covariant distribution shifts frequently encountered in CA tasks: *topic shift, domain shift*, and *language shift*.

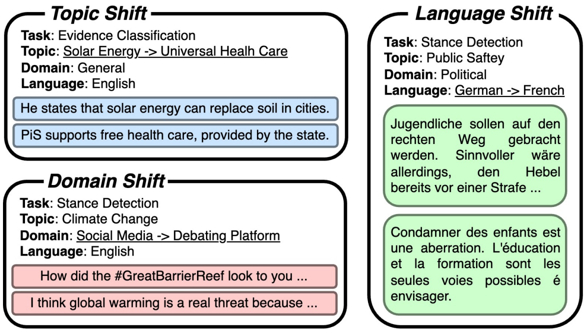

Figure 1: Common OOD types of computational argumentation covering *evidence classification* (Shnarch et al., 2018), and mono or multilingual *stance detection* (Hardalov et al., 2021; Vamvas and Sennrich, 2020).

* Corresponding author andreas.waldis@live.com
[1]We provide data and code at online.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14878–14898
August 11-16, 2024

Figure 1 illustrates these three types of OOD scenarios, in which researchers aimed at developing systems for CA tasks that generalize across unseen topics (Shnarch et al., 2018; Toledo et al., 2019), text domains (Lauscher et al., 2020; Hardalov et al., 2021), or languages (Eger et al., 2018; Vamvas and Sennrich, 2020). These studies have observed that CA systems often fail to handle OOD scenarios.

Given this variety of OOD scenarios in CA and the need for data efficiency, this paper aims to answer the following research question: "*how to handle different types of OOD scenarios in computational argumentation using LMs*"? Most previous work on evaluating the generalization and robustness of NLP models has either predominantly focused on a single type of OOD scenario, such as domain shift (Blitzer et al., 2007; Hardalov et al., 2021), or on general OOD that does not distinguish among various types of OOD scenarios (Yuan et al., 2023). However, these studies overlook the heterogeneous nature of OOD and, thereby, limit the transferability of the corresponding findings to the spectrum of shift types in CA tasks. This study introduces a detailed evaluation framework encompassing a list of holistic performance measures (§ 3) to pinpoint crucial generalization flaws such as misalignment between performance and training loss in models. In addition, we feature a heterogeneous collection of eleven CA tasks (§ 4) covering three types of OOD scenarios. We evaluate these tasks with an extensive experimental setup (§ 5) covering twelve LMs of various sizes and eight learning paradigms, including gradient-based learning like vanilla fine-tuning (FT) and prompt-based fine-tuning (P+FT), as well as in-context learning (ICL). From the observed results (§ 6), we conduct an in-depth analysis (§ 7) to gain a better understanding of how learning paradigms and LMs differ for different types of OOD for computational argumentation.

In contrast to Yuan et al. (2023), suggesting in-context learning (ICL) surpasses fine-tuning LMs for addressing general OOD, we find different learning paradigms excel in different types of OOD for CA tasks. In particular, ICL outperforms gradient-based learning for domain shifts where train and test label distributions heavily differ. However, gradient-based learning surpasses ICL for topic shifts characterized by a clear semantic divergence in the covered topics between the training and testing datasets.

In summary, our main contributions are:

1. We propose an evaluation framework covering eleven CA tasks across three types of OOD scenarios. Along with a comprehensive assessment of LM's OOD capabilities, it provides a clear picture of the generalization challenges in CA and offers guidance to practitioners in tackling these challenges.
2. Results of extensive experiments offer valuable insights and show that different learning paradigms effectively manage OOD scenarios for CA under different conditions. Particularly, in-context learning should be preferred for domain shifts, while gradient-based learning is the first choice for generalization across semantic differences (topic shifts).
3. We shed light on the unused potential of base-sized LMs for OOD scenarios. We demonstrate that training a fraction of the parameters of base-sized LMs with LoRA achieves performance comparable to full LM tuning, and such parameter-efficient training offers better stability than larger LMs.
4. This work emphasizes the critical role of OOD heterogeneity in tackling generalization challenges within CA tasks. This paves the way for future research to conduct detailed and targeted examinations of OOD scenarios in other research areas.

2 Related Work

Out-of-Distribution Generalization Studies in NLP target OOD generalization from different perspectives, focusing on the robustness of LMs (Hendrycks et al., 2019; Jin et al., 2020; Zhou et al., 2020; Wang et al., 2021) or OOD detection (Koner et al., 2021; Cho et al., 2023). Similar to *computer vision* (Tseng et al., 2020), NLP studies primarily focus on considering covariant distribution shifts (Zhang et al., 2020) and analyze single types of them in isolation, such as domain across datasets (Hardalov et al., 2021; Yang et al., 2023; Yuan et al., 2023), language (K et al., 2020; Conneau et al., 2020a), topic (Stab et al., 2018; Allaway and McKeown, 2020). This shortage of comprehensively analyzing OOD hinders analytical or methodological advancements in a challenging field such as *computational argumentation* since generalization of methods is limited when relying on shift-specific

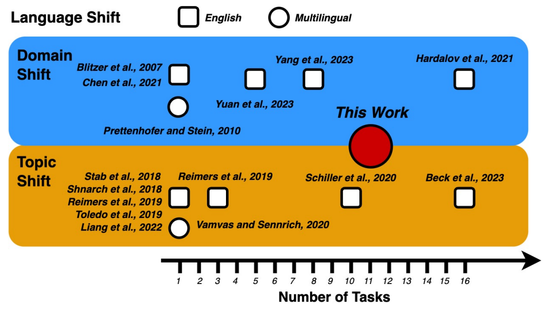

Figure 2: Comparison of our study, covering topic (orange), domain (blue), or language (square/circle) covariant distributions shifts with previous studies that mainly consider single shifts.

features (Liang et al., 2022; Xu et al., 2018; Peng et al., 2018; Rietzler et al., 2020).

Prompt-based Fine-tuning Commonly, pre-trained LMs are fine-tuned by providing a natural language input and optimizing regarding an arbitrary label (Devlin et al., 2019). Instead, prompt-based fine-tuning (Liu et al., 2021a) (P+FT) allows relying upon acquired competencies during pre-training, both for encoding the input and predicting the label by formulating the task as a cloze test. This procedure allows LMs to reach comparable performance to their large-sized counterparts (Schick and Schütze, 2021a,b) with the same limited data as in few-shot settings. Despite their success for few-shot scenarios, little work analyzed how P+FT generalizes differently than FT or how it performs considering complete datasets, exceptionally Raman et al. (2023) showed the robustness of P+FT against adversarial attacks.

With this work (Figure 2), we address the need to comprehensively evaluate OOD abilities of LMs with a particular focus on *computational argumentation*. In particular, we assess a variety of LMs using in-context and gradient-based learning paradigms, considering three types of OOD scenarios covering eleven different tasks.

3 Methodology

3.1 OOD Types

We distinguish between two generalization scenarios: in-distribution (ID) and out-of-distribution (OOD) generalization. While ID assumes train and test data being independent and identically distributed, OOD accounts for practical challenges where we expect apparent distribution shifts between the training and testing instances (Shen et al., 2021). To capture the success of a classifier $f(y|x)$ in such scenarios, we measure its ability to transfer learning from train instances X_{train} to test instances X_{test}. However, OOD potentially introduces covariant, label, and concept shifts between train and test data (Zhang et al., 2020). In this work, we focus on three types of covariant shifts (*topic shift, domain shift*, and *language shift*), as illustrated in Figure 1, due to their frequent prevalence in computational argumentation tasks.

3.2 OOD Evaluation Protocol

Generalization success is typically measured with single task metrics, like the F_1 macro score. However, solely relying on one metric ignores known stability issues, such as apparent deviations regarding randomness (Mosbach et al., 2021). Thus, we compose a set of three requirements that a superior learning model should fulfill: good task performance (*Applicability*), better alignment between optimization and evaluation (*Reliability*), and *Stability* regarding data and randomness. We ground this evaluation for a specific task on a given set of runs ($r \in R$), trained for one distinct fold and seed over a number of epochs ($e \in E$). Note that we formalize these requirements with OOD classification in mind and, therefore, rely on F_1 macro score as the reference metric. However, these requirements can be generalized to other types of OOD tasks using the corresponding reference metrics, such as ROUGE for text generation.

