Dive into the Chasm: Probing the Gap between In- and Cross-Topic Generalization

Andreas Waldis[*1,2], **Yufang Hou**[3,1], **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

Pre-trained language models (LMs) perform well in In-Topic setups, where training and testing data come from the same topics. However, they face challenges in Cross-Topic scenarios where testing data is derived from distinct topics - such as *Gun Control*. This study analyzes various LMs with three probing-based experiments to shed light on the reasons behind the In- vs. Cross-Topic generalization gap. Thereby, we demonstrate, for the first time, that generalization gaps and the robustness of the embedding space vary significantly across LMs. Additionally, we assess larger LMs and underscore the relevance of our analysis for recent models. Overall, diverse pre-training objectives, architectural regularization, or data deduplication contribute to more robust LMs and diminish generalization gaps. Our research contributes to a deeper understanding and comparison of language models across different generalization scenarios. [1]

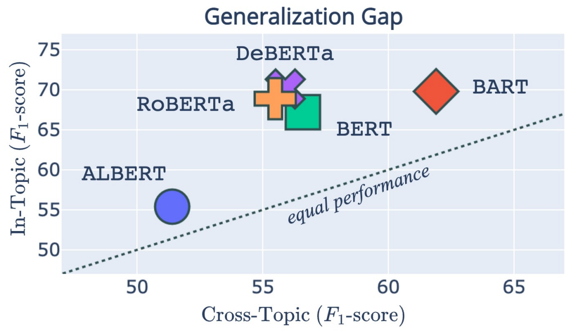

Figure 1: Generalization gap of fine-tuning LMs on argumentative *stance detection* (Stab et al., 2018) in the In- or Cross-Topic evaluation setup. The dashed line marks the ideal case of equal performance.

1 Introduction

Probing (Belinkov et al., 2017; Conneau et al., 2018a) is widely used to analyze pre-trained language models (LMs) (Devlin et al., 2019; Liu et al., 2019; He et al., 2021; Radford et al., 2019). It enables a better understanding of how LMs encode information and how it evolves in the architecture by studying linguistic properties such as part-of-speech or dependency-tree parsing (Tenney et al., 2019a,b). However, probing methods (Hewitt and Liang, 2019a; Hewitt and Manning, 2019; Voita and Titov, 2020a; Elazar et al., 2021) mainly rely on the general In-Distribution (ID) scenario, where we distribute train and test instances independent and identically. As a result, other more realistic Out-of-Distribution (OOD) scenarios (Shen et al., 2021), like generalizations regarding forthcoming topics or temporal changes in the language, remain underexplored by probing.

Addressing this research gap, we propose - for the first time - a probing-based approach to comprehensively analyze LMs in a challenging OOD setup. More precisely, we rely on Cross-Topic[2] evaluation where we deliberately withhold instances from specific topics for testing. Following (Habernal and Gurevych, 2016; Stab et al., 2018), we define *topic* as the query used to compose a specific dataset - such as arguments covering *gun control* or *marijuana legalization*. This evaluation setup is highly relevant for challenging Argument Mining (AM) downstream tasks (Slonim et al., 2021). It allows for simulating, in a controlled setup, how well LMs handle topic-shifts when unseen semantic features (such as topic-specific vocabulary) arise in future and new topics. Previous studies found that Cross-Topic argument mining is challenging compared to the In-Topic setup (Stab et al., 2018; Waldis and Gurevych, 2023). The major reason lies in the apparent generalization gaps between randomly composing training and testing data (In-Topic) and using distinct groups of topics for training and testing

* Corresponding author andreas.waldis@live.com
[1]We provide data and code at https://github.com/UKPLab/eacl2024-cross-topic-probing.

[2]Also known as Cross-Target in *Stance Detection* research.

Findings of the Association for Computational Linguistics: EACL 2024, pages 2197–2214
March 17-22, 2024

(Cross-Topic). Figure 1 shows such performance gap when fine-tuning on the *UKP ArgMin* dataset (Stab et al., 2018) - labeling arguments as in favor, against, or neutral to one of eight topics. Notably, we observe gaps between In- and Cross-Topic varying considerably across LMs - with BART outperforming the others in the Cross-Topic setup.

Such inconsistencies underline the need to investigate such crucial generalization capabilities. Thus, we propose extensive probing-based experiments to examine the gap between In- and Cross-Topic generalization and show that embedding spaces of LMs vary considerably regarding their generalizability and robustness. In detail, we propose three probing-based experiments to answer the following research questions, considering three linguistic probes (dependency-tree parsing, part-of-speech tagging, and named-entity recognition) based on *UKP ArgMin* dataset:

How do generalization gaps of LMs differ after pre-training? (§ 4) We find generalization gaps substantially differ across LMs while becoming more prominent for tasks with more semantically difficulties, such as NER. In addition, we crucially observe that probing generally underperforms on lexical unseen instances (like highly rare entities), and deduplicating pre-training data provides more robust embedding space when evaluating larger and more recent LMs.

How do LMs depend on topic-specific vocabulary? (§ 5) Next, we assess the influence of topic-specific tokens by removing them using amnesic probing and LMs significantly differing in their reliance on and robustness concerning such semantic features. Interestingly, pre-training objectives or architectural regularization influence robustness, suggesting their potential importance in building robust and competitive LMs.

How do generalization gaps evolve during fine-tuning? (§ 6) Finally, we re-probe tuned LMs on the *UKP ArgMin* dataset and find that In-Topic fine-tuning erases more linguistic properties than Cross-Topic fine-tuning.

To sum up, we expand the probing scope to Cross-Topic generalization and highlight probing as a universal tool complementing the study of language models beyond general evaluation setups. While we focus on an in-depth analysis of In- vs. Cross-Topic generalization gaps, our experimental setup generalizes to other types of OOD scenarios where one verifies generalization regarding other text genres (like the *social media* domain), languages, or temporal changes in the languages (Conneau et al., 2018b; Hardalov et al., 2021; Röttger and Pierrehumbert, 2021; Yang et al., 2023).

2 In- and Cross-Topic Probing

The following section formally outlines the probing setup and tasks before elaborating on the generalization gap and comparing the evaluation of In- and Cross-Topic probing.

2.1 Probing Setup and Tasks

We define a probe f_p comprised of a frozen encoder h and linear classifier c without any intermediate layer. This classifier is trained to map instances $X = \{x_1, \ldots, x_n\}$ to targets $Y = \{y_1, \ldots, y_n\}$ for a given probing task. Using a frozen LM as h, the probe converts x_i into a vector h_i. In detail, we encode the entire sentence, which wraps x_i, and average relevant positions of x_i to find h_i. Relevant positions for the considered probing task are either single tokens for *part-of-speech tagging (POS)*), a span for *named entity recognition (NER)*, or the concatenation of two tokens for *dependency tree parsing (DEP)*. Then, the classifier c utilizes h_i to generate a prediction $\hat{y}_i$, as shown in Equation 1.

$$\hat{y}_i = f_p(x_i) = c(h(x_i)) \tag{1}$$

2.2 Generalization Gap

Generalization gaps arise when comparing evaluation setups focusing on different capabilities for the same task. This work focuses on gaps in using data from the same (In-Topic) or different topics (Cross-Topic) for training and testing. We define such topics $T = \{t_1, \ldots, t_m\}$ as the query to collect instances and thereby given by specific datasets (Habernal and Gurevych, 2016; Stab et al., 2018) - such as arguments covering *gun control* or *marijuana legalization*. The In- vs. Cross-Topic gap is visible in Figure 2, which shows how NER instances (in blue) are distributed in the semantic space. For Cross-Topic, entities cover only specific topics and thereby are less broadly spread, while In-Topic ones are spread more broadly since they cover all datasets' topics. Simultaneously, we note more lexically *unseen* entities (in red) during training for Cross-Topic. Ideally, generalization gaps do not exist since pre-trained language models (LMs)

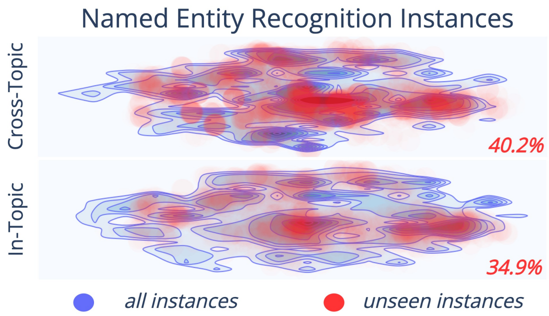

Figure 2: Density plot of In- and Cross-Topic NER test instances (blue), encoded with *bert-base-uncased* and reduced with the same t-SNE model (van der Maaten and Hinton, 2008). While the number of instances is the same, Cross-Topic embodies, with 40.2%, more *unseen* instances than In-Topic (34.9%).

overcome such distribution shifts between different evaluation setups. However, practically, these gaps vary for different models (Figure 1).

