Domain Pre-training Impact on Representations

Cesar Gonzalez-Gutierrez **Ariadna Quattoni**
Universitat Politècnica de Catalunya, Barcelona, Spain
cesar.gonzalez.gutierrez@upc.edu,
aquattoni@cs.upc.edu

Abstract

This empirical study analyzes how the choice of pre-training corpus affects the quality of learned transformer representations. We focus specifically on the representation quality achieved through pre-training alone. Our experiments demonstrate that pre-training on a small, specialized corpus can produce effective representations, and that the effectiveness of combining a generic and a specialized corpora depends on the distributional similarity between the target task and the specialized corpus.

1 Introduction

Since 2018, pre-training (PT) has become a standard step in model development, demonstrating effective transfer learning for diverse natural language understanding tasks (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2018). This approach leverages textual corpora to induce useful representations by minimizing a self-supervised language modeling loss function. These representations are typically leveraged via transfer learning for downstream tasks using supervised data.

When selecting a PT corpus, we generally have three options: 1) use a large, generic corpus $\boldsymbol{G}$, similar to training a foundational model; 2) use a smaller, *specialized* corpus $\boldsymbol{S}$, which is expected to be more relevant to the target task; 3) combine both $\boldsymbol{G}$+$\boldsymbol{S}$, as in domain-adaptive PT (Gururangan et al., 2020), where a model initially pre-trained on $\boldsymbol{G}$ is further refined using $\boldsymbol{S}$.

For example, when developing a toxicity filter for a new forum with only a few labeled comments, one option is to use a pre-trained representation computed on a generic corpus $\boldsymbol{G}$ (such as BERT). Alternatively, we could construct a specialized corpus $\boldsymbol{S}$ by collecting posts from the forum itself or from other related sources. We can then pre-train using either $\boldsymbol{G}$ alone or a combination of both $\boldsymbol{G}$+$\boldsymbol{S}$. We refer to the textual distribution from which a PT corpus is sampled as its *domain*.

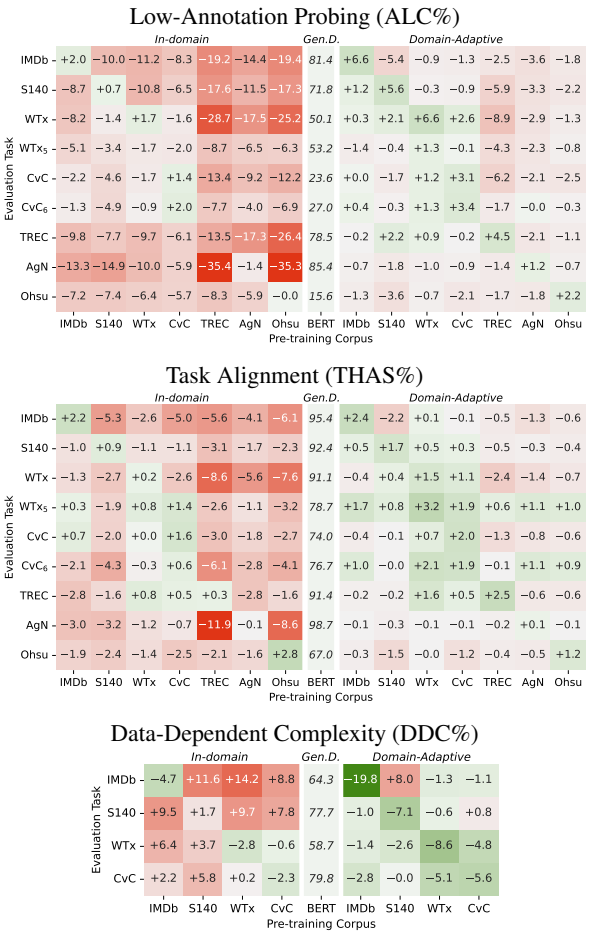


Figure 1: Cross-domain pre-trained representation performance. We report absolute performance for GD embeddings (center column), and relative improvement for ID (left matrix) and DA (right matrix).