Applicability captures the task-specific performance. Specifically, we measure the average task-specific metric, here F_1 macro score (μ_{F_1}), across all runs r covering different folds and seeds.

Reliability requires that the *learning* process (optimization objective) is reflected in the obtained task *performance*. We evaluate the model using the development dataset that embodies the same OOD type as the training dataset. Specifically, we approximate, after each epoch e of a run r, *learning* as the loss $\beta = \{\forall e \in E(r) | f^{\text{dev}}_{\text{loss}}(e)\}$ and *performance* using task metric (F_1 macro) $\gamma = \{\forall e \in E(r) | f^{\text{dev}}_{F_1}(e)\}$. Then, we calculate the Kendall correlation between β and γ and average it for every r as μ_τ. Ideally, we expect a negative correlation ($\tau = -1$), indicating that improvements in *learning* are reflected in better *performance* and vice-versa. However, since we determine final labels using the argmax operation, dev loss and performance can increase simultaneously. For example, while predicting the same class ($\hat{y} = c_0$) the class probabilities can change from $(95\%, 5\%)$ to $(90\%, 10\%)$. At the same time, the cross-entropy changes from 0.074 to 0.15. Therefore, we assume the model is becoming less sure about the prediction. This aspect is particularly relevant for OOD generalization, where overfitting to distributional properties of training data, such as unique vocabulary, likely introduces uncertainty during inference.

Stability demands a low impact from varying data and randomness on both *Applicability* and *Reliability*. As recommended by Reimers and Gurevych (2017), we measure the standard deviation of σ_{F_1} and $\sigma_{\mathcal{T}}$ across R runs covering different data folds and seeds.

4 CA Tasks Across OOD Types

In this section, we present the selection of computational argumentation tasks (§ 4.1) and subsequently show their heterogeneous distribution shifts, focusing on covariant and label properties (§ 4.2).

4.1 Task Selection

We choose eleven tasks from computational argumentation and related fields (Stede, 2020) that inherent OOD as a fundamental challenge. We broadly categorize them according to their targeted covariant distribution shift, either **topic**, **domain**, or **language**. For example, *domain* for stance detection across datasets (Hardalov et al., 2021), sentiment analysis across *languages* (Prettenhofer and Stein, 2010), or argument quality across *topics* Toledo et al. (2019). Figure 2 compares our study with previous research in terms of the number of tasks and the range of OOD types covered. Below we briefly describe each task:

Argument Quality (*arg-qua*) Toledo et al. (2019) analyzed 9,100 argument pairs across **22 topics** to determine which one has higher quality.

Argument Similarity (*arg-sim*) Reimers et al. (2019) annotated 3,595 arguments pairs of **28 topics** to decide whether they are similar or not.

Argument Classification (*arg-cls*) Stab et al. (2018) annotated the stance of arguments (*pro*, *con*, *neutral*) regarding one of **eight topics**.

Evidence Classification (*evi-cls*) Shnarch et al. (2018) presented 5,785 sentences annotated as relevant or not for one out of **118 topics**.

Sentiment Classification (*review*) Blitzer et al. (2007) annotated 8,000 reviews as positive or negative for **four domains** (Amazon product groups).

Multi-Dataset Stance Detection (*stance*) Following Hardalov et al. (2021), we use the *semeval* (Mohammad et al., 2016), *emergent* (Ferreira and Vlachos, 2016), and *iac* dataset (Walker et al., 2012) to evaluate stance detection across **three domains** (*social media, news*, and *debating*). All of them provide the same labels (*pro*, *con*, *neutral*).

Multi-Dataset Entailment (*entail*) Following Yang et al. (2023), we consider three medium-sized datasets (*rte* (Wang et al., 2018), *SciTail* (Khot et al., 2018), *hans* (McCoy et al., 2019)) to evaluate textual-entailment across **three domains**.

Multi-Lingual Stance Detection (*x-stance*) This dataset (Vamvas and Sennrich, 2020) includes 63,000 multilingual comments (*de*, *fr*, *it*) annotated as *favor* or *against* regarding **12 topics**.

Multi-Lingual Sentiment Classification (*x-review*) Prettenhofer and Stein (2010) presents a set of 43,000 positive or negative reviews covering **four languages** (*de*, *en*, *fr*, *jp*) and **three domains** (Amazon product groups).

While the first seven English-only datasets mentioned above, include annotations for one considered shift (*topic* or *domain*), the selected multilingual datasets come with multiple such annotations. This enables formulating four OOD tasks from two datasets addressing two shift types each: language and domain shifts for *x-review* and topic and language shifts for *x-stance*.

4.2 Tasks Characteristics

In this subsection, we delve into the nature of the distribution shifts embodied by the selected tasks.

Shift Characteristics We focus on covariant properties of the input x (such as semantics) and the label y to describe the characteristic of distribution shifts between training and testing instances. Table 1 show these properties, with higher values denoting increased challenge levels. First, we assess the separability of train and test instances based on their semantic representation. Following Sun

	Shift Type	Separability	Δ Flesch	Δ Words	KL
arg-qua	Top.	78.6	1.5	2.2	0.1
arg-sim	Top.	75.8	4.6	0.27	0.4
arg-cls	Top.	28.7	2.0	0.6	1.6
evi-cls	Top.	56.3	2.4	0.7	7.1
review	Dom.	52.7	6.5	60.5	0.0
stance	Dom.	86.7	2.7	60.8	70.8
entail	Dom.	40.4	5.1	31.2	12.8
x-stance	Lang./Top.	0.05/19.8	16.6/1.3	6.6/0.3	0.6/0.4
x-review	Lang./Dom.	0.07/72.4	11.0/1.8	60.0/6.5	0.0/0.0

Table 1: Distribution shift characteristics between train and test splits of the eleven tasks (averaged across all folds): Separability, differences between train and test instances regarding Flesch score, number of words, and the class distribution (KL divergence).

et al. (2022), we embed[2] all instances and apply k-means clustering (Lloyd, 1982; MacQueen, 1967) to form two clusters. The alignment of these clusters with the train/test split is measured using the adjusted Rand index (Hubert and Arabie, 1985). A higher score suggests a more pronounced semantic shift between train and test sets. Subsequently, we examine biases in surface-level text features introduced during training by calculating differences in average readability (Flesch, 1948) and word count (Δ Flesch, Δ Words) between training and testing instances. Furthermore, we evaluate distributional disparities in class labels using Kullback-Leibler (KL) divergence (Boyd and Vandenberghe, 2004). Higher KL values indicate more pronounced imbalances, complicating the task, as LMs often develop biases towards the training label distribution.

Task Difficulties Drawing from the above analyses, we categorize tasks into distinct groups. *arg-qua*, *arg-sim*, and *stance* demonstrate high semantic differences, with separability scores ranging between 75.8 and 86.7. Tasks like *review*, *stance*, *entail*, and *x-review* present surface-level challenges due to varying readability (Δ Flesch) and text lengths (Δ Word Count). Additionally, tasks such as *evi-cls*, *stance*, and *entail* show notable label distribution imbalances, reflected in high KL divergence values, thereby adding further complexity. Notably, *stance* emerges as particularly challenging, exhibiting distinct semantic, surface form, and label differences between training and testing instances, coupled with significant divergence from the LMs' pre-trained text understanding.

[2]Following Reimers and Gurevych (2019), we use `paraphrase-multilingual-mpnet-base-v2` for embedding.

5 Experimental Setup

This section outlines the experimental setup covering the models, learning paradigms, and evaluation.

Models We primarily experiment with base-sized LMs, including **BERT** (Devlin et al., 2019), **RoBERTa** (Liu et al., 2019), and **DeBERTa-v3** (He et al., 2021b) and their multilingual counterparts (Devlin et al., 2019; Conneau et al., 2020b; He et al., 2021b). For additional experiments, we consider base-sized version of **ALBERT** (Lan et al., 2020), **DeBERTa** (He et al., 2021a), **ELECTRA** (Clark et al., 2020), and 3B version of **T5** (Raffel et al., 2020) and **FLAN-T5** (Chung et al., 2022). Further we consider **GPT-3.5** (Ouyang et al., 2022), **Llama-2-Chat** (70B) (Touvron et al., 2023), and **Orca-2** (13B) (Mitra et al., 2023).