2.3 Difference between In- and Cross-Topic Evaluation

By evaluating probing tasks for In- and Cross-Topic, we examine the varying generalization gaps between these setups across different LMs.

Cross-Topic With Cross-Topic evaluation, we investigate how well a probe generalizes when the train, dev, and test instances cover distinct sets of topics $\{T^{(train)}, T^{(dev)}, T^{(test)}\}$. A probe f_p must generalize across the distribution shift in this setup. This shift originates because distinct topics cover different specific vocabulary Z - i.e., $Z_{(test)}$ for topics in $T^{(test)}$. We formally describe this shift, denoted as ΔZ, as the relative complement between topic-specific vocabulary from train and test instances - $\Delta Z = Z_{(train)} \setminus Z_{(test)}$. For Cross-Topic, we expect ΔZ to be large (Figure 2).

In-Topic In contrast, ΔZ is smaller for the In-Topic setup because instances from every split (train/dev/test) cover the same topics. We expect similar topic distribution and minor semantic differences within these splits compared to Cross-Topic (Figure 2). Thus, we see fewer difficulties for In-Topic because a classifier does not need to generalize across a large distribution shift ΔZ.

Topic-Specific Vocabulary As discussed previously, we see topic-specific vocabulary as one main reason for generalization gaps between In- and Cross-Topic because ΔZ differs for these setups considering a dataset d covering topics $T = t_1, \ldots, t_m$. The topic-specificity of a token z_i is a latently encoded property within the encodings h_i for a token w_i. To capture this property on the token level, we adopt the approach of Kawintiranon and Singh (2021) and use the maximum log-odds-ratio r_i of a token regarding a set of topics T. Firstly, we calculate the odds of finding the token w_i in a topic t_j as $o_{(w_i,t_j)} = \frac{n(w_i,t_j)}{n(\neg w_i,t_j)}$, where $n(w_i, t_j)$ is the number of occurrences of w_i in t_j, and $n(\neg w_i, t_j)$ is the number of occurrences of every other token $\neg w_i$ in t_j. We then compute r as the maximum log-odds ratio of w_i for all topics in T as $r_{(w_i,T)} = max_{t_j \in T}(log(\frac{o(w_i,t_j)}{o_{(w_i,\neg t_j)}}))$.

Model	# Params	Objectives	Data
`ALBERT` (Lan et al., 2020)	12M	MLM + SOP	16GB
`BART` (Lewis et al., 2020)	121M	DAE	160GB
`BERT` (Devlin et al., 2019)	110M	MLM + NSP	16GB
`DeBERTa` (He et al., 2021)	100M	MLM	80GB
`RoBERTa` (Liu et al., 2019)	110M	MLM	160GB
`ELECTRA` (Clark et al., 2020)	110M	MLM+DISC	16GB
`GPT-2` (Radford et al., 2019)	117M	LM	40GB

Table 1: Overview of the used LMs trained on MLM, LM, DISC, NSP, SOP, or DAE objectives.

3 Experimental Setup

We propose three experiments to analyze the varying generalization gap between LMs after pre-training (§ 4), their dependence on topic-specific vocabulary (§ 5), and the evolution of these gaps during fine-tuning (§ 6). We outline general details about these experiments, while details and results are provided in the subsequent sections.

Models We examine how various LMs (Table 1) with varying pre-training objectives or architectural designs differ regarding our probing tasks. We cover LMs pre-trained using masked language modeling (MLM), next sentence prediction (NSP), sentence order prediction (SOP), language modeling (LM), discriminator (DISC), and denoising autoencoder (DAE) objectives. As in previous work (Koto et al., 2021), we group them into the ones pre-trained using token- (MLM) and sentence-objectives (NSP, SOP, or DAE) and four purely token-based pre-trained (MLM, LM, DISC). We consider the base-sized variations to compare their specialties in a controlled setup. Apart from these seven contextualized LMs, we use a static LM with `GloVe` (Pennington et al., 2014).

Data We require a dataset with distinguishable topic annotations to evaluate probing tasks in the In- and Cross-Topic evaluation setup. Therefore, we mainly[3] rely on the *UKP ArgMin* dataset (Stab et al., 2018), which provides 25,492 arguments annotated for their argumentative stance (*pro*, *con*, or *neutral*) towards one of eight distinct topics like *Nuclear Energy* or *Gun Control*. Using these instances, we heuristically generate at most 40,000 instances for the three linguistic properties *dependency tree parsing* (**DEP**), *part-of-speech tagging* (**POS**), or *named entity recognition* (**NER**) using spaCy.[4] Additionally, we consider the main task of the *UKP ArgMin* dataset (Stab et al., 2018) - *argumentative stance detection* (**Stance**). Therefore, we have a topic-dependent reference probe to relate the results of other probes and evaluate the generalization ability of LMs on real-world tasks after pre-training. We use a three-folded setup for all these four probing tasks to consider the full data variability for both In- and Cross-Topic evaluation. Details about the compositions of these folds and how we ensure a fair comparison between In- and Cross-Topic are provided in the Appendix (§ A.2) as well as examples for probing tasks (Appendix § A.1).

Evaluation We primarily report the macro F_1 score averaged over the results of evaluating each of the three folds three times using different random seeds. Following recent work (Voita and Titov, 2020b; Pimentel et al., 2020), we additionally report information compression I (Voita and Titov, 2020b) for a holistic evaluation. It measures the effectiveness of a probe as the ratio ($\frac{u}{mdl}$) between uniform code length $u = n * log_2(K)$ and minimum description length mdl, where u denotes how many bits are needed to encode n instances with label space of K. We follow *online* variation of mdl and use the same ten-time steps $t_{1:11} = \{\frac{1}{1024}, \frac{1}{512}, ..., \frac{1}{2}\}$, where we train a probe for every t_j with a fraction of instances and evaluate with the same fraction of non-overlapping instances. Exemplary, for, t_9 we use the first fraction of $\frac{1}{4}$ instances to train and another fraction of $\frac{1}{4}$ to evaluate. We find the final mdl as the sum of the evaluation losses of every time step $t_{1:11}$. For Cross-Topic, we group training instances into two groups of distinct topics and sample the same fraction of instances to train and evaluate. Thus, we ensure a similar distribution shift between training and evaluation fractions as in all instances.

[3]We verified our findings with another dataset in the Appendix § B.1.

[4]We show in the Appendix (§ B.8) that the heuristically generated labels are reliable, and our results are well aligned with previous work.

	DEP		**POS**		**NER**		**Stance**		*Average*		
	In	*Cross*	*In*	*Cross*	*In*	*Cross*	*In*	*Cross*	*In*	*Cross*	Δ
ALBERT	**43.8**	**39.5**	**80.2**	**78.0**	**48.6**	45.8	54.8	**45.9**	***56.9***	***52.3***	*-4.6*
BART	36.5	36.9	75.4	74.1	**48.7**	**45.3**	**60.8**	44.4	*55.3*	*50.2*	*-5.1*
BERT	25.4	25.6	68.5	67.5	45.4	41.6	56.9	43.0	*49.0*	*44.4*	*-4.6*
DeBERTa	32.8	29.9	73.7	74.6	**48.8**	42.4	**59.8**	**45.8**	*53.4*	*48.2*	*-5.2*
RoBERTa	25.1	23.6	64.0	65.5	**48.4**	42.1	51.8	40.1	*47.3*	*42.8*	*-4.5*
ELECTRA	33.6	33.6	75.3	75.3	41.5	41.2	46.6	43.1	*49.3*	*48.3*	***-1.0***
GPT-2	25.2	23.9	63.5	61.9	45.5	38.6	51.1	38.4	*46.3*	*40.7*	*-5.6*
GloVe	12.1	11.9	26.5	26.2	43.4	37.5	41.6	34.1	*30.9*	*27.4*	*-3.5*
Avg. Δ	*-1.2*		*-0.5*		*-4.5*		*-11.0*		-	-	-

Table 2: In- and Cross-Topic probing results for eight LMs. We report the macro F_1 over three random seeds, the average difference between the two setups (last row), and their average per LM (last three columns). The best results within a gap of 1.0 are marked by columns.