The goal of this paper is to understand how pre-training corpus choice affects the quality of the learned representation. We focus on measuring representation quality after self-supervised pre-training, without subsequent supervised training (as in fine-tuning). Following prior work, we use three representation quality metrics: a standard

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 22033–22049
November 4-9, 2025

probing technique (Ettinger et al., 2016; Adi et al., 2017), and two label-representation alignment metrics, one based on hierarchical clustering structures (Gonzalez-Gutierrez et al., 2023) and the other based on data-dependent complexity (Yauney and Mimno, 2021).

We conduct an empirical analysis using a wide range of data sources and tasks, evaluating representation quality under different PT scenarios. We derive the following conclusions: 1) If the specialized corpus $\boldsymbol{S}$ is close to the target task distribution, and it is not too small, pre-training on $\boldsymbol{S}$ can be as effective as pre-training on $\boldsymbol{G}$. 2) Pre-training on both $\boldsymbol{G}$ and $\boldsymbol{S}$ (i.e. domain-adaptive pre-training) can sometimes improve performance compared to using $\boldsymbol{G}$ alone, but it can also lead to decreased performance. 3) The success of the domain-adaptive strategy depends on the similarity between the target task distribution and that of $\boldsymbol{S}$. We validate this claim by showing a consistent correlation between distributional similarity and domain-adaptive performance improvement, using two different domain similarity measures.

2 Pre-training Effect on Transformer Embeddings

We begin this section with a description of the experimental setting followed by a discussion of the main results.

2.1 Experimental Setting

Pre-training scenarios:

- In-domain (**ID**): Pre-training a model from scratch on a domain-specific corpus $\boldsymbol{S}$.
- General domain (**GD**): The common PLM training approach of using a large generic corpus $\boldsymbol{G}$.
- Domain-adaptive (**DA**): Starting with a PLM trained on $\boldsymbol{G}$, continuing pre-training on $\boldsymbol{S}$.

Corpora. To simulate domain-specific corpora ($\boldsymbol{S}_{\text{domain}}$) for pre-training, we used the unlabeled text in the following classification datasets: IMDb (Maas et al., 2011), Sentiment 140 (Go et al., 2009), Wiki Toxic (Wulczyn et al., 2017), Civil Comments (Borkan et al., 2019), TREC (Li and Roth, 2002; Hovy et al., 2001), AG-News (Zhang et al., 2015), and Ohsumed (Hersh et al., 1994). Test partitions are excluded as held-out sets and are only used to compute the representation's quality metrics.

As general corpora $\boldsymbol{G}$, we employ the models pre-trained on their generic corpora: BookCorpus plus English Wikipedia for BERT, and WebText for GPT-2. Table 1 shows a summary of PT corpora with their text type and size.

Corpus	Text	#words
IMDb	movie reviews	18M
Sentiment 140	tweets	21M
Wiki Toxic	user discussions	11M
Civil Comments	user comments	100M
TREC	short questions	56k
AG-News	news	4.5M
Ohsumed	medical abstracts	1.8M
BookCorpus	books	800M
Wikipedia (EN)	encyclopedia	2.5B
WebText	miscellaneous	9.1B

Table 1: Pre-training corpora.

Dataset	Task	$\lvert\mathcal{Y}\rvert$	prior	len.	# train / test
IMDb	sentiment	2	0.5	233	25k / 25k
Sentiment 140	sentiment	2	0.5	14	1.6M / 498
Wiki Toxic	toxicity	2	0.096	68	160k / 64k
Wiki Toxic 5	toxicity	5	imb.	52	9.8k / 3.4k
Civ. Com.	toxicity	2	0.08	53	1.9M / 97k
Civ. Com. 6	toxicity	6	imb.	49	145k / 7.4k
TREC	topic	6	imb.	10	5.5k / 500
AG-News	topic	4	$1/\lvert\mathcal{Y}\rvert$	38	120k / 7.6k
Ohsumed	topic	23	imb.	175	10k / 13k

Table 2: Evaluation dataset summary with number of classes, label prior, average length and partition size.

Models. For pre-training, our main experiments use the BERT masked language model (MLM) architecture (Devlin et al., 2019). In A.3, we present a parallel study using GPT-2 autoregressive models. These are two well-studied transformer architectures (Ethayarajh, 2019; Rogers et al., 2020) and have a size suitable for running multiple PT experiments. Following Liu et al. (2019b), we omit the next sentence prediction (NSP) objective in BERT and focus solely on MLM.