Learning Paradigms We assess the generalization capabilities of LMs under various learning paradigms. This includes vanilla fine-tuning (**FT**), prompt-based fine-tuning (**P+FT**), and in-context learning (**ICL**). Further, we consider linear probing (**LP**) and cloze prompting (**P**) as lower bounds to capture the LM's pre-trained capabilities. In LP and FT, we train task-specific classification heads, in which the LM remains either frozen (LP) or trainable (FT)[3]. For P and P+FT, we embed the input into a cloze and let the pre-trained MLM head to predict the masking token and keep the LM frozen (P) or trainable (P+FT). In addition, we verify scaling gradient-based methods to bigger LMs using parameter efficient methods, including **LoRA** (Hu et al., 2022), **P-Tuning** (Liu et al., 2021b), and **Prompt-Tuning** (Lester et al., 2021). Finally, using ICL, we verify the capabilities of large LMs with task-specific instructions and demonstrations. Appendix § A.5 and § A.6 provide more details about these learning paradigms.

Evaluation We enforce distribution shifts for OOD evaluation by composing train/dev/test splits, including instances with distinct distributional properties, such as unique topics or text domains (Figure 1). We utilize multi-fold cross-validation (CV) to account for data variability and ensure each distinct distributional property (like a unique topic) is tested precisely once [4]. We evaluate all tasks on all learning paradigms, taking LP and P as a

[3]We use [SEP] to concatenate the input with its topic, if available

[4]See Appendix § A.3 for more details.

lower bound and ID fine-tuning (FT-ID) as an upper bound. We assess every task using three random seeds to account for randomness. Using these runs, we employ comprehensive performance measurement including average *Stability* (μ_{F_1}), *Reliability* (μ_τ), and the *Stability* (σ_{F_1},σ_τ) - as previously defined in § 3.2.

6 Results

This section reports results on a detailed (Table 2) and aggregated level (Figure 3) and discusses ***six key findings***.

i) Generalization flaws and the efficacy of prompt-based fine-tuning. The aggregation of the comprehensive evaluation (Figure 3) reveals crucial generalization flaws of OOD fine-tuning (blue). Compared to ID fine-tuning (red), it provides a lower *Applicability* (F_1 score), optimization (loss) and performance (F_1 score) are less aligned (lower *Reliability*), and measurements are less stable across different seeds and folds (*Stability*). In particular, we see this misalignment of loss and performance - a violation of a fundamental generalization assumption - crucially affects vanilla fine-tuning's (FT) degraded OOD generalization capabilities. Turning to prompt-based fine-tuning (P+FT, green), it partially overcomes these flaws. Paired with DeBERTa-v3 and RoBERTa, it achieves higher absolute performance (*Applicability*), a better *Reliability*, and fewer deviations regarding data and randomness (*Stability*).

ii) Superiority of DeBERTa-v3. Next, we focus on the detailed results (Table 2) to compare the different LMs. Overall, we note the superior performance of DeBERTa-v3 compared to RoBERTa and BERT for all learning paradigms across all tasks.[5] In particular, when paired with prompt-based fine-tuning (P+FT), DeBERTa-v3 provides 3.3 better *Applicability* (μ_{F_1}), 2.9 better *Reliability* (μ_τ), and similar *Stability* with -0.4 (σ_{F_1}) and $+0.3$ (σ_τ) than RoBERTa with P+FT. Moreover, we see DeBERTa-v3 with P+FT outperforms FT in ten out of eleven tasks and reaches ID performance for two tasks (*arg-sim* and *review*).

iii) Label differences cause significant generalization gaps. Table 2 reveals significant generalization gaps between OOD (FT and P+FT) and ID

[5]These findings extend to ID scenarios — see Appendix § B.1.

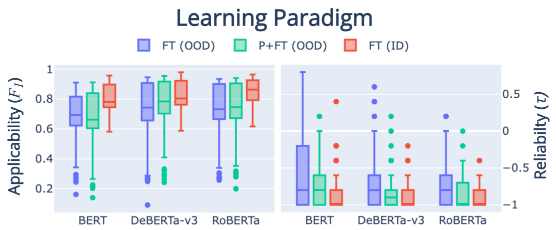

Figure 3: Aggregated results of ID and OOD vanilla fine-tuning (FT-ID and FT) and OOD prompt-based fine-tuning (P+FT) across eleven tasks (§ 4) for *Applicability* (F_1), *Reliability* (τ), and *Stability* (deviation of F_1 and τ).

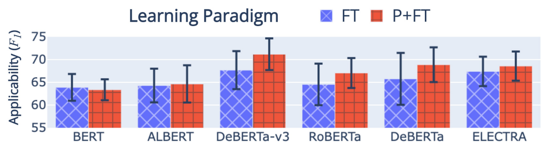

Figure 4: Average *Applicability* of comparing various LMs tuned on the English-only tasks using vanilla fine-tuning (FT) or prompt-based fine-tuning (P+FT).

results (FT-ID) for *stance* and *entail*. These difficulties correlate with their previously identified label differences between train and test instances based on their high KL divergences (Table 1). These generalization issues are also visible when we compare linear probing (LP) and cloze prompting (P) for *stance* and *entail* with other tasks. Since these two paradigms largely evaluate the pre-trained capabilities of LMs, we expect a big gap between them and full LM tuning paradigms (FT and P+FT) when they are done without significant generalization flaws. However, this gap is smaller for *stance* and *entail* than for other tasks, indicating that PF and P+FT exhibit higher generalization problems for these two tasks. Still, we see again that P+FT partially overcomes such generalization flaws and provides, paired with DeBERTa-v3, improvements of 3.4 (*stance*) and 4.6 (*entail*) compared to FT.

iv) Pre-training influences the success of prompt-based fine-tuning. Next, we compare the gap between vanilla fine-tuning (FT) and prompt-based fine-tuning (P+FT) for three additional base-sized LMs to better understand the efficacy of DeBERTa-v3 paired with P+FT. In particular, we focus on its design properties like token-only pre-training objective, disentangled attention (DA), ELECTRA-style training, and extensive vocabulary. To determine which design choice of DeBERTa-v3 has the greatest impact on its superior OOD performance, we evaluate additional LMs on the English-

	arg-qua *Top.*	arg-sim *Top.*	arg-cls *Top.*	evi-cls *Top.*	review *Dom.*	stance *Dom.*	entail *Dom.*	x-stance *Lang./Top.*	x-review *Lang./Dom.*	↑ *Applicability* $\mu_{F_1} \pm \sigma_{F1}$	↓ *Reliability* $\mu_\tau \pm \sigma_\tau$
LP$_{\text{BERT}}$	48.4	57.1	42.7	65.6	81.0	27.9	46.3	52.5/56.7	67.5/73.3	56.3 ± 0.8	-58.4 ± 6.2
P$_{\text{BERT}}$	40.5	50.4	40.1	49.2	72.9	25.0	41.2	34.5/48.6	45.6/54.5	45.7 ± 0.2	-
FT$_{\text{BERT}}$	75.5	68.4	57.5	74.7	89.3	31.1	50.7	62.0/63.9	77.7/84.4	66.8 ± 0.9	-56.8 ± 12.3
P+FT$_{\text{BERT}}$	76.2	66.0	59.8	75.7	89.3	28.5	48.0	59.5/63.6	79.6/83.9	66.4 ± 1.1	-61.7 ± 12.4
FT-ID$_{\text{BERT}}$	87.9	76.4	67.3	78.9	90.4	61.1	93.6	67.6	87.0	78.9 ± 0.4	-96.1 ± 6.5
LP$_{\text{DeBERTa-v3}}$	53.0	70.0	55.1	67.9	88.6	23.4	58.0	55.4/59.7	78.7/83.6	63.0 ± 0.5	-64.3 ± 4.3
P$_{\text{DeBERTa-v3}}$	54.2	58.6	40.3	57.2	61.9	26.5	54.6	51.1/51.2	49.5/52.0	50.6 ± 1.0	-
FT$_{\text{DeBERTa-v3}}$	78.4	75.4	64.0	77.3	93.4	29.6	55.6	**69.8**/69.3	91.3/90.9	72.3 ± 1.1	-72.6 ± 13.4
P+FT$_{\text{DeBERTa-v3}}$	**78.5**	**79.1**†	**74.6**	**78.6**	**94.2**†	**33.0**	**60.2**	69.7/**69.9**	**91.8**/ **91.4**	74.6 ± 0.9	-78.4 ± 8.4
FT-ID$_{\text{DeBERTa-v3}}$	89.0	78.4	75.2	80.6	93.9	63.3	95.4	72.2	92.1	82.2 ± 0.4	-97.7 ± 6.5
LP$_{\text{RoBERTa}}$	51.8	55.3	41.6	62.5	85.7	28.7	39.2	55.1/57.5	82.8/82.5	58.4 ± 0.6	-56.3 ± 6.2
P$_{\text{RoBERTa}}$	48.3	55.3	42.9	51.8	80.5	24.0	40.9	42.4/48.7	67.2/73.4	52.3 ± 0.0	-
FT$_{\text{RoBERTa}}$	70.9	73.0	56.9	77.5	92.2	30.0	51.3	62.2/66.8	89.6/90.1	69.1 ± 2.5	-69.7 ± 10.4
P+FT$_{\text{RoBERTa}}$	77.6	74.3	66.0	77.9	92.0	29.1	52.4	67.4†/67.5†	89.7/90.0	71.3 ± 0.5	-75.5 ± 8.1
FT-ID$_{\text{RoBERTa}}$	84.0	79.4	71.0	80.9	92.9	64.7	94.1	66.3	91.0	80.5 ± 1.9	-96.6 ± 4.7