4 The Generalization Gap of LMs

The first experiment shows that the generalization gap already exists after pre-training and varies regarding specific LMs and probing tasks. We analyze general (Table 2 and Figure 3) and fine-grained (Table 3) results and discuss them for the different evaluating setups, probing tasks, and LMs. While firstly focusing on mid-size LMs usable for fine-tuning, we close how probing performance scales to large LMs in § 4.

Design We probe eight LMs on the probing tasks DEP, POS, NER, and Stance and verify them by observing significant performance drains using random initialized LMs (Appendix § B.2). For a holistic evaluation, we provide general results and group instances into two categories: *seen* and *unseen*. We define *seen* instances as already processed during training but in another context. For example, the pronoun *he* might appear in both training and test data, but in distinct sentences. By evaluating the LMs on *seen* instances, we gain insights into the influence of token-level lexical information versus context information from surrounding tokens. In contrast, *unseen* instances were not encountered during the training of a probe. They allow assessing whether LMs generalize to tokens that are similar to some extent (such as *Berlin* and *Washington*) but not seen during training.

Results for Evaluation Setups Upon analyzing Table 2, we observe clear generalization gaps between In- and Cross-Topic evaluation for all tasks and LMs. As in Figure 3, the magnitude of this gap

		DEP			POS			NER		
		all	Δ *seen*	Δ *unseen*	*all*	Δ *seen*	Δ *unseen*	*all*	Δ *seen*	Δ *unseen*
	Instance Ratio	-	85%	15%	-	86%	14%	-	65%	35%
In-Topic	ALBERT	43.8	+0.21	-3.2	80.2	+0.41	-17.7	48.6	+1.1	-5.8
	BART	36.5	+0.13	-3.0	75.4	+0.20	-16.5	48.7	+1.3	-7.0
	BERT	25.4	-0.02	-0.8	68.5	+0.20	-16.5	45.4	+1.0	-5.8
	DeBERTa	32.8	+0.07	-1.5	73.7	+0.09	-12.7	48.8	+1.0	-5.6
	RoBERTa	25.1	-0.01	-0.9	64.0	-0.04	-15.5	48.4	+1.0	-5.7
	Average	-	*-0.08*	*-1.9*	-	*+0.17*	*-15.8*	-	*+1.1*	*-6.0*
	Instance Ratio	-	78%	22%	-	81%	19%	-	51%	49%
Cross-Topic	ALBERT	39.5	+0.03	-2.3	78.0	+0.51	-12.9	45.8	+2.2	-5.3
	BART	36.9	+0.01	-4.0	74.1	+0.24	-16.5	45.3	+2.4	-5.8
	BERT	25.6	-0.09	-0.7	67.5	+0.20	-14.0	41.6	+1.9	-5.1
	DeBERTa	29.9	-0.07	-1.3	74.6	+0.14	-11.7	42.4	+2.0	-5.2
	RoBERTa	23.6	-0.22	-0.3	65.5	+0.00	-14.7	42.1	+1.9	-5.2
	Average	-	*-0.08*	*-1.7*	-	*+0.22*	*-14.0*	-	*+2.1*	*-5.3*

Table 3: Performance difference of *seen* and *unseen* instances compared to the full set (*all*). We report for ALBERT, BART, BERT, DeBERTa, & RoBERTa, and include the ratio of *seen* and *unseen* instances.

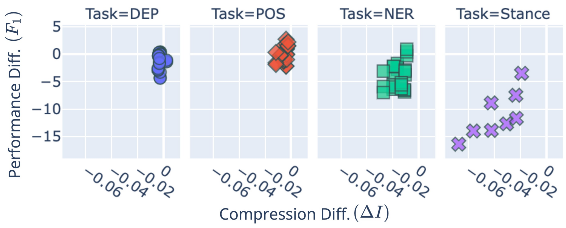

Figure 3: Comparision of the difference in ΔF_1 and ΔI between Cross-Topic and In-Topic for all eight LMs on the four probing tasks.

(ΔF_1) correlates with the difference in compression (ΔI). Interestingly, we find a stronger correlation between F_1 and I for Cross-Topic ($\rho = 0.72$) as compared to In-Topic ($\rho = 0.69$). Thus, a higher performance level, like for In-Topic, leaves less room for compression improvements.

Further, we examine the performance of *seen* and *unseen* instances in Table 3. It shows that *seen* performs slightly better than *all*, while *unseen* ones underperform the complete set (*all*) and *seen* instances. Considering the average over LMs, there are fewer relative gains for *seen* for In-Topic and more loss for *unseen* instances (+1.2, -6.0 for NER) compared to Cross-Topic (+2.0, -5.3 for NER). This observation relates to the lower percentage of *unseen* instances (i.e., made of topic-specific terms) for In- compared to Cross-Topic. We see *unseen* instances on In-Topic are harder and cover rare vocabulary, and *seen* instances on Cross-Topic are easier and made of general terms - which confirm our theoretical and semantic assumptions (§ 2).

Results for Probing Tasks Considering Table 2 and Figure 3, we note higher generalization gaps (*Avg.* Δ of -4.5 and -11.0) for semantic tasks (NER and Stance) than for syntactic ones (DEP and POS) - *Avg.* Δ of -1.2 and -0.5. We verify this trend with results by observing a more pronounced gap for semantic NER classes (like ORG) than for syntactic ones (like ORDINAL) in the Appendix (§ B.5).

Next, we separately compare tasks for *seen* and *unseen* instances. DEP shows the slightest performance difference compared to *all*. We assume that the pairwise nature of the task leads to a larger shared vocabulary between *unseen* and training instances - since a pair can be *unseen*, but it may contain a frequent word like *of*. In contrast, apparent differences between NER and POS are visible - with less performance drain on *unseen* instances for NER than POS. Therefore, we assume for NER a higher semantic overlap with training instances since they could include - as being an n-gram - words from the training vocabulary. In contrast, tokens of *unseen* POS instances are always single words; thus, we assume a smaller semantic overlap with the training.

Results for Encoding Models We now compare LMs amongst themselves. The four best-performing LMs of In-Topic differ up to 7.6 (ALBERT - BERT), while for Cross-Topic, this difference narrows to 4.1 (ALBERT - ELECTRA). These results confirm the varying generalization gap between them and, again, that we can not transfer conclusions from one evaluation setup to another. For example, the probing performance of BART for In-Topic Stance is the best and the third best for Cross-Topic.

Generally, we do not see a clear correlation between better average performance and a smaller generalization gap. LMs like DeBERTa perform better for In- and Cross-Topic but show a bigger gap (-5.1) compared to lower performing LMs like ELECTRA (-1.0), but there are also worse LMs with a bigger gap (GPT-2, -5.6) or better ones with a smaller gap (ALBERT, -4.6). Overall, we see the generalization gap being more pronounced for better-performing LMs.

Considering absolute performance, ALBERT and BART performs the best for both evaluation setups, while ELECTRA excels POS and DEP, and DeBERTa performs for NER and Stance. In contrast, BERT, RoBERTa, GPT-2, and GloVeunderperform the others. Thus, LMs with architectural regularization, such as layer-wise parameter sharing (ALBERT), encoder-decoder layers (BART), disentangled attention (DeBERTa), or discriminator (ELECTRA), tend to provide

	DEP		POS		NER		Stance		*Average*		
	In	*Cross*	*In*	*Cross*	*In*	*Cross*	*In*	*Cross*	*In*	*Cross*	Δ
ALBERT	43.8	39.5	**80.2**	78.0	48.6	45.8	54.8	45.9	*56.9*	*52.3*	*-4.6*
BART	36.5	36.9	75.4	74.1	48.7	45.3	60.8	44.4	*55.3*	*50.2*	*-5.1*
PYTHIA (12B)	38.3	35.4	79.5	77.7	57.3	50.5	65.2	41.6	*60.1*	*51.3*	*-8.8*
PYTHIA-DD (12B)	**45.3**	**45.4**	79.8	79.2	**64.5**	**55.8**	66.1	**50.4**	***63.4***	***57.9***	*-6.2*
LLAMA-2 (13B)	**44.4**	41.8	**81.0**	**80.6**	48.7	45.3	**66.8**	44.2	*60.2*	*53.0*	*-7.2*
LLAMA-2 Chat (13B)	**45.4**	41.7	**80.7**	**80.1**	49.2	42.9	**67.2**	43.2	*60.6*	*52.0*	*-8.7*

Table 4: Results (macro F_1) of the four probing tasks using the two overall best-performing LMs (ALBERT and BART) in the In- and Cross-Topic setup based on the *ArgMin* dataset (Table 2) and three large LMs.

higher Cross-Topic performance. Similarly, ALBERTor DeBERTagenerally achieve more performance gains for *seen* instances and fewer performance drops for *unseen* ones than models without regularization such as BERT or RoBERTa. We hypothesize that architectural and regularization aspects give LMs a more generalizable and robust encoding space.