Embeddings. To obtain sentence embeddings, we extract mean token embeddings from the last layer. A.6 presents a study using alternative representation functions, showing that different layer and token selection strategies yield similar trends, albeit with varying absolute performance scores. We adopt this commonly used method (Reimers and Gurevych, 2019) as our representative extraction strategy, though similar conclusions can be drawn using other representation functions.

Evaluation Metrics. To quantify the changes underwent by representations during PT, we employ three evaluation metrics targeted at the model's em-

bedding space quality w.r.t. a task: probing and two label-representation alignment scores. In A.4, we explore an intrinsic clustering quality metric that does not depend on task labels.

Probing (Ettinger et al., 2016; Adi et al., 2017) uses weak classifiers to evaluate the task performance attributable to the representation. We use low-annotation probes to test the representation's ability to uncover structures that enable learning from few samples. In particular, the probes are MaxEnt classifiers on top of the embeddings trained with sample sizes ranging from 100 to 1000, increasing in steps of 100. We report the area under this learning curve (ALC), using accuracy for the binary balanced and multi-class datasets, and F1 of the target class for binary imbalanced datasets (Civil Comments and Wiki Toxic).

Task Hierarchical Alignment Score (THAS; Gonzalez-Gutierrez et al., 2023) quantifies the alignment between hierarchical clustering structures in the representation space and task labels. This metric measures the degree of cluster *purity* at different hierarchical levels. A representation capable of perfectly separating n classes into n pure clusters will obtain the maximum score. More precisely, we used agglomerative clustering on the embeddings and, for each partition, measured the area under the precision-recall curve using in-cluster class prevalence as label predictions for each data point. Fig. 9 (A.5) shows the curves from which ALC and THAS aggregate metrics where computed.

Data-Dependent Complexity (DDC; Yauney and Mimno, 2021) quantifies the compatibility between a representation and a binary classification task. It captures patterns through the eigen decomposition of a kernel matrix that measures sample similarity in the representation space. Label alignment is evaluated based on the extent to which the label vectors can be reconstructed from their projections onto the top eigenvectors of the kernel matrix. Following Yauney and Mimno (2021), DDC score is computed as the ratio of the real annotation's DDC to the average DDC over random annotations.

Evaluation Tasks. The three metrics described above are computed on the test partitions of the evaluation tasks listed in Table 2. These evaluation tasks correspond to the original annotations of the text corpora used to pre-train our models (Table 1). In addition, we constructed two new multi-class tasks from the two toxicity benchmarks, which involve predicting the specific subtype of toxicity in each comment: Wiki Toxic_5 and Civil Comments_6. Further details on the construction of these datasets are provided in A.1.

We repeat each experiment five times with different random seeds and report the average performance. Additional experimental details are provided in A.1.

2.2 Main Results

We study how the choice of pre-training corpus affects representation quality. More specifically, our goal is to understand the necessary conditions for cross-domain generalization. To this end, we first pre-train models on each of the seven corpora $\boldsymbol{S}_{\text{domain}}$ described in Table 1. This is done under two settings: in-domain (training solely on $\boldsymbol{S}_{\text{domain}}$), and domain-adaptive (training with both $\boldsymbol{G}$+$\boldsymbol{S}_{\text{domain}}$). We then evaluate each resulting representation across all nine target tasks using three representation quality metrics: low-annotation probing, task alignment, and DDC. Additionally, we compute the representation quality for the general-domain representation (BERT) as a baseline.

For each metric, we compute a column of absolute scores for the GD baseline, shown in the center of Fig. 1. We then generate two relative improvement matrices, M_{ID} and M_{DA}, corresponding to the in-domain and domain-adaptive settings, respectively (left and right matrices in Fig. 1). Each matrix entry $M(i, j)$ represents the performance difference between the evaluated representation and the GD baseline. For example, $M_{\text{DA}}(\text{WTx}, \text{CvC})$ denotes the difference in performance evaluated on the WTx task when using domain-adaptive pre-training on the $\boldsymbol{S}_{\text{CvC}}$ corpus, compared to using the GD representation. Higher scores indicate improvement for low-annotation probing and task alignment, while lower values indicate DDC improvement (lower DDC scores are better).