Table 2: OOD results using linear probing (**LP**), prompting (**P**), vanilla fine-tuning (**FT**), and prompt-based fine-tuning (**P+FT**),and ID fine-tuning (**FT-ID**). We report average *Applicability* (μ_{F_1}), *Reliability* (μ_τ), *Stability* ($\sigma_{F_1}, \sigma_\tau$). The best performance within one LM is underlined, overall is marked in **bold**, and † indicates that OOD surpasses ID.

only tasks involving topic and domain shifts to test these properties (Figure 4). First, we found that DeBERTa(-v3), RoBERTa, and ELECTRA benefit more from prompt-based fine-tuning when pre-trained with token-only objectives (like masked language modeling or replaced token detection). In contrast, LMs such as BERT or ALBERT, trained with additional sentence objectives like next sentence prediction or sentence order prediction, exhibit minor gains or perform worse with P+FT than FT. Second, we do not find DeBERTa-v3 gains from ELECTRA-style pre-training, as the FT vs. P+FT gap is more pronounced for DeBERTa than ELECTRA itself. Third, DeBERTa (with DA) performs better than RoBERTa (without DA) on both FT and P+FT. DeBERTa's disentangled attention (DA) mechanism impacts its superior OOD performance since both models are pre-trained on the same datasets with masked language modeling. Finally, we see DeBERTa-v3's extensive vocabulary (120k tokens) as another factor in its success, as it outperforms its ancestor (DeBERTa) with 50k tokens. These results show how pre-training crucially shapes LMs differently beyond their performance on downstream applications (Wang et al., 2018). We see these insights to be well aligned with other work, particularly in the examination of how the internal representations of LMs vary among different pre-training setups (Waldis et al., 2024).

v) No free lunch for in-context learning or gradient-based methods. We show in Figure 5 results of evaluating English-only tasks using in-context learning (**ICL**) with GPT-3.5 (turbo), Llama-2-chat (70B), and Orca-2 (13B).[6] Comparing these LMs for average *Applicability*, notably Orca-2 (66.2) outperforms GPT-3.5 (64.9) and Llama-2 (60.4). We see this strongly related to the reasoning-oriented pre-training of Orca-2. Moreover, ICL does not reach the average performance level of the best gradient-based (FT and P+FT) approach based on DeBERTa-v3. However, we note the superiority of ICL in scenarios involving heavy domain shifts, particularly in cross-dataset tasks, such as *stance* and *entail*, where heavy differences between train and test label distribution (high **KL** divergence) cause substantial generalization flaws for gradient-based learning methods. These flaws are visible when comparing the gap between P and P+FT. Since we tune the LM for P+FT, we expect a significant gap for successful generalization. However, these gaps are relatively small for *stance* and *entail* - from Table 2 $+6.5$ for *stance* and $+5.6$ for *entail* with DeBERTa-v3 compared to $+34.3$ for *arg-cls*. In addition, there is no clear gain of using P+FT for the relatively easy and popular sentiment analysis task (*review*). Due to its popularity, we assume this task is well covered in the enormous pre-training corpus of large LMs - such as GPT-3.5. In contrast, we note the superiority of P+FT for topic shift scenarios, which predominantly involve challenges of a semantic nature and moderate label distribution differences.

[6]Please find details in the Appendix (§ A.6).

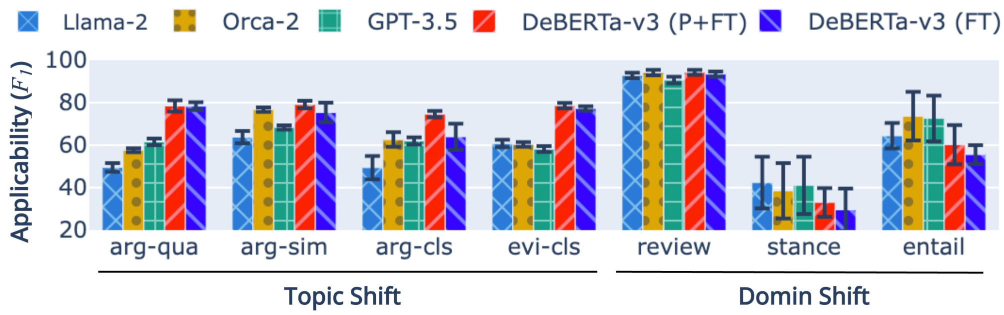

Figure 5: Comparison of ICL using ChatGPT, and DeBERTa-v3 using vanilla fine-tuning (FT) and prompt-based fine-tuning (P+FT).

	arg-qua	arg-sim	arg-cls	evi-cls	review	stance	entail	↑ *Applicability*	↓ *Reliability*
P+FT$_{\text{DeBERTa-v3}}$	78.5	79.1	74.6	78.6	94.2	33.0	60.2	71.2±1.3	**-84.4**±8.3
+P-Tuning	56.3	54.9	38.1	54.7	53.7	**33.7**	43.5	47.8±1.2	-22.0±20.3
+Prompt-Tuning	54.8	54.0	38.2	54.6	53.4	32.3	43.2	47.2±**0.7**	-6.0±30.5
+LoRA	78.1	78.8	73.4	77.9	94.9	33.1	60.8	71.0±1.0	-75.5±**4.9**
***P+FT**$_{\text{DeBERTa-v3 (300m)}}$	81.4	80.0	**78.7**	**79.8**	95.3	31.2	**62.6**	72.7±1.2	-78.1±8.0
***P+FT**$_{\text{T5 (3b)}}$	79.6	80.6	75.7	76.5	**95.7**	26.6	56.6	70.2±0.8	-73.1±17.7
***P+FT**$_{\text{Flan-T5 (3b)}}$	**81.8**	**82.3**	78.5	79.3	96.3	31.0	62.4	**73.1**±1.0	-75.0±22.2

Table 3: Comparison of full-parameter to efficient training using DeBERTa-v3 (rows one to four) and large LMs using LoRA (*) in rows five to seven. Best performance is marked in **bold**.

vi) Few parameters are enough for competitive performance and allow to scale to larger LMs. Next, we compare the performance of different parameter-efficient tuning strategies with full model tuning. As shown in rows two to four in Table 3, we see LoRA with $r = 4$ outperforms P-Tuning and Prompt-Tuning on most tasks. Further, it performs on par with full-parameter tuning (first row) regarding *Applicability*, provides better *Reliability*, but degraded *Stability*. Further experiments considering bigger LMs show that LoRA allows their efficient use for OOD scenarios. Precisely, the large version of DeBERTa-v3 with 300 million provides 1.7 higher *Applicability* and 2.6 better *Reliability* than the base version (86m). Simultaneously, this scaling effect does not continue. T5 or Flan-T5, with three billion parameters, seem to be generally more affected by random seeds and different folds (*Stability*) without apparent *Applicability*. From these observations, OOD fine-tuning still leaves a large potential of LMs unexploited, while larger LMs improve the performance but introduce new *Stability* flaws.