Results for Larger Models We compare in Table 4 four open accessible large LMs with the two best performing models (ALBERT and BART). In general, we see the performance scales with the higher number of parameters, but more noticeable for In- than Cross-Topic tasks. Therefore, the generalization gap of large LMs tend to be bigger than for LMs. Regarding the different large LMs, PYTHIA (Biderman et al., 2023) and LLAMA-2 (Touvron et al., 2023) outperform the others on In-Topic tasks while performing on par with ALBERT. Further, we notice data deduplication during pre-training (PYTHIA-DD) results in the best performing model and actively reduces the generalization gap from 8.8 to 6.2. In addition, instruction fine-tuning does not heavily affect the performance but tends to increase the generalization gap from 7.2 (LLAMA-2) to 8.7 (LLAMA-2 Chat).

5 The Dependence on Topic-Specific Vocabulary

To this point, we saw that the generalization gap varies between different LMs and probing tasks. Since topic-specific vocabulary crucially affects generalization gaps, we analyze the varying dependence on the topic-specific vocabulary of LMs using *Amnesic Probing* (Elazar et al., 2021). We observe apparent differences among LMs and assume their embedding space clearly differs beyond single evaluation metrics. Therefore, we emphasize considering various LMs when using *Amnesic Probing*. Additional insights of comparing *seen* and *unseen* instance and distinct NER classes are provided in the Appendix (§ B.4, § B.6).

Design To measure how LMs depend on topic-specific vocabulary, we employ *Amnesic Probing* (Elazar et al., 2021) to remove the latently encoded topic-specificity z_i from the embeddings h_i of a token w_i. More precisely, we compare how the performance of a probing task (like NER) changes when we remove z_i. A more negative effect indicates a higher dependence on topic-specific vocabulary, while this property is a hurdle when performance improves. We first train a linear model on token-level topic-specificity r (§ 2.3). To shape it as a classification task, we categorize r into three classes (*low*, *medium*, *high*). [5] Next, we find a projection matrix P that projects all embeddings h_i - gathered as H - using the learned weights W_l of l to the null space as $W_lPH = 0$. Using P we update h_i by neutralizing topic-specificity from the input as $h_i^{'} = Ph_i$ before training the probe. Following (Elazar et al., 2021), we verified our results by measuring less effect of removing random information from h_i (see Appendix § B.3).

Results Considering Figure 4, we see ALBERT, BART, and BERT depend less on topic-specific vocabulary. Their diverse pre-training (token- and sentence-objectives or sentence denoising) results in a more robust embedding space. Surprisingly, they show positive effects (3.2 for DEP for BART) when removing topic-specificity. This could remove potentially disturbing parts of the embedding space. Similarly, GPT-2 is less affected by the removal - we assume this is due to its generally lower performance level. Therefore, it has less room for performance drain, and capturing topic-specificity is less powerful.

Comparing In- and Cross-Topic setups shows a narrowing generalization gap for more affected models (like RoBERTa and GloVe on NER or NER). Simultaneously, less affected LMs either maintain the gap or enlarge it slightly - like BART on DEP, NER, or NER. Further, DeBERTa, RoBERTa, ELECTRA, and GloVe rely more on topic-specific vocabulary since they show significant performance loss (up to 34.6 for GloVe on POS) when removing this information. Specifically, GloVe as a static language model, and RoBERTa is affected the highest for all tasks. ELECTRA shows similar behavior but is less pro-

[5] Please find examples in the Appendix § A.6.

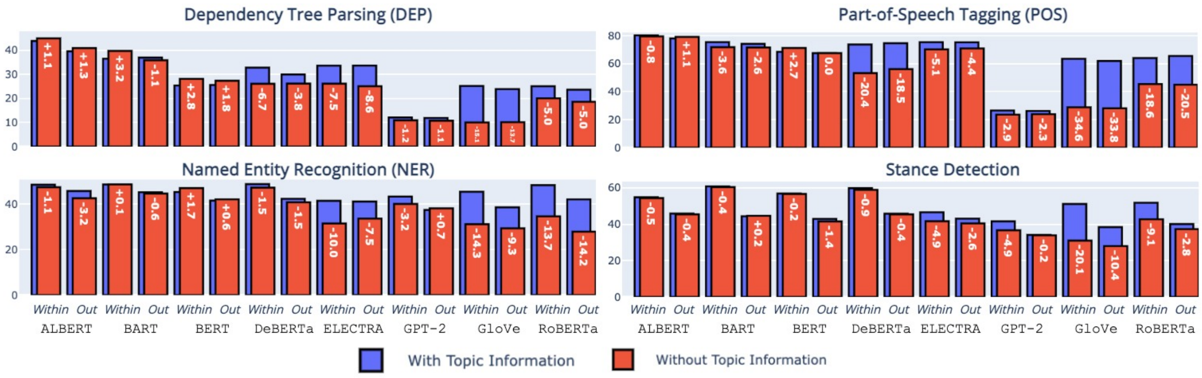

Figure 4: Comparison of the probing results with (blue bars) or without (red bars) topic information. The white text indicates the difference between these two scenarios ($\Delta F_1^{\setminus T}$).

nounced for POS. Thus, its reconstruction pre-training objective provides a more robust embedding space than purely MLM (DeBERTa or RoBERTa). Comparing DeBERTa and RoBERTa, DeBERTa is less affected by the removal of semantic tasks (NER and NER). We hypothesize that distinguishing between token content and token position via disentangled attention makes DeBERTa more robust for the semantic than for syntactic tasks (DEP and POS).

		Stance	DEP	POS	NER	*Avg.*	DEP	POS	NER
		F_1 *fine-tuned*	ΔF_1 *probing*				$\Delta F_1^{\setminus T}$		
In-Topic	ALBERT	55.4 +0.6	-27.3	-40.2	-25.0	-30.8	-0.6	-3.0	-0.1
	BART	69.8 +9.0	-17.3	-32.2	-4.0	-17.8	-0.8	-4.0	+0.3
	BERT	67.2 +10.3	-7.5	-24.8	+1.0	-10.4	+0.4	+0.7	+1.1
	DeBERTa	**70.1 +10.3**	-13.2	-25.3	-8.8	-15.8	-0.8	-3.8	-0.4
	RoBERTa	68.9 +17.1	-19.7	-48.6	-29.7	-27.2	-0.8	-3.0	-0.7
	Avg.	66.3 +9.5	-16.6	-32.6	-12.1	-20.4	-0.5	-2.6	+0.1
Cross-Topic	ALBERT	51.4 +5.5	-14.4	-20.3	-12.6	-15.8	+1.6	-1.3	+2.1
	BART	**61.9 +17.5**	-16.5	-33.9	-5.4	-18.6	-1.0	-3.5	-1.6
	BERT	56.6 +13.6	-5.7	-19.5	+0.6	-8.2	+0.7	+0.6	+1.2
	DeBERTa	55.9 +10.1	-13.4	-33.4	-11.8	-19.5	-1.2	-8.6	+1.6
	RoBERTa	55.5 +15.4	-16.6	-48.3	-23.1	-23.5	-1.9	-4.8	-0.3
	Avg.	56.3 +12.6	-13.0	-29.3	-9.1	-17.1	-0.4	-3.5	+0.6

Table 5: Results of evaluating our probing setup on fine-tuned LMs on NER. The first column shows these fine-tuned results and the gained improvement compared to probing for NER on pre-trained LMs (Table 2). Next, we show performance differences between pre-trained and fine-tuned LMs (ΔF_1 *probing*) and how removing topic-specificity affects the fine-tuned LMs ($\Delta F_1^{\setminus T}$).

6 The Evolution of the Generalization Gap during Fine-Tuning

Finally, we re-evaluate fine-tuned LMs using our proposed probing setups and show that fine-tuning leads to a drain in probing performance. We use these results to retrace apparent differences between evaluation setups and the varying generalization gap between LMs. This is relevant for a broader understanding of how fine-tuning affects LMs (Mosbach et al., 2020; Kumar et al., 2022a), and what they learn during fine-tuning (Merendi et al., 2022; Ravichander et al., 2021).

Design We fine-tune the LMs on an argumentative *stance detection* task and re-evaluate them on DEP, POS, and NER probing tasks. To be consistent with our probing setup, we used the same folds for fine-tuning. Further details are in the Appendix (§ A.5). We compare these results with the probing performance of their pre-trained counterparts (§ 4 and § 5) and correlate this change with the generalization gap observed on the downstream task. We limit our analysis to ALBERT, BERT, BART, DeBERTa, and RoBERTa.