Focusing on the **domain-relevant scenario** (i.e., matrix entries where the pre-training corpus matches the evaluation task domain), and particularly within the in-domain setting (left matrices), we find that, in most cases, pre-training on a smaller, specialized dataset yields representations comparable in quality to those produced by a GD model trained on a much larger corpus. However, the size of the corpus $\boldsymbol{S}_{\text{domain}}$ can be a limiting factor. Representations learned from the smallest corpora ($\boldsymbol{S}_{\text{TREC}}$, $\boldsymbol{S}_{\text{Ohsu}}$) fail to match or improve upon

the performance of the GD representation. Similarly, the limited size of the Wiki Toxic$_5$ dataset makes it a challenging task for probing.

Still focusing on the domain-matched cases, we observe that the domain-adaptive strategy (right matrices) yields the most substantial improvements in representation quality (Gururangan et al., 2020).

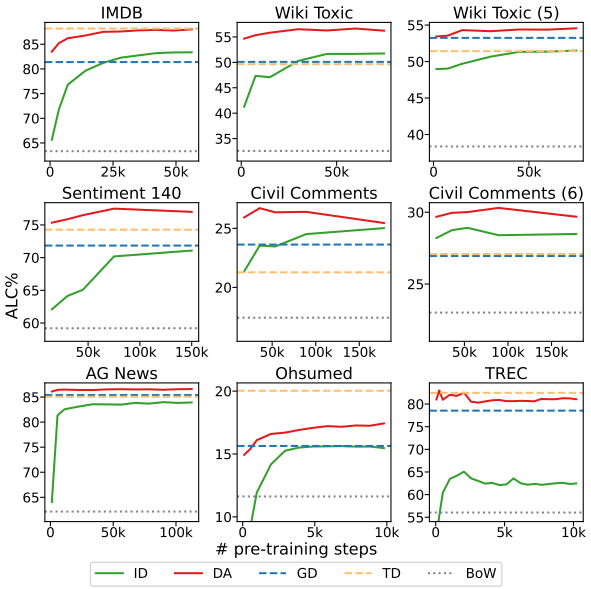


Figure 2: Probe performance (ALC%) as a function of PT steps. A text domain (TD) and sparse bag-of-words (BoW) baselines are shown for comparison.

Fig. 2 complements the domain-matched results showing low-annotation probing performance as a function of the number of PT steps. Analogous graphs for task alignment are shown in Fig. 5 (see A.2.1). We observe that ID pre-training can achieve performance comparable to GD, given a sufficient number of PT iterations.

We additionally included text domain (TD) baselines, i.e., models pre-trained on domains closely related to the text type of each task: reviews for IMDb, banned community comments for the toxicity datasets, tweets for Sentiment140, news articles for AG-News, medical texts for Ohsumed, and question-answer pairs for TREC. These models are described in more detail in A.1.

For the Ohsumed dataset, the TD baseline indicates that $\boldsymbol{G}$ is not well suited to the medical domain, where a text-specific model yields significantly better representations. At the same time, it also shows that $\boldsymbol{S}_{\text{Ohsu}}$ is not large enough to produce high-quality representations on its own, suggesting that corpus size is a limiting constraint.

Turning to **cross-domain performance**, where the pre-training corpus domain $\boldsymbol{S}_{\text{domain}}$ differs from

Metric	Binary Tasks ncvg	Binary Tasks E[acc$_{\text{L1}}$]	All Tasks ncvg	All Tasks E[acc$_{\text{L1}}$]
ΔProbe	75.85	80.30	73.91	57.18
ΔTHAS	66.67	64.89	55.39	56.26
ΔDDC	71.71	96.00		

Table 3: Spearman correlation (%) between similarity of pre-training and target task distributions and representation quality improvement gains.