7 Analysis

Next, we discuss ***four aspects*** which differentiate learning paradigms based on *arg-cls* where we observe prominent differences.

i) The bias regarding surface features. We show in Figure 7 average word counts and input complexities of test instances for ID and OOD vanilla fine-tuning (FT-ID and FT) and prompt-based fine-tuning with (P+FT) using DeBERTa-v3 and in-context learning (ICL) with Orca-2 and GPT-3.5. LMs predict shorter and more complex instances (higher Flesch score) more likely correct, and vice versa for wrong ones - compared to the dataset average (dashed line). However, P+FT exhibits less bias on surface correlations than FT and shows similar patterns as FT-ID, hinting at the superior abilities of P+FT. In contrast, ICL predictions, in particular of Orca-2, are less biased for both surface features. Still, deviations from the dataset average suggest fundamental bias in such features.

ii) P+FT provides more prediction confidence and relies less on surface features. From Table 4, P+FT provides higher average confidence (defined as the logit of the predicted label) than FT and a similar one as ID fine-tuning (FT-ID). Thus, we assume P+FT is less confused by the distribution shift, which is also visible in the lower correlation between confidence and surface features (Flesch score or word counts) than FT. For example, FT seems less confident when the input is longer and more complex.

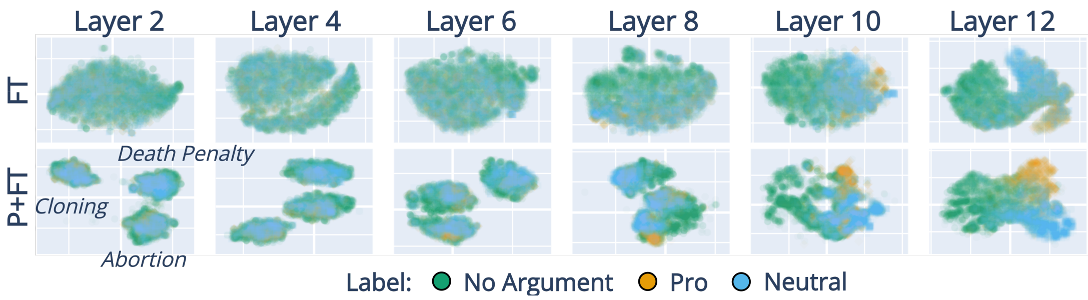

Figure 6: Overview of the T-SNE reduced embeddings of the *CLS* token for FT and *MASK* P+FT for every second layer where instance labels are colorized.

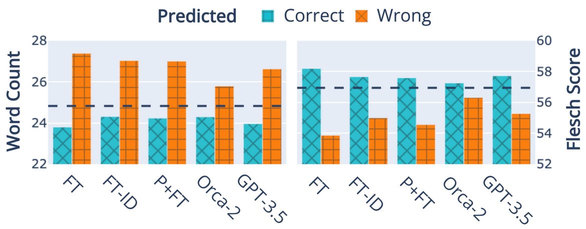

Figure 7: Average word count and input complexity (Flesch score) for correct and wrong predictions for DeBERTa-v3 with ID and OOD vanilla fine-tuning (FT-ID and FT), prompt-based fine-tuning (P+FT) and in-context learning (ICL) using Orca-2, and GPT-3.5.

	FT-ID	FT	P+FT	raw	P+raw
Average Confidence	97.6	95.9	97.8	-	-
Confidence×Flesch	5.1	8.6	4.1	-	-
Confidence×Word Count	-10.3	-13.2	-6.3	-	-
Average Attribution	16.2	15.5	13.0	16.3	13.2
Correct Attribution	16.4	15.8	13.1	-	-
Wrong Attribution	15.2	14.9	12.7	-	-

Table 4: Analysis and correlation (×) of the prediction confidence and token attribution for DeBERTa-v3. *raw* and *P+raw* provide results of the solely pre-trained LM.

iii) Prompt-based fine-tuning considers input differently. Next, we analyze how LMs attribute to the input tokens. We follow Kobayashi et al. (2020) and calculate the attribution of a token using the attention and the norm of the token embeddings. FT-ID and FT have higher average attributions than P+FT, indicating that token attributions are more evenly distributed since they are normalized using the Euclidean norm within a given input sentence. This is already visible when comparing the token attribution before fine-tuning (*raw* vs. *P+raw*). Apparent differences are also visible when we compare how attributions of correct or wrong predicted instances differ. While P+FT shows maximum 0.4 differences (P+FT), this rises to 1.0 for FT-ID. With these results, we assume LMs applied in prompt-based or vanilla fine-tuning fundamentally differ in how inputs are processed.

iv) P+FT retains more semantic information. Figure 6 visualizes the layer-wise embeddings of the classification proxy tokens - *CLS* for FT and *MASK* for P+FT. It shows that P+FT retains more semantic information (about topics) until the last layers, while FT eliminates them across all layers during training. Hinting, again, at substantial differences between FT and P+FT.

8 Conclusion

This work marks the most extensive study to date addressing the heterogeneous types of OOD scenarios in CA by systematically evaluating different OOD types. We evaluate a multiplicity of LMs and learning paradigms on eleven CA tasks. With this extensive evaluation, we shed light on the challenges of having diverse covariant distribution shifts in CA. In addition, we provide takeaways of general relevance, such as the superiority of ICL for domain shifts, where gradient-based learning fails to generalize effectively due to significant label discrepancies between the training and testing data. In contrast, gradient-based learning surpasses ICL when generalization across significant semantic differences is required, like in cases of topic shifts. With the rise of larger LMs, systematic evaluation of distribution shifts becomes even more important, necessitating the consideration of additional factors such as computational efficiency, task complexity, and data contamination. Finally, our findings highlight the untapped potential of base-sized models, which points towards a need for further advancements in gradient-based learning paradigms.

Ethical Considerations and Limitations

8.1 Higher Input Length

By embedding the input into a prompt, we sacrifice potential input tokens. Since the used tasks have relatively short inputs, this is not crucial for this work. However, this can be an essential limitation for other tasks when inputs get longer.

8.2 Efficiency

We always refer to efficient fine-tuning when discussing efficient methods in this work. Therefore, we did not consider efficient methods to make inferences on larger LMs more feasible. We see this as another crucial and essential aspect of real-world applications. Simultaneously, we think performance and efficiency will alternate in the future. Therefore, we keep that for future work.

8.3 Large Language Models

We show the competitive performance of ChatGPT compared to gradient-based approaches by only relying on four demonstrations and without any tuning. Simultaneously, we need to assume that the pre-training corpus of ChatGPT leaks crucial aspects - like broadly covers controversially discussed topics like *Nuclear Energy* or includes instances of popular datasets (like *RTE* (Wang et al., 2018) or *SemEval2016* (Mohammad et al., 2016)) word-by-word. When we have in mind that we use OOD to verify generalization capabilities required for upcoming scenarios, we need to examine the performance of ChatGPT carefully and whether it was able to learn the task or just remembered some semantic aspects of the pre-training.

Acknowledgements

We thank Cecilia Liu, Thy Thy Tran, and Kexin Wang for their valuable feedback. Andreas Waldis has been funded by the Hasler Foundation Grant No. 21024. Yufang Hou is supported by the Visiting Female Professor Programme from TU Darmstadt.

References

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Aristotle and George A. Kennedy. ca. 350 B.C.E., translated 2007. *On Rhetoric: A Theory of Civic Discourse*. Oup Usa.

John Blitzer, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Stephen P Boyd and Lieven Vandenberghe. 2004. *Convex optimization*. Cambridge university press.

Hyunsoo Cho, Choonghyun Park, Junyeop Kim, Hyuhng Joon Kim, Kang Min Yoo, and Sang goo Lee. 2023. Probing out-of-distribution robustness of language models with parameter-efficient transfer learning. *ArXiv*, abs/2301.11660.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8440–8451. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. 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.

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.

Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1163–1168, San Diego, California. Association for Computational Linguistics.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics*, 43(1):125–179.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *EMNLP*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021a. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *CoRR*, abs/2111.09543.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021b. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. 2019. Using pre-training can improve model robustness and uncertainty. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2712–2721. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8018–8025. AAAI Press.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5189–5197. AAAI Press.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. 2021. Oodformer: Out-of-distribution detection transformer. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 209. BMVA Press.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations.

Anne Lauscher, Lily Ng, Courtney Napoles, and Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4563–4574, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Anne Lauscher, Henning Wachsmuth, Iryna Gurevych, and Goran Glavaš. 2022. Scientia potentia est—on the role of knowledge in computational argumentation. *Transactions of the Association for Computational Linguistics*, 10:1392–1422.

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, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3045–3059. Association for Computational Linguistics.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):1–25.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *CoRR*, abs/2107.13586.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT understands, too. *CoRR*, abs/2103.10385.

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

Stuart P. Lloyd. 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

J MacQueen. 1967. Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297. University of California Los Angeles LA USA.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, et al. 2023. Orca 2: Teaching small language models how to reason. *arXiv preprint arXiv:2311.11045*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Minlong Peng, Qi Zhang, Yu-gang Jiang, and Xuanjing Huang. 2018. Cross-domain sentiment classification with target domain specific information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2505–2513, Melbourne, Australia. Association for Computational Linguistics.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1118–1127, Uppsala, Sweden. Association for Computational Linguistics.

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

Mrigank Raman, Pratyush Maini, J. Zico Kolter, Zachary C. Lipton, and Danish Pruthi. 2023. Model-tuning via prompts makes NLP models adversarially robust. *CoRR*, abs/2303.07320.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Copenhagen, Denmark. Association for Computational Linguistics.

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

3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 567–578, Florence, Italy. Association for Computational Linguistics.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4933–4941. European Language Resources Association.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Timo Schick, Helmut Schmid, and Hinrich Schütze. 2020. Automatically identifying words that can serve as labels for few-shot text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5569–5578, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2022. On the effect of sample and topic sizes for argument mining datasets. *arXiv preprint arXiv:2205.11472*.

Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2021. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*.

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize, Yufang Hou, Leshem Choshen, Ranit Aharonov, and Noam Slonim. 2018. Will it blend? blending weak and strong labeled data in a neural network for argumentation mining. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 599–605, Melbourne, Australia. Association for Computational Linguistics.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Manfred Stede. 2020. Automatic argumentation mining and the role of stance and sentiment. *Journal of Argumentation in Context*, 9(1):19–41.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - new datasets and methods. 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 5625–5635, Hong Kong, China. Association for Computational Linguistics.

Stephen E. Toulmin. 1960. The uses of argument.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. 2020. Cross-domain few-shot classification via learned feature-wise transformation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jannis Vamvas and Rico Sennrich. 2020. X -stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS*

2020, Zurich, Switzerland, June 23-25, 2020 [online only], volume 2624 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Frans H Van Eemeren, Rob Grootendorst, and Tjark Kruiger. 2019. *Handbook of argumentation theory: A critical survey of classical backgrounds and modern studies*, volume 7. Walter de Gruyter GmbH & Co KG.

Andreas Waldis, Yufang Hou, and Iryna Gurevych. 2024. Dive into the chasm: Probing the gap between in- and cross-topic generalization. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2197–2214, St. Julian's, Malta. Association for Computational Linguistics.

Marilyn A. Walker, Jean E. Fox Tree, Pranav Anand, Rob Abbott, and Joseph King. 2012. A corpus for research on deliberation and debate. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pages 812–817. European Language Resources Association (ELRA).

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021. Infobert: Improving robustness of language models from an information theoretic perspective. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chang Xu, Cécile Paris, Surya Nepal, and Ross Sparks. 2018. Cross-target stance classification with self-attention networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 778–783, Melbourne, Australia. Association for Computational Linguistics.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750, Toronto, Canada. Association for Computational Linguistics.

Lifan Yuan, Yangyi Chen, Ganqu Cui, Hongcheng Gao, Fangyuan Zou, Xingyi Cheng, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Revisiting out-of-distribution robustness in nlp: Benchmark, analysis, and llms evaluations. *arXiv preprint arXiv:2306.04618*.

Aston Zhang, Zachary Chase Lipton, Mu Li, and Alexander J. Smola. 2020. Dive into deep learning. *Journal of the American College of Radiology : JACR*.

Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian J. McAuley, Ke Xu, and Furu Wei. 2020. BERT loses patience: Fast and robust inference with early exit. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

A Additional Details of the Experiments

A.1 Training Setup

For all our experiments, we use NVIDIA RTX A6000 GPUs with CUDA (11.7), python (3.8.10), transformers (4.28.0), PyTorch (1.13.1), and openprompt (1.0.1).

A.2 Hyperparameters

We use for the experiments fixed hyperparameters; AdamW (Loshchilov and Hutter, 2019) as optimizer; a batch size of 16; a learning rate of 0.00002; a dropout rate of 0.1; a warmup rate of 10% of the steps; random seeds: $[0, 1, 2]$. In the case of parameter-efficient tuning, we use a learning rate of a learning rate of 0.0002. Moreover, we use the following tags from the huggingface model hub:

- `albert-base-v2`
- `bert-base-uncased`
- `aajrami/bert-mlm-base`
- `microsoft/deberta-base`
- `microsoft/deberta-v3-base`
- `roberta-base`
- `google/electra-base-discriminator`
- `t5-3b`
- `google/flan-t5-xl`
- `TheBloke/Llama-2-70B-Chat-AWQ`
- `TheBloke/Orca-2-13B-AWQ`

A.3 Fold Composition

With our evaluation, we want to cover a given dataset fully. Therefore, we conduct a multi-folded evaluation that covers every instance of the dataset once in one of the tests splits X_{test}. For a fair comparison, we use the same number of folds for OOD and ID and synchronize their dimension, i.e., train, dev, and test split of the first fold have the same number of instance for OOD and ID.

We show with Figure 8 an example of a dataset with a topic shift. We colorize topics and indicate train, dev, and test splits with solid, dashed, and dotted lines. First, we sort all dataset instances according to their assigned topic for OOD while we randomly shuffle them for ID. Then we compose the test splits $X^{(\text{OOD})}_{\text{test}}$ and $X^{(\text{ID})}_{\text{test}}$in a way to cover every instance exactly once. Next, we form the train and dev splits by randomly distribution the left-over topics (OOD) or instances (ID) for all folds. When composing these splits, we compose the ID splits to match the respective OOD splits of the same fold. For example considering the first fold, the splits $X^{(\text{OOD})}_{\text{train-1}}$, $X^{(\text{OOD})}_{\text{dev-1}}$, and $X^{(\text{OOD})}_{\text{test-1}}$ have the same number of instances as the splits $X^{(\text{ID})}_{\text{train-1}}$, $X^{(\text{ID})}_{\text{dev-1}}$, and $X^{(\text{ID})}_{\text{test-1}}$.

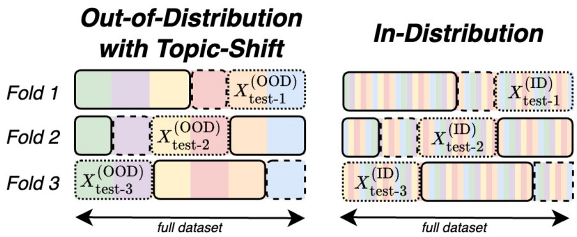

Figure 8: Example of the composition of the different folds when we target the topic shift of a dataset with three folds. Distinct topics are colorized, while solid, dashed, or dotted lines indicate train, dev, and test splits.

Based on the number of unique distribution shift properties (topics, domains, or languages), we use the different number of folds to distribute these properties as even as possible across the different test splits $X^{(\text{OOD})}_{\text{test}}$. Therefore, we use whenever possible a three-folded setup. However, when the number of distribution properties is equal to four (i.e., four domains), we conduct a four-folded evaluation. Please find this concrete number of folds per task in the source code.