Results Table 5 shows that fine-tuning clearly boost the performance on NER compared to the probing performance (§ 4) but leads to a clear performance drop (ΔF_1) for both evaluation setups and the probing tasks. Cross-Topic achieved more gains on average (+12.6) and fewer drains (-17.1) on the three linguistic properties than In-Topic (+9.5, -20.4). On average, we assume that In-Topic fine-tuning affects the encoding space of LMs more heavily than Cross-Topic. Regarding the different probing tasks, the performance drain is more pronounced for syntactic tasks (DEP and POS) than semantic tasks (NER). This hints that LMs acquire competencies of a semantic nature - which holds for *stance detection*. Similarly, removing topic-specificity influences fine-tuned LMs the least for NER. At the same time, this removal is more pronounced for Cross-Topic. This confirms the assumption that the Cross-Topic setup has smaller effects on LMs internals since we saw big impacts of this removal (§ 5).

Considering the single LMs, we see apparent differences. For example, ALBERT, with its shared architecture and priorly best-performing LM, experiences big probing performance drains and the smallest fine-tuning gains (+0.6, +5.5). In contrast, we note effective fine-tuning of BERTwith +10.3 for In- and +13.6 for Cross-Topic, and that it lost the least probing performance. Comparing RoBERTa and DeBERTa reveals again the effectiveness of architectural regularization of DeBERTa. RoBERTa shows the most gains when fine-tuning on NER and almost catching up with DeBERTa. However, it experiences a more clear performance drain (-27.2, -23.5) regarding the probing tasks for In- and Cross-Topic compared to DeBERTa (-15.8, -19.5). Next, we focus on BART and its superior Cross-Topic performance on NER. It seems already well-equipped for this downstream task due to its high In-Topic probing performance on NER. Therefore, it can learn the task more robustly during fine-tuning.

7 Related Work

The rise of LMs (Devlin et al., 2019; Liu et al., 2019; Radford et al., 2019; He et al., 2021) enabled big success on a wide range of tasks (Wang et al., 2018, 2019). Nevertheless, they still fall behind on more realistic Cross-Topic, like generalizing towards unseen topics (Stab et al., 2018; Gulrajani and Lopez-Paz, 2021; Allaway and McKeown, 2020). One primary reason is that LMs often rely on unwanted spurious correlations. Despite LMs seeing such vocabulary during pre-training, they failed to consider test vocabulary in the required fine-grained way (Thorn Jakobsen et al., 2021; Reuver et al., 2021). Further, Kumar et al. (2022b) found linear models can outperform fine-tuning LMs when considering out-of-distribution data. Thus, a broader understanding of LMs in challenging evaluation setups is crucial.

Probing (Belinkov et al., 2017; Conneau et al., 2018a; Peters et al., 2018) helps to analyze inners of LMs. This includes to examine how linguistic (Tenney et al., 2019a,c), numeric (Wallace et al., 2019), reasoning (Talmor et al., 2020), or discourse (Koto et al., 2021) properties are encoded. Other works focus on specific properties used for other tasks (Elazar et al., 2021; Lasri et al., 2022), or fine-tuning dynamics (Merchant et al., 2020; Zhou and Srikumar, 2022; Kumar et al., 2022b). However, these works target the commonly used *In-Topic* setup and less work considering Cross-Topic setups. Aghazadeh et al. (2022) analyzed metaphors across domains and language, or Zhu et al. (2022) cross-distribution probing for visual tasks. They found that models generalize to some extent across distribution shifts in probing-based evaluation. Nevertheless, these works focus on specialized tasks and consider the generalizations across distributions in isolation. In contrast, we propose with our experiments a more holistic probing-based evaluation of LMs, covering different generalization aspects after pre-training and fine-tuning.

8 Conclusion

Discussion We analyzed and compared In- and Cross-Topic evaluation setups and found generalization gaps significantly differing regarding specific LMs and probing tasks.[6] Further, we make various crucial observations contributing to a better understanding of the generalizability of LMs: **(1)** diverse pre-training objectives and architectural regularization tend to positively affect the robustness of LMs and their embedding space, such as depending less on topic-specific vocabulary; **(2)** probing performance falls short for rare vocabulary, underscoring the need to explore token-level properties; **(3)** probing performance, but also generalization gaps, tend to scale for larger LMs, while deduplication of pre-training data improves their robustness and narrows these gaps; and **(4)** In-Topic fine-tuning tend to vanish linguistic properties more prominently than for the Cross-Topic setup.

To conclude, we highlight the practical utility of probing to analyze and compare the capacities of various LMs from a different perspective - considering different generalization scenarios. Thereby, our work points out the importance of probing as a universally applicable method, regardless of size or being static or contextualized, to complement existing work on analyzing language models (Wang et al., 2018; Liang et al., 2022).

Outlook With our findings in mind, we regularly see probing LMs and large LMs and consider forthcoming learning paradigms as indispensable for a holistic evaluation of their verity and multiplicity. Therefore, we will continue to analyze language models, including a broader set of tasks and focus-

[6]We verified our results using a second dataset from the social media domain (Conforti et al., 2020) - details in the Appendix § B.1.

ing on general and rare vocabulary to increase our understanding of how, why, and where they differ.

Acknowledgements

We thank Irina Bigoulaeva, Tim Baumgärtner, Tilman Beck, and the anonymous reviewers for their valuable feedback. This work has been funded by the Hasler Foundation Grant No. 21024.

Ethical Considerations and Limitations

Automatic Annotations for Linguistic Properties Our experiments require all instances origin in the same datasets with topic annotations. Thanks to this condition, we align all our experiments, like probing LMs, with the same data as they got pre-trained. Therefore, we minimize other influences like semantic shifts of other datasets. However, there are no corresponding annotations for linguistic properties, which forces us to rely on automatically gathered annotations. This work addresses this issue by transparently stating the libraries and models we used to derive these annotations and providing the source code and the extracted labels in our repository. We compared our results (§ B.8) with previous work (Tenney et al., 2019a,c; Hewitt and Liang, 2019b) and found our results well aligned. Further, we verify the probing task results on the different LMs with randomly initialized counter-parts (§ B.2) and confirm our findings with a second dataset (§ B.1).

Definition of Topic-Specific Vocabulary This work considers a topic as a semantic grouping provided by a given dataset. As previously mentioned, this focus on the context of one dataset allows in-depth and controlled analysis, like examining the change of LMs during fine-tuning. On the other hand, we need to re-evaluate other datasets since the semantic space and granularity of the topic are different in almost every other dataset. Nevertheless, results in the Appendix (§ B.1) let us assume that our findings correlate with other datasets and domains. Further, we consider only token-level specific vocabulary, as done previously in literature (Kawintiranon and Singh, 2021). We think that considering n-grams could give a better approximation of topic-specific terms. Still, we do not consider them because *Amnesic Probing* (Elazar et al., 2021) require token-level properties to apply resulting intervention on token-level tasks like POS.

Impact of LMs Design choices This work analyzes LMs regarding different properties like pre-training objectives or architectural regularization. However, we do not claim the completeness of these aspects nor a clear causal relationship. Making such a final causal statement would require significant computational resources to pre-train models to verify single properties with full certainty. Instead, we use same-sized model variations, evaluate all probes on three folds and three random seeds to account for data variability and random processes, and verify our results on a second dataset. Nevertheless, we use them to correlate results on aggregated properties (like having diverse pre-training objectives or not) and not on single aspects, like the usefulness of the *Sentence-Order* objective.

References

Ehsan Aghazadeh, Mohsen Fayyaz, and Yadollah Yaghoobzadeh. 2022. Metaphors in pre-trained language models: Probing and generalization across datasets and languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2037–2050, Dublin, Ireland. Association for Computational Linguistics.

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.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

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.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on Twitter. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018a. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. 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.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Ishaan Gulrajani and David Lopez-Paz. 2021. In search of lost domain generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. 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 9011–9028. Association for Computational Linguistics.

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

John Hewitt and Percy Liang. 2019a. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2733–2743. Association for Computational Linguistics.

John Hewitt and Percy Liang. 2019b. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4129–4138. Association for Computational Linguistics.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge enhanced masked language model for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4725–4735, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3849–3864. Association for Computational Linguistics.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022a. 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.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. 2022b. 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. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

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.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yüksekgönül, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models. *CoRR*, abs/2211.09110.