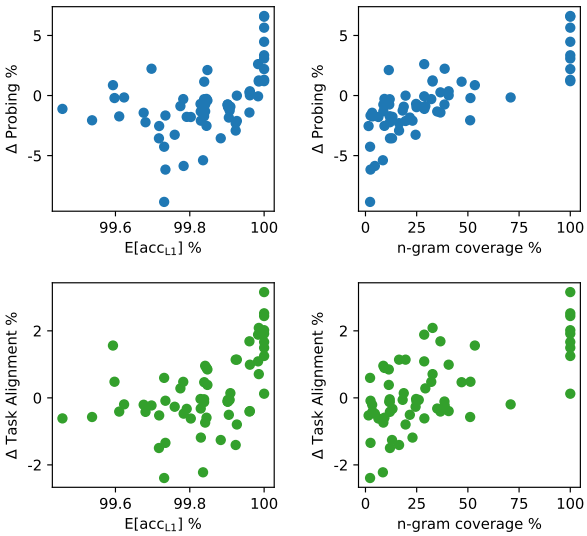


Figure 3: Correlation between similarity of pre-training corpus and target task distributions and representation quality differences using all datasets.

the task domain $\boldsymbol{S}_{\text{task}}$, the outcomes are mixed. In some cases, domain-specific pre-training improves representation quality; in others, it degrades performance.

Notably, when the two domains are highly related (such as $\boldsymbol{S}_{\text{WTx}}$ and $\boldsymbol{S}_{\text{CvC}}$), the resulting representations generally support better performance on relevant probes (Wiki Toxic, Wiki Toxic$_5$, Civil Comments, and Civil Comments$_6$). In contrast, domains like $\boldsymbol{S}_{\text{Ohsu}}$ and $\boldsymbol{S}_{\text{AgN}}$, which are unrelated to most tasks, perform poorly in cross-domain probes. Similarly, the smallest corpus from a dissimilar domain ($\boldsymbol{S}_{\text{TREC}}$), yields the weakest representations.

A natural explanation for these observations is that the effectiveness of domain-specific pre-training depends on the distributional similarity between the pre-training domain and the target task domain (Grangier and Iter, 2022).

To test this hypothesis, we computed two distributional similarity metrics. The first, n-gram coverage (ncvg), measures the proportion of n-grams observed in the target task distribution that also

appear in the pre-training distribution:

$$\text{ncvg} = \frac{|N_t \cap N_c|}{|N_t|},$$

where N_t is the set of unique n-grams (up to size n) observed in the target distribution $\boldsymbol{S}_t$, and N_c is the analogous set in the pre-training corpus $\boldsymbol{S}_c$.

A classical metric for measuring distribution distance is the Kullback and Leibler's (1951) divergence between the n-gram distributions P_t and P_c:

$$KL = \sum_{n \in N_t} P_t(n) \log \frac{P_t(n)}{P_c(n)}$$

However, this metric is problematic in our case due to sparsity: for many n-grams, $P_c(n)$ is zero. To address this, we propose a more robust metric, suitable for sparse distributions. This metric estimates the expected L1-accuracy when using $P_c(n)$ as a proxy for $P_t(n)$:

$$E[\text{acc}_{\text{L1}}] = 1 - \sum_{n \in N_t} P_t(n) \cdot |P_t(n) - P_c(n)|$$

Notice that both metrics are asymmetric.

To validate our hypothesis, we computed the correlation between representation quality improvements under the domain-adaptive scenario and these two similarity metrics. As shown in Table 3, there is a strong correlation between distributional similarity and gains in representation quality. Fig. 3 shows the scatter plots of these metrics computed across all evaluation tasks. Additional correlation graphs for the binary tasks are provided in A.2.2.

3 Related Work

In theoretical studies, Ge et al. (2024); Deng et al. (2024) attributed the advantage of PT to the induction of useful representations and to learning complexity reduction. Tripuraneni et al. (2020) showed that shared representations enable transfer learning, improving generalization across tasks, even when annotation coverage is sparse (Du et al., 2021). Domain adaptation depends on the diversity of those tasks and the relative sample sizes (Grangier and Iter, 2022). In this work, we empirically investigate the adaptability of transformer representations and their ability to generalize across diverse tasks in a controlled experimental setting.

Domain-specific data has been used to train different PLMs from scratch (Beltagy et al., 2019; Lee et al., 2019, *inter alia*) or via domain adaptation (Gururangan et al., 2020; Han and Eisenstein, 2019). Aharoni and Goldberg (2020) studied the implicit notion of domain in PLMs, Krishna et al. (2023) studied task-domain PT, and Chronopoulou et al. (2022) leveraged the cross-domain overlap using adapters. In contrast, we investigate the representation changes that enable domain adaptation in different PT scenarios.