A.4 Dataset Details

As a part of this work, we propose eleven different OOD classification tasks based on 13 different datasets. In the following, we provide additional details. Table 5 shows an overview of these tasks and examples for every task. Furthermore, we show in Figure 9 how these task examples diverge from the LMs' pre-trained textual understanding based on Wikipedia, which is a major pre-training dataset for BERT, RoBERTa, and DeBERTa. Specifically, we compute the pseudo perplexity (Salazar et al., 2020), determined as the cross-entropy of each token, for 500 randomly chosen instances per task. For English tasks, `bert-base-uncased` is used, while `bert-base-multilingual-uncased` is

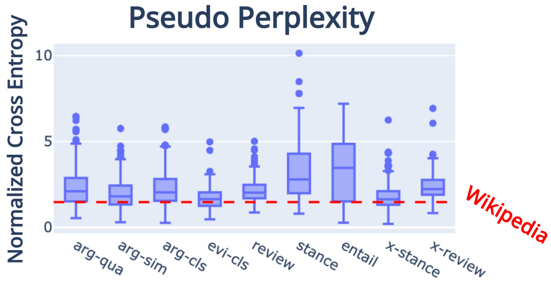

Figure 9: Pseudo perplexity of the selected tasks compared to pre-training data from Wikipedia (red line).

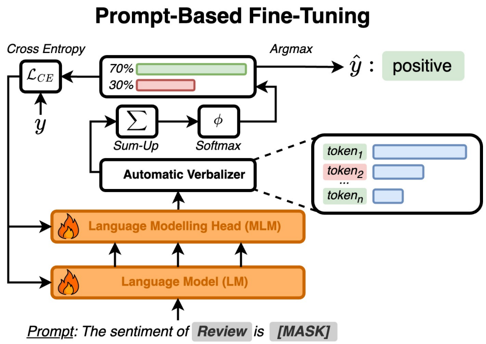

Figure 10: An exemplary overview of prompt-based fine-tuning (P+FT) for the sentiment classification task, following Schick and Schütze (2021a): re-formulating the task as cloze (prompt); gathering relevant tokens (verbalizer) for the specific classes - *positive* (green) or *negative* (red); finding the final prediction by summing up the probability of relevant tokens per class; backpropagating the error through the MLM head and the LM.

deployed for multilingual tasks. We then compare the averaged cross-entropy of these tasks (refer to Figure 9) against 500 randomly selected samples from the Wikipedia pre-training corpus (Devlin et al., 2019). The comparison indicates a notable divergence in the chosen tasks from a common pre-training dataset Wikipedia, except for *evi-cls*, which leverages Wikipedia data, and *x-stance*, which aligns closely with Wikipedia's text genre.

Finally, Table 6 lists the prompt templates used for all tasks and languages.

A.5 Prompt-Based Fine-Tuning

In this paper, we adopt prompt-based fine-tuning (**P+FT**) (Liu et al., 2021a) as an alternative approach to vanilla fine-tuning (**FT**). Unlike FT, P+FT relies on the pre-trained masked-language modeling (MLM) head and avoids using new classification heads.

Forward Pass We show in Figure 10 an exemplary overview of P+FT for sentiment analysis given two classes $K = \{positive, negative\}$. In detail, we wrap the review with a cloze template and add a masking token as the prediction proxy. Next, the LM processes this prompt and outputs the most probable tokens T along with their log probabilities L using the MLM head. Then, the verbalizer selects the relevant tokens A within L and assigns them a class mapping - like *positive* (green) or *negative* (red). In contrast to other probing work (Schick and Schütze, 2021a; Raman et al., 2023), we automatically select indicative tokens Schick et al. (2020) using the likelihood ratio regarding every class k in K based on the train instances. With these token-class mappings, we sum up the log-probabilities for every class k in K as $w_k = \sum_{a \in A(k)} L(a)$ and apply the softmax (Equation 1) to find $\hat{y}$.

$$\hat{y} = \arg\max_{k \in K} \frac{exp(w_k)}{\sum_{k' \in K} exp(w_{k'})} \quad (1)$$

Backward Pass While the forward pass represents the prompting paradigm (**P**), we analyze and evaluate LMs without parameter optimization. However, to fine-tune the LM and MLM head, we calculate the cross-entropy loss $\mathcal{L}_{CE}$ and update the weights through back-propagation. Note that we initialized the automatic verbalizer before the training and did not update it anymore afterward.

A.6 In-Context Learning Setup

As reported in § 6, we evaluated ChatGPT using in-context learning (ICL). In detail, we provide four demonstration samples from the training instances for every test instance. We use the templates reported in Table 7 for every training demonstration instance and the test instance, where we exclude the LABEL for the test one. For a fair comparison with gradient-based approaches (FT, P+FT), we allow to sample these demonstration instances from the entire training set. We use BM25 to calculate the similarity between the test and train instances. Afterward, we use the top-4 most similar train instances as a demonstration for a given test instance.

B Additional Results

In addition to results shown in the main paper (§ 6), we show in the following the effectiveness of P+FT for ID scenarios (§ B.1) and that other methods to prevent freshly initialized classification heads underperforms prompt-based fine-tuning (§ B.2).

B.1 In-Distribution Results

Table 8 shows the superior performance of prompt-based fine-tuning transfers to ID scenarios.

B.2 Classification Head Pre-Initialisation

In addition to prompt-based fine-tuning (**P+FT**), we experimented with pre-initializing the classification head using a linear probe (**LP+FT**) following Kumar et al. (2022). As reported in Table 9, we did not find a positive effect of using LP+FT.

	Dataset	Description	Distribution Shift
arg-qua	Argument Quality (Toledo et al., 2019)	Choose which argument out of two has the higher quality: TOPIC : *we should ban fossil fuels* ARG-1 : *fossil fuels pollute and cause a lot of diseases* ARG-2 : *fossil fuel companies often have incredibly bad and dangerous working conditions* LABEL : ARG-1	**Topical** (22 topics)
arg-sim	Argument Similarity (Reimers et al., 2019)	Decide whether two arguments are similar or not-similar : TOPIC : *organ donating* ARG-1 : *One organ and tissue donation can save or enhance the lives of nearly 100 people* ARG-2 : *By donating your organs after you die, you can save or improve as many as 50 lives* LABEL : similar	**Topical** (28 topics)
arg-cls	Argument Classification (Stab et al., 2018)	Classify an argument as pro , con , or no-argument given a topic: TOPIC : *abortion* ARG : *Now our nonprofit really needs your help* LABEL : similar	**Topical** (8 topics)
evi-cls	Evidence Classification (Shnarch et al., 2018)	Decide whether a text is relevant evidence for a topic or not-relevant TOPIC : *we should limit executive compensation* TEXT : *On April 7, 2009, Blankfein recommended guidelines to overhaul executive compensation* LABEL : not-relevant	**Topical** (118 topics)
review	Sentiment Classification (Blitzer et al., 2007)	Classify product review as positive or negative : DOMAIN : *dvd* REVIEW : *If you don't own this dvd ... my opinion it is the best american animated film ever released* LABEL : positive	**Domain** books, dvd, electronics, kitchen & housewares
stance	Stance Detection	Classify a text as either pro , con , or neutral regarding a topic: TOPIC : *climate change is a real concern* TEXT : *Be kind to the earth beneath your feet. #environment* LABEL : pro	**Domain** News (Ferreira and Vlachos, 2016) Debating (Walker et al., 2012) Social Media (Mohammad et al., 2016)
entail	Entailment	Predict whether two sentences do entail or not-ential each other: DOMAIN : RTE SENTENCE-1 : *No Weapons of Mass Destruction Found in Iraq Yet* SENTENCE-2 : *Weapons of Mass Destruction Found in Iraq* LABEL : not entail	**Domain** RTE (Wang et al., 2018) SciTail (Khot et al., 2018) HANS (McCoy et al., 2019)
x-review	Multilingual Sentiment Classification (Prettenhofer and Stein, 2010)	Classify product review as positive or negative : DOMAIN : *books* LANGUAE : *de* REVIEW : *Ich war vor 5 Jahren in Indien ... Ich kann dieses Buch nur empfehlen.* LABEL : positive	**Domain** books, dvd, music **Lingual** de, en, fr, jp
x-stance	Multilingual Stance Detection (Vamvas and Sennrich, 2020)	Classify a text as either favor , or against regarding a given topic: TOPIC : *encomonmy* LANGUAE : *it* TEXT : *Non penso che tale ampliamento sia necessario, né urgente.* LABEL : against	**Domain** books, dvd, music **Lingual** de, fr, it

Table 5: Overview and examples of the used datasets and information about the enforced distribution shift.