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. *CoRR*, abs/1907.11692.

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.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Federica Merendi, Felice Dell'Orletta, and Giulia Venturi. 2022. On the nature of BERT: correlating fine-tuning and linguistic competence. In *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, pages 3109–3119. International Committee on Computational Linguistics.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. 2020. On the Interplay Between Fine-tuning and Sentence-level Probing for Linguistic Knowledge in Pre-trained Transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2502–2516, Online. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018. Dissecting contextual word embeddings: Architecture and representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1499–1509, Brussels, Belgium. Association for Computational Linguistics.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4609–4622, Online. Association for Computational Linguistics.

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

Abhilasha Ravichander, Yonatan Belinkov, and Eduard H. Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3363–3377. Association for Computational Linguistics.

Myrthe Reuver, Suzan Verberne, Roser Morante, and Antske Fokkens. 2021. Is stance detection topic-independent and cross-topic generalizable? - a reproduction study. In *Proceedings of the 8th Workshop on Argument Mining*, pages 46–56, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Röttger and Janet Pierrehumbert. 2021. Temporal adaptation of BERT and performance on downstream document classification: Insights from social media. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2400–2412, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, Liat Ein-Dor, Roni Friedman-Melamed, Assaf Gavron, Ariel Gera, Martin Gleize, Shai Gretz, Dan Gutfreund, Alon Halfon, Daniel Hershcovich, Ron Hoory, Yufang Hou, Shay Hummel, Michal Jacovi, Charles Jochim, Yoav Kantor, Yoav Katz, David Konopnicki, Zvi Kons, Lili Kotlerman, Dalia Krieger, Dan Lahav, Tamar Lavee, Ran Levy, Naftali Liberman, Yosi Mass, Amir Menczel, Shachar Mirkin, Guy Moshkowich, Shila Ofek-Koifman, Matan Orbach, Ella Rabinovich, Ruty Rinott, Slava Shechtman, Dafna Sheinwald, Eyal Shnarch, Ilya Shnayderman, Aya Soffer, Artem Spector, Benjamin Sznajder, Assaf Toledo, Orith Toledo-Ronen, Elad Venezian, and Ranit Aharonov. 2021. An autonomous debating system. *Nat.*, 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.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2020. oLMpics-On What Language Model Pre-training Captures. *Transactions of the Association for Computational Linguistics*, 8:743–758.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019c. What do you learn from context? probing for sentence structure in contextualized word representations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Terne Sasha Thorn Jakobsen, Maria Barrett, and Anders Søgaard. 2021. Spurious correlations in cross-topic argument mining. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 263–277, Online. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Elena Voita and Ivan Titov. 2020a. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 183–196. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020b. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Andreas Waldis and Iryna Gurevych. 2023. Bridging topic, domain, and language shifts: An evaluation of comprehensive out-of-distribution scenarios. *CoRR*, abs/2309.08316.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. Do NLP models know numbers? probing numeracy in embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5306–5314. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

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.

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.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes BERT. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1046–1061, Dublin, Ireland. Association for Computational Linguistics.

Zining Zhu, Soroosh Shahtalebi, and Frank Rudzicz. 2022. OOD-probe: A neural interpretation of out-of-domain generalization. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*.

A Additional Details of the Experiments

A.1 Probing Tasks

Table 6 shows examples and additional details of the different probing tasks.

A.2 Fold Composition

We rely on a three-folded evaluation for In- and Cross-Topic for a generalized performance measure. These folds cover every instance exactly once in a test split. In addition, we require that In- and Cross-Topic train/dev/test splits have the same number of instances for a fair comparison, as visualized in Figure 5. For Cross-Topic, we make sure that every topic $\{t_1, ..., t_m\}$ is covered precisely once by one of the three test splits $X^{(test)}_{cross}$. To compose $X^{(train)}_{cross}$ and $X^{(dev)}_{cross}$, we randomly distribute the remaining topics for every fold. For In-Topic, we randomly[7] form subsequent test splits $X^{(test)}_{in}$ for every fold from all instances $\{x_1, ..., x_m\}$. $X^{(train)}_{in}$ and $X^{(dev)}_{in}$ are then randomly composed for every fold using the remaining instance set following the dimension of $X^{(train)}_{cross}$ and $X^{(dev)}_{cross}$.

A.3 Training Setup

For all our experiments, we use NVIDIA RTX A6000 GPUs, python (3.8.10), transformers (4.9.12), and PyTorch (1.11.0).

A.4 Probing Hyperparameters

Further, we use for the training of the probes the following fixed hyperparameters: 20 epochs, where we find the best one using dev instances; AdamW (Loshchilov and Hutter, 2019) as optimizer; a batch size of 64; a learning rate of 0.0005; a dropout rate of 0.2; a warmup rate of 10% of the steps; random seeds: $[0, 1, 2]$

In addition, we use the following tags from the huggingface model hub:

- `albert-base-v2`
- `bert-base-uncased`
- `facebook/bart-base`
- `microsoft/deberta-base`
- `roberta-base`

[7]We expect that all folds cover all topics given the small number of topics (8) and the big number of instances.

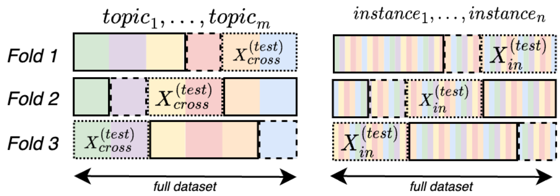

Figure 5: Overview of the In- and Cross-Topic setup using three folds. The colour indicates a topic; solid lines train-, dotted lines dev-, and dashed lines test-splits.

- `google/electra-base-discriminator`
- `gpt2`
- `EleutherAI/pythia-12b`
- `EleutherAI/pythia-12b-deduped`
- `meta-llama/Llama-2-13b-hf`
- `meta-llama/Llama-2-13b-chat-hf`
- `google/t5-xxl-lm-adapt`
- `allenai/tk-instruct-11b-def`

A.5 Fine-Tuning Hyperparameters

To fine-tune on *stance detection*, we use the following setup: 5 epochs, where we find the best one using dev instances; AdamW (Loshchilov and Hutter, 2019) as optimizer; a batch size of 16; a learning rate of 0.00002; a warmup rate of 10% of the steps; random seeds: $[0, 1, 2]$.

A.6 Token-Level Examples for Topic Relevance

In § 5, we use the binned topic-specificity (§ 5) for each token. We show in Table 7 examples for three bins *low*, *medium*, and *high*. The first bin (*low*) is made of tokens, which barely occur in the dataset. The second one (*medium*) consists of tokens which are part of most topics. Finally, the last bin (*high*) includes tokens with a high topic relevance for ones like *Cloning* or *Minimum Wage*.

B Further Results

B.1 Generalization Across Datasets

With Table 8, and Figure 6 we verify the results of § 4, § 5, and § 4 using another *stance detecion*

Task	Example	Label	# Instances	# Labels
`DEP`	I think there is a lot <u>we</u> can <u>learn</u> from Colorado and Washington State.	*nsubj*	40,000	41
`POS`	I think there is a lot <u>we</u> can learn from Colorado and Washington State.	*PRON*	40,000	17
`NER`	I think there is a lot we can learn from Colorado and <u>Washington State</u>.	*PERS*	25,892	17
`Stance`	<u>I think there is a lot we can learn from Colorado and Washington State.</u>	*PRO*	25,492	3

Table 6: Overview and examples of the different probing tasks.

low	***medium***	***high***
fianc, joking, validate, latitude, poignantly, informative ameliorate, bonding, mentors brigade, emancipation, deriving, ignatius, 505, nominations, electorate, SWPS, 731	as, on, take, some, like, how, so, one, these, instead, while, ago where, came, still, many, come, engage, seems	cloning, uniform, wage, marijuana, minimum, gun, cloned, wear, clone, nuclear, energy, penalty, uranium, legalization, cannabis, execution, wast, employment

Table 7: Examples of tokens with a *low*, *medium,* or *high* token relevance following § 4.