Representation properties have been studied using probing tasks (Ettinger et al., 2016; Adi et al., 2017; Conneau et al., 2018; Hewitt and Manning, 2019) or analyzing their relation to annotations (Gonzalez-Gutierrez et al., 2023; Yauney and Mimno, 2021; Zhou and Srikumar, 2021). Representation learning dynamics has been explored across various syntactic (Chiang et al., 2020; Saphra and Lopez, 2019), semantic (Templeton et al., 2024; Liu et al., 2021, 2019a), or multilingual model capabilities (Wang et al., 2024; Blevins et al., 2022). The role of representations in generalization has been studied for linguistic phenomena (Choshen et al., 2022; Warstadt et al., 2020) or factual knowledge (Zhang et al., 2021). These works have not explored dynamics in varying PT scenarios, with a focus on cross-domain generalization.

4 Conclusion

This paper compared pre-trained representations obtained under different pre-training scenarios. We used three representation quality metrics to evaluate how effectively transformer-based representations, learned from diverse pre-training corpora, can be leveraged for a target task. We draw two main conclusions:

1. The success of cross-domain representation transfer can be predicted by the degree of similarity between the n-gram distributions of the pre-training and target domains. While this conclusion may seem intuitive, to the best of our knowledge, this is the first empirical study to provide a thorough analysis of cross-domain adaptation of transformer representations to substantiate it.

2. Pre-trained representations learned from relatively small, domain-specific corpora can be highly competitive. This suggests that the relevance of the PT data may be more important than its size. High-quality models can thus be developed using only domain-specific data, without requiring extensive GD corpora and with a fraction of the computational resources.

Limitations

This empirical study explored the properties of transformer-based representations using MLM (BERT) and autoregressive (GPT-2) pre-trained models. While our focus is on any transformer-based representation, we have not compared our results with other transformer architectures. We believe the findings are representative, but a broader experimental setup would allow for more robust conclusions.

Although current state-of-the-art LMs present interesting properties that deserve our attention, the methods presented in this work do not scale. In other words, pre-training very large LMs is not feasible with medium-size computational resources.

Fine-tuning is a widely used approach for adapting pre-trained representations for downstream tasks. However, this work focuses solely on the changes induced by pre-training, without any supervised learning. Our aim is to understand how self-supervised pre-training alone shapes representations to support cross-domain transferability. The impact of fine-tuning on representation quality is an important research direction in its own right, which we leave for future work.

We adopted an empirical framework that allows us to assess representation performance not only through probing tasks, but also by measuring properties of the representation space in relation to the target task (such as task alignment) to better account for structural changes in representations that may explain performance differences due to pre-training. While a more diverse set of tasks (as in Conneau et al. 2018 and similar works) could provide a more comprehensive evaluation of the capabilities of representations, appropriate parallel metrics that capture changes at the representation level are currently lacking. As a result, our experimental design is limited to classification tasks. Expanding to a broader task set could could support a more general and robust evaluation, but at the cost of reduced interpretability.

In particular, the DDC metric we use to evaluate embedding performance is restricted to binary classification tasks. Although this metric is consistent with the other representation quality measures employed, this constraint may limit the generalizability of our findings.

Potential Risks

We do not foresee any potential societal risks derived from the use of the methods presented in this work.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme under grant agreement No 853459. The authors gratefully acknowledge the computer resources at ARTEMISA, funded by the European Union ERDF and Comunitat Valenciana as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV). This research is supported by a recognition 2021SGR-Cat (01266 LQMC) from AGAUR (Generalitat de Catalunya).

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *Preprint*, arXiv:1608.04207.

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Jason Ansel, Edward Yang, Horace He, Natalia Gimelshein, Animesh Jain, Michael Voznesensky, Bin Bao, Peter Bell, David Berard, Evgeni Burovski, Geeta Chauhan, Anjali Chourdia, Will Constable, Alban Desmaison, Zachary DeVito, Elias Ellison, Will Feng, Jiong Gong, Michael Gschwind, Brian Hirsh, Sherlock Huang, Kshiteej Kalambarkar, Laurent Kirsch, Michael Lazos, Mario Lezcano, Yanbo Liang, Jason Liang, Yinghai Lu, C. K. Luk, Bert Maher, Yunjie Pan, Christian Puhrsch, Matthias Reso, Mark Saroufim, Marcos Yukio Siraichi, Helen Suk, Shunting Zhang, Michael Suo, Phil Tillet, Xu Zhao, Eikan Wang, Keren Zhou, Richard Zou, Xiaodong Wang, Ajit Mathews, William Wen, Gregory Chanan, Peng Wu, and Soumith Chintala. 2024. Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2*, ASPLOS '24, page 929–947, New York, NY, USA. Association for Computing Machinery.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono- and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

Cheng-Han Chiang, Sung-Feng Huang, and Hung-yi Lee. 2020. Pretrained language model embryology: The birth of ALBERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6813–6828, Online. Association for Computational Linguistics.