Task	Prompt
arg-qua	ARG-1 is MASK than ARG-2 regarding TOPIC
arg-sim	ARG-1 is MASK than ARG-2 regarding TOPIC
arg-cls	The attitude of ARG is MASK regarding TOPIC
evi-cls	TEXT is MASK evidence regarding TOPIC
review	The sentiment of REVIEW is MASK
stance	The attitude of TEXT is MASK regarding TOPIC
entail	SENTENCE-1 ? MASK , SENTENCE-2
x-stance	de: Die Haltung von ARG ist MASK zu TOPIC fr: L'attitude de ARG est MASK envers TOPIC it: L'atteggiamento di ARG MASK verso TOPIC
x-review	de: Die Stimmung von REVIEW ist MASK en: The sentiment of REVIEW is MASK fr: Le sentiment de REVIEW est MASK jp: REVIEW の感情は MASK です

Table 6: Overview of the used prompt templates for all tasks and languages for the prompt-tuning setup.

Task	Prompt
arg-qua	Given the following two arguments and the topic they cover, which one has the higher quality? Options are first or second. Argument 1: ARG-1 Argument 2: ARG-2 : Topic: TOPIC Label: LABEL
arg-sim	Are the following arguments similar regarding the given topic? Options are yes or no. Argument 1: ARG-1 Argument 2: ARG-2 : Topic: TOPIC Label: LABEL
arg-cls	What is the attitude of the following argument regarding the given topic? Options are neutral, favor, or against. Argument: ARG Topic: TOPIC Label: LABEL
evi-cls	Corresponds the following evidence to the given topic? Options are yes or no. Evidence: TEXT Topic: TOPIC Label: LABEL
review	What is the sentiment of the following text? Options are positive or negative. Review: TEXT Label: LABEL
stance	What is the attitude of the following text regarding the given topic? Options are neutral, favor, or against. Text : TEXT Topic: TOPIC Label: LABEL
entail	Can we conclude an entailment from the following two texts? Options are yes or no. Text 1: TEXT-1 Text 2: TEXT-2 : Topic: TOPIC Label: LABEL

Table 7: Overview of the used prompting templates for the in-context learning setup.

	arg-qua	arg-sim	arg-cls	evi-cls	review	stance	entail	x-stance	x-review	*Applicability*	*Reliability*
	Top.	*Top.*	*Top.*	*Top.*	*Dom.*	*Dom.*	*Dom.*	*Lang./Top.*	*Lang./Dom.*	$\mu_{F_1} \pm \sigma_{F1}$	$\mu_\tau \pm \sigma_\tau$
LP$_{\text{BERT}}$	55.7	69.9	58.5	70.4	85.5	55.3	72.1	58.9	76.3	67.0 ± 0.2	-71.3 ± 3.8
P$_{\text{BERT}}$	47.7	50.4	36.5	54.7	60.2	44.7	49.2	49.2	57.2	48.7 ± 0.0	-
FT$_{\text{BERT}}$	87.9	76.4	67.3	78.9	90.4	61.1	93.4	67.6	87.0	78.9 ± 0.4	-83.7 ± 6.5
P+FT$_{\text{BERT}}$	88.0	76.1	67.7	79.1	90.4	62.8	93.4	67.0	87.0	79.1 ± 0.3	-78.7 ± 8.1
LP$_{\text{DeBERTa-v3}}$	55.1	72.5	60.3	71.1	89.3	53.2	87.1	59.6	85.5	70.4 ± 0.1	-74.6 ± 3.0
P$_{\text{DeBERTa-v3}}$	55.1	60.5	41.7	61.5	63.3	46.1	57.8	52.2	53.4	54.6 ± 0.5	-
FT$_{\text{DeBERTa-v3}}$	89.0	78.4	75.2	80.6	93.9	63.3	96.7	72.5	92.1	82.4 ± 0.4	-92.3 ± 6.5
P+FT$_{\text{DeBERTa-v3}}$	90.3	81.5	78.9	81.5	94.8	70.1	96.5	71.4	92.3	84.1 ± 0.3	-91.0 ± 7.0
LP$_{\text{RoBERTa}}$	54.0	67.0	57.8	69.6	88.4	53.3	73.0	59.4	86.2	67.6 ± 0.2	-80.6 ± 3.1
P$_{\text{RoBERTa}}$	54.9	57.1	45.3	54.7	79.5	46.3	55.6	49.2	76.1	58.0 ± 0.0	-
FT$_{\text{RoBERTa}}$	84.0	79.4	71.0	80.9	92.9	64.7	94.9	58.6	91.0	79.7 ± 1.9	-90.0 ± 4.7
P+FT$_{\text{RoBERTa}}$	88.2	79.6	72.5	80.8	92.7	67.0	95.2	69.8	91.1	81.9 ± 0.3	-85.7 ± 6.3

Table 8: In-distribution (ID) results for BERT, DeBERTa-v3, and RoBERTa using linear probing (**LP**), prompting (**P**), fine-tuning (**FT**), and prompt-based fine-tuning (**P+FT**). We report average *Applicability* (μ_{F_1}), *Reliability* (μ_τ), *Stability* ($\sigma_{F_1}, \sigma_\tau$). Best OOD performance within one LM are underlined and **bold** highlights best OOD performance across LMs.

	arg-qua	arg-sim	arg-cls	evi-cls	review	stance	entail	x-stance	x-review	*Applicability*	*Reliability*
	Top.	*Top.*	*Top.*	*Top.*	*Dom.*	*Dom.*	*Dom.*	*Lang./Top.*	*Lang./Dom.*	$\mu_{F_1} \pm \sigma_{F1}$	$\mu_\tau \pm \sigma_\tau$
FT$_{\text{BERT}}$	75.5	68.4	57.5	74.7	89.3	31.1	50.7	62.0/63.9	77.7/84.4	66.8 ± 0.9	-56.8 ± 12.3
LP+FT$_{\text{BERT}}$	75.7	66.5	57.3	74.1	89.3	34.2	50.4	60.8/64.1	77.1/84.0	66.7 ± 1.4	-56.5 ± 14.3
P+FT$_{\text{BERT}}$	76.2	66.0	59.8	75.7	89.3	28.5	48.0	59.5/63.6	79.6/83.9	66.4 ± 1.1	-61.7 ± 12.4
FT$_{\text{DeBERTa-v3}}$	78.4	75.4	64.0	77.3	93.4	29.6	55.6	69.8/69.3	91.3/90.9	72.3 ± 1.1	-72.6 ± 13.4
LP+FT$_{\text{DeBERTa-v3}}$	78.4	75.6	63.7	76.5	93.6	30.1	54.7	69.6/69.1	91.1/91.1	72.1 ± 1.3	-70.8 ± 11.9
P+FT$_{\text{DeBERTa-v3}}$	78.5	79.1	74.6	78.6	94.2	33.0	60.2	69.7/69.9	91.8/91.4	74.6 ± 0.9	-78.4 ± 8.4
FT$_{\text{RoBERTa}}$	70.9	73.0	56.9	77.5	92.2	30.0	51.3	62.2/66.8	89.6/90.1	69.1 ± 2.5	-69.7 ± 10.4
LP+FT$_{\text{RoBERTa}}$	76.0	73.9	54.3	77.2	92.1	27.3	47.6	62.3/67.0	89.1/89.2	68.7 ± 1.7	-71.0 ± 11.7
P+FT$_{\text{RoBERTa}}$	77.6	74.3	66.0	77.9	92.0	29.1	52.4	67.4/67.5	89.7/90.0	71.3 ± 0.5	-75.5 ± 8.1

Table 9: Comparing vanilla (**FT**), linear-probing fine-tuning afterward (**LP+FT**), and prompt-based fine-tuning (**P+FT**) for BERT, DeBERTa-v3, and RoBERTa. We report average *Applicability* (μ_{F_1}), *Reliability* (μ_τ), *Stability* ($\sigma_{F_1}, \sigma_\tau$).