	DEP		**POS**		**NER**		**NER**		*Average*		
	In	*Cross*	*In*	*Cross*	*In*	*Cross*	*In*	*Cross*	*In*	*Cross*	Δ
ALBERT	**33.5**	**32.9**	**75.1**	**74.2**	30.9	28.6	**57.3**	32.8	***49.1***	*42.1*	*-7.0*
BART	**32.9**	**33.1**	63.2	62.1	**32.4**	**30.5**	51.9	**47.2**	*45.1*	***43.2***	*-1.9*
BERT	21.6	21.2	54.8	55.9	27.2	27.8	47.4	32.1	*37.8*	*34.2*	*-3.6*
DeBERTa	26.9	27.6	69.6	67.9	29.4	28.5	49.5	35.7	*43.9*	*40.0*	*-3.9*
RoBERTa	20.4	19.9	54.7	53.5	26.1	25.5	37.0	37.8	*35.6*	*34.2*	*-1.4*
ELECTRA	26.6	26.6	69.6	68.6	21.7	24.1	35.1	36.7	*38.2*	*39.0*	***+0.8***
GPT-22	16.9	16.5	42.2	42.2	25.1	24.0	40.8	32.6	*31.2*	*28.8*	*-2.4*
GloVe	12.9	12.2	23.5	22.6	28.1	24.6	45.2	34.2	*27.4*	*23.4*	*-4.0*
Avg. Δ	*-0.3*		*-0.7*		*-0.9*		*-9.5*		-	-	-

Table 8: Results of the four probing tasks using eight LMs in the In- and Cross-Topic setup. We report the mean F_1 (macro averaged) over three random seeds, the average difference between the two evaluation setups per task (last row), and their average per LM (last two columns). Best-performing results within a margin of 1pp are marked for every task and setup.

dataset. Namely, we use the *wtwt* (*will-they-wont-they*) (Conforti et al., 2020) dataset which covers 51.284 tweets annotated either *support*, *refute*, *comment*, or *unrelated* towards five financial topics. The overall performance comparison between In- and Cross-Topic shows the same trend as we already saw in § 4, but on a lower level. We assume this is mainly due to this dataset's more specific domain (twitter) compared to *UKP ArgMin*. Focusing on the influence of topic-specific vocabulary verifies the previously presented results (§ 5) again. LMs pre-trained with purely token-based objectives highly depend on topic-specific vocabulary.

B.2 Comparison of Probing Tasks against Random Initialized LMs

We show in Table 9 and Table 10 the results of running the three linguistic probes on the seven contextualized LMs in their random initialized version. For In- and Cross-Topic, there is a clear performance drop of having random initialized models.

	DEP		**POS**		**NER**	
	Random	Δ	*Random*	Δ	*Random*	Δ
ALBERT	1.4	-42.4	6.8	-41.8	3.4	-76.8
BART	1.4	-35.1	5.0	-43.7	2.7	-72.7
BERT	2.7	-22.7	9.4	-36.0	4.6	-63.9
DeBERTa	7.0	-25.8	16.3	-32.5	16.1	-57.6
RoBERTa	2.2	-22.9	11.0	-37.4	4.7	-59.3
ELECTRA	1.7	-31.9	8.4	-33.1	3.8	-71.5
GPT-2	5.8	-19.4	12.3	-33.2	12.5	-51.0

Table 9: Results of evaluating DEP, POS, and NER using the seven contextual LMs (random initialized) for In-Topic and the difference to their pre-trained counterparts in Table 2.

B.3 The Effect of Removing Random Information

We saw in § 5 that removing topic-specificity has a big impact for some models (like RoBERTa or ELECTRA) but at the same time can even boost the performance of others like BERT. As suggested in Elazar et al. (2021), we apply a sanity check by removing random information from the encodings of LMs. Following the results in Figure 7, removing random information (green bars) performs in between the scenarios with (blue bars) or without (red bars) topic information for cases where we see a clear negative effect when removing topic information. In contrast, removing random information

	DEP		**POS**		**NER**	
	Random	Δ	*Random*	Δ	*Random*	Δ
ALBERT	1.4	-38.1	6.2	-39.6	3.4	-74.6
BART	1.5	-35.4	5.0	-40.3	2.9	-71.2
BERT	2.1	-23.5	9.6	-32.0	4.5	-63.0
DeBERTa	6.8	-23.1	14.0	-28.4	17.2	-57.4
RoBERTa	2.6	-21.0	10.0	-32.1	5.2	-60.3
ELECTRA	3.0	-30.6	9.8	-31.4	4.1	-71.2
GPT-2	5.8	-18.1	13.6	-25.0	11.0	-50.9

Table 10: Results of evaluating DEP, POS, and NER using the seven contextual LMs (random initialized) for Cross-Topic and the difference to their pre-trained counterparts in Table 2.

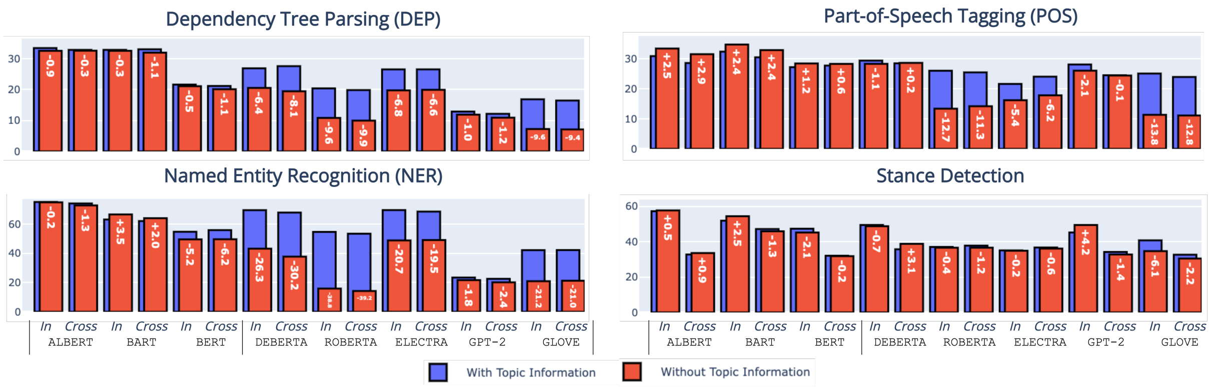

Figure 6: Comparison of the probing results with (blue bars) or without (red bars) topic-specificity for the *will-they-wont-they* dataset (Conforti et al., 2020). The white text indicates the difference between these two scenarios.

can produce a more pronounced effect when we see performance improvements. This observation backs our assumption that removing information can have a regularization effect.

B.4 The Effect of Removing Topic Information on *Seen* and *Unseen* Instances

We show in Figure 8 that a performance drop affects *seen* and *unseen* instances for In- and Cross-Topic equally. Exceptionally, we see *unseen* ones are more affected on POS for DeBERTa and RoBERTa. This result indicates that these LMs fall short of generalizing towards rare vocabularies - like *unseen* instances of POS.

B.5 Analysis of Per-Class Results for NER

When considering the per-class results of NER in Table 11, we see the classes CARDINAL, MONEY, ORG, and PERSON show the biggest differences between In- and Cross-Topic. For ORG and PERSON, we see their topic-specific terms as the main reason for the performance gap. In contrast, we were surprised about the high difference for CARDINAL. We think this is mainly because this class embodies all numbers belonging to no other class. For MONEY, we see its uneven distribution over topics as the main reason for the performance difference - one topic covers more than 50% of the instances. These entities are highly topic-specific from a statistical point of view.

Despite having almost the same performance for In-Topic, BART and DeBERTa tend to outperform ALBERT on classes with more semantic complexities - like GPE, ORG or PERSON. For Cross-Topic, we see ALBERT performing better in classes unevenly distributed instances over topics - like MONEY. Further, it outperforms BART and DeBERTa on less semantical classes (CARDINAL, ORDINAL, PERCENT).

		CARDINAL	DATE	GPE	MONEY	NORP	ORDINAL	ORG	PERCENT	PERSON
In	ALBERT	95.0	95.3	89.4	95.0	91.3	97.8	80.2	99.2	82.7
	BART	94.8	94.6	89.7	95.6	91.6	97.3	81.0	99.4	83.5
	DeBERTa	95.3	95.6	90.0	96.5	91.5	97.4	81.1	99.2	83.7
Cross	ALBERT	91.2	95.0	88.6	55.6	90.8	98.1	78.8	98.9	81.7
	BART	90.1	94.2	88.9	35.0	90.7	97.6	79.1	98.8	81.8
	DeBERTa	88.3	95.3	88.6	0.0	90.5	97.5	79.8	98.6	81.8

Table 11: Per-class results of ALBERT, BART, and DeBERTa on NER for In- and Cross-Topic.