Leshem Choshen, Guy Hacohen, Daphna Weinshall, and Omri Abend. 2022. The grammar-learning trajectories of neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8281–8297, Dublin, Ireland. Association for Computational Linguistics.

Alexandra Chronopoulou, Matthew Peters, and Jesse Dodge. 2022. Efficient hierarchical domain adaptation for pretrained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1336–1351, Seattle, United States. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. 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.

Yuyang Deng, Junyuan Hong, Jiayu Zhou, and Mehrdad Mahdavi. 2024. On the Generalization Ability of Unsupervised Pretraining. *Proceedings of machine learning research*, 238:4519–4527.

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.

Simon Shaolei Du, Wei Hu, Sham M. Kakade, Jason D. Lee, and Qi Lei. 2021. Few-shot learning via learning the representation, provably. In *International Conference on Learning Representations*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 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)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. 2024. On the provable advantage of unsupervised pretraining. In *The Twelfth International Conference on Learning Representations*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision. *CS224N project report, Stanford*, 1(12):6.

Cesar Gonzalez-Gutierrez, Audi Primadhanty, Francesco Cazzaro, and Ariadna Quattoni. 2023. Analyzing text representations by measuring task alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 70–81, Toronto, Canada. Association for Computational Linguistics.

David Grangier and Dan Iter. 2022. The trade-offs of domain adaptation for neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. 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 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature*, 585(7825):357–362.

William Hersh, Chris Buckley, T. J. Leone, and David Hickam. 1994. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR '94*, pages 192–201, London. Springer London.

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, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. 2001. Toward semantics-based answer pinpointing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Kundan Krishna, Saurabh Garg, Jeffrey Bigham, and Zachary Lipton. 2023. Downstream datasets make surprisingly good pretraining corpora. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12207–12222, Toronto, Canada. Association for Computational Linguistics.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Xin Li and Dan Roth. 2002. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Zeyu Liu, Yizhong Wang, Jungo Kasai, Hannaneh Hajishirzi, and Noah A. Smith. 2021. Probing across time: What does RoBERTa know and when? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 820–842, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

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.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Naomi Saphra and Adam Lopez. 2019. Understanding learning dynamics of language models with SVCCA. 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 3257–3267, Minneapolis, Minnesota. Association for Computational Linguistics.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Nilesh Tripuraneni, Michael Jordan, and Chi Jin. 2020. On the theory of transfer learning: The importance of task diversity. *Advances in neural information processing systems*, 33:7852–7862.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Di Wu, Wasi Uddin Ahmad, and Kai-Wei Chang. 2024. Pre-trained language models for keyphrase generation: A thorough empirical study. *Preprint*, arXiv:2212.10233.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

Gregory Yauney and David Mimno. 2021. Comparing text representations: A theory-driven approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5527–5539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 649–657, Cambridge, MA, USA. MIT Press.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Yichu Zhou and Vivek Srikumar. 2021. DirectProbe: Studying representations without classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5070–5083, Online. Association for Computational Linguistics.

A Appendix

A.1 Experimental Details

Text Corpora Domain-specific corpora were constructed concatenating the text fields within a dataset, if they contained multiple. We used all the partitions present in the dataset except the test set. This includes training, development (if available), and unsupervised (IMDb only) partitions.

Toxicity datasets (Wiki Toxic and Civil Comments) were pre-processed to remove any markup and non-alphanumeric characters except relevant punctuation.

Multi-class Dataset Variants As part of our evaluation tasks defined in §2, we constructed two new multi-label datasets: Wiki Toxic_5 and Civil Comments_6. These datasets are available on the HuggingFace Hub:

- Wiki Toxic_5: cglez/wiki_toxic_multi
- Civil Comments_6: cglez/civil_comments_multi

These datasets are based on Wiki Toxic and Civil Comments toxicity datasets respectively, which contain a general toxicity annotation with additional subtype annotations: insult, threat, obscene, attack on identity groups, or sexually explicit. We excluded the severe toxicity label for its rarity and not being well distinguished with other classes. The subscript in the dataset denotes the total number of label classes.

To construct these datasets, we first found those samples that contain a single subtype annotation, i.e. without overlapping with other subtypes. Such unambiguous toxicity subtype samples are assigned their corresponding subtype label. Samples flagged with the general toxicity label, but that were not annotated with any specific subtype, constitute another class called 'other'. Table 4 contains a class statistics summary.

Dataset	Label	prior%	#train	#test
Wiki Toxic 5	identity attack	2.1	193	81
	insult	16.2	1530	603
	obscene	24.0	2233	931
	threat	1.5	146	55
	other	56.2	5707	1710
Civ. Com. 6	identity attack	6.4	9298	455
	insult	70.4	101787	5234
	obscene	2.6	3788	208
	sexual explicit	1.9	2720	140
	threat	2.5	3676	184
	other	16.2	23459	1175

Table 4: Multi-label toxicity dataset statistics.

According to these statistics, although a subset of texts are common to the original datasets, these new tasks differ substantially from the originals in terms of number of classes, prior distribution and difficulty.

Pre-training Table 5 presents the parameters used in model pre-training.

Parameter	Value
architecture	$\text{BERT}_{\text{BASE}}$
hidden size	768
max. tokens	512
vocabulary size	30,522
activation	gelu
dropout	0.1
batch size GPU	32
grad. accumulation	3
optimizer	AdamW
learning rate	5e-5
weight decay	linear
precision	fp16
architecture	GPT-2
hidden size	768
max. tokens	1024
vocabulary size	50,257
activation	gelu_new
dropout	0.1
batch size GPU	8
grad. accumulation	12
optimizer	AdamW
learning rate	5e-5
weight decay	linear
precision	fp16

Table 5: Model pre-training parameters.

We pre-trained models for different total number of updates depending on the size of the dataset (see Fig. 4). Devlin et al. (2019) pre-trained BERT for 1M update steps, approximately 40 epochs over the 3.3B word corpus. In comparison, the training length and computation resources needed for our models is orders of magnitude smaller.

Probes MaxEnt probes were implemented using Scikit-learn toolbox (Pedregosa et al., 2011) and NumPy (Harris et al., 2020).

Models Table 6 summarizes the models pre-trained for our experimental study. These models, along with their intermediate pre-training weights, are available for download and use through the HuggingFace Hub platform. More details on how to use these models and their variants can be found in their respective model cards on HuggingFace.

Our experiments employed the BERT implementation from the HuggingFace Transformers library (Wolf et al., 2020) with PyTorch (Ansel et al., 2024)

Model	PT	Corpus	Model Card
$BERT_{BASE}$	ID	AG-News	cglez/bert-ag_news-uncased
		Civil Comments	cglez/bert-civil_comments-uncased
		IMDb	cglez/bert-imdb-uncased
		Ohsumed	cglez/bert-ohsumed-uncased
		Sentiment 140	cglez/bert-s140-uncased
		TREC	cglez/bert-trec-uncased
		Wiki Toxic	cglez/bert-wiki_toxic-uncased
	DA	AG-News	cglez/bert-dapt-ag_news
		Civil Comments	cglez/bert-dapt-civil_comments-uncased
		IMDb	cglez/bert-dapt-imdb-uncased
		Ohsumed	cglez/bert-dapt-ohsumed-uncased
		Sentiment 140	cglez/bert-dapt-s140-uncased
		TREC	cglez/bert-dapt-trec-uncased
		Wiki Toxic	cglez/bert-dapt-wiki_toxic-uncased
GPT-2	ID	AG-News	cglez/gpt2-ag_news
		IMDb	cglez/gpt2-imdb
		Ohsumed	cglez/gpt2-ohsumed
		TREC	cglez/gpt2-trec
		Wiki Toxic	cglez/gpt2-wiki_toxic
	DA	AG-News	cglez/gpt