		CARDINAL	DATE	GPE	MONEY	NORP	ORDINAL	ORG	PERCENT	PERSON
In	BART	-0.23	0.04	0.15	0.15	0.02	-0.04	0.08	-0.13	0.20
	BERT	1.65	-0.15	-0.04	28.00	-0.14	-0.58	0.06	0.00	0.22
	DEBERTA	-1.14	-0.13	-1.48	-7.74	-14.40	-0.30	-0.82	-0.12	-0.10
	ROBERTA	-6.00	-3.00	-7.82	-24.09	-90.61	-98.06	-2.66	-0.51	-0.58
Cross	BART	-0.48	0.01	-0.13	2.45	-0.06	-0.52	-0.38	-0.09	-0.03
	BERT	-0.05	-0.05	1.00	0.00	8.95	-0.60	0.29	0.00	0.00
	DEBERTA	-0.07	-0.16	-2.52	0.00	-21.88	-0.35	-0.91	-0.01	0.07
	ROBERTA	-9.04	-2.63	-7.45	0.00	-85.23	-98.07	-2.99	-35.97	-0.46

Table 12: Class-wise effect on the performance when removing topic information of BART, BERT, DeBERTa, and RoBERTa on NER for In- and Cross-Topic.

B.6 Effect of Removing Token-Level Topic Information of Per-Class Results for NER

Similar to the previous analysis, there are apparent effects of removing topic information when considering NER classes separately. Table 12 shows these results for BART, BERT, DeBERTa, and RoBERTa. Like the overall result, BART, DeBERTa, and RoBERTa perform less when removing topic information. Whereby the effect is the most pronounced for RoBERTa with the highest performance drop for In- and Cross-Topic on classes like NORP or ORDINAL. In addition, these results show that the performance gain from removing topic information within BERT happens on MONEY for In-Topic and NORP for Cross-Topic.

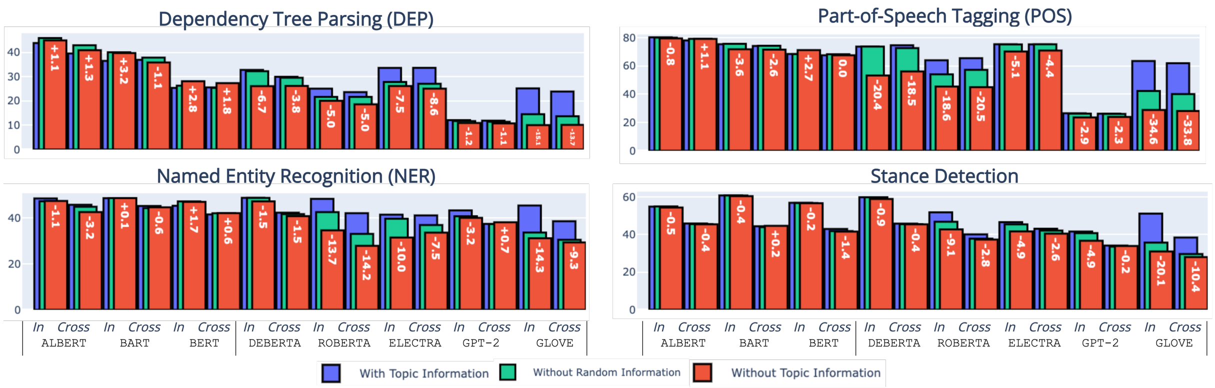

Figure 7: Comparison of the probing results with (blue bars) and without (red bars) topic information, or without random information (green bars). The white text indicates the difference between the blue and red bars.

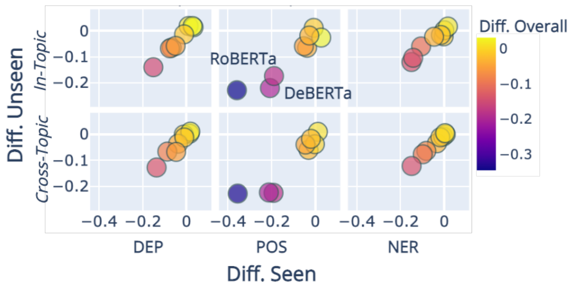

Figure 8: Performance difference for *seen* (x-axis) and *unseen* (y-axis) instances when removing topic information or not. One dot represents one LM.

		CARDINAL	DATE	GPE	MONEY	NORP	ORDINAL	ORG	PERCENT	PERSON
In	ALBERT	-34.2	-25.4	-26.9	-95.0	-51.9	-60.3	-22.4	-99.2	-21.8
	BART	-8.5	-7.2	-7.5	-7.2	-10.4	-36.6	-4.1	-3.8	-2.7
	BERT	-1.9	-2.0	-2.0	34.8	-4.4	-17.9	-0.8	-3.9	-1.1
	DEBERTA	-15.1	-6.8	-8.7	-19.5	-43.7	-60.8	-8.8	-24.8	-8.3
Cross	ALBERT	-21.5	-10.4	-19.1	-55.6	-34.4	-13.1	-10.7	-81.0	-9.2
	BART	-9.2	-7.4	-7.0	-16.3	-11.2	-24.4	-3.9	-4.5	-2.1
	BERT	-2.5	-1.2	-1.2	3.6	-2.2	-9.7	-0.8	-2.6	-0.5
	DEBERTA	-18.2	-6.2	-12.7	0.0	-50.6	-76.0	-11.7	-73.5	-6.8

Table 13: Per-class difference before and after fine-tuning on *stance detection* of ALBERT, BART, BERT, and DeBERTa on NER for In- and Cross-Topic.

B.7 The Effect of Fine-Tuning on NER Classes

Analysing the results (Table B.7) for every NER class gives additional insights into where the fine-tuning had the most significant effect. We generally see the biggest effect on classes with less semantic meaning, like ORDINAL, PERCENT, or MONEY. At the same time, GPE, PERSON, and ORG are less affected as classes with more attached semantics. Regarding the different LMs, ALBERT and DeBERTa show the most performance training, while BERT gains performance for the MONEY class.

	DEP		POS		NER	
	In	*Cross*	*In*	*Cross*	*In*	*Cross*
ALBERT	85.2	83.9	93.8	93.6	86.9	85.0
BART	80.9	81.0	92.6	92.0	87.1	84.5
BERT	76.1	76.1	89.2	88.6	85.2	82.9
DeBERTa	81.2	79.9	92.8	93.1	87.5	84.0
RoBERTa	75.9	75.5	89.6	90.1	86.3	83.2
ELECTRA	81.1	80.7	92.3	92.2	82.8	82.2
GPT-2	69.8	69.1	85.8	85.7	84.6	81.1
GloVe	39.5	38.5	46.6	45.9	78.8	77.2
Average	73.7	73.1	85.3	85.2	84.9	82.5
BERT 80k	80.5	79.1	92.0	91.5	-	-
BERT 160k	84.3	84.2	93.1	92.8	-	-
BERT 320k	86.3	85.6	93.7	93.3	-	-
BERT (Tenney et al., 2019c)	93.0		97.0		96.1	
BERT (Tenney et al., 2019a)	95.2		96.5		96.0	
BERT (Hewitt and Liang, 2019b)	89.0		97.2		-	

Table 14: Accuracy results for In- and Cross-Topic probing results for eight LMs, across three random seeds. Further, we report results of gradually increasing the number of consider instance (BERT 80k, BERT 160k, and BERT 320k), as well as reference performance of previous work (Tenney et al., 2019c,a; Hewitt and Liang, 2019b).

B.8 Annotation Verification

To evaluate probing tasks in the In- and Cross-Topic setup, we rely on data with topic annotations on the instance level - like the *UKP ArgMin* (Stab et al., 2018) or the *wtwt* (Conforti et al., 2020) dataset. Since these datasets do not include linguistic annotations, we make use of spaCy[8] to automatically derive the labels for *dependency tree parsing (DEP)*, *part-of-speech tagging (POS)*, or *named entity recognition (NER)*. We used the `en_core_web_sm` model, which provides reliable labels with a detection performance in terms of accuracy of 97.0 for POS, 90.0-92.0 for DEP, and an F1 score of 85.0 for NER (details available online). Note, this performance referees to iden-

[8]https://spacy.io/

tify valid candidates (like entities for NER) given a piece of text, and assign the corresponding labels, such as person or organization. In contrast, in probing, we consider only the second step: assigning the right label of a valid candidate. Therefore, we can not directly compare recognition and probing performance.

Considering our results (§ 4), we see these derived labels as reliable and well aligned with previous work (Tenney et al., 2019c,a; Hewitt and Liang, 2019b), even though we mainly report F_1 score. One reason for that is the similar performance ranking (DEP $<$ NER $<$ POS) as in previous work, considering F_1 score as well as the accuracy score reported in Table 14. Another reason is the narrowing accuracy performance gap between our experiments and previous work when we gradually increase the number of consider instance from 40k to 80k, 160k, until 320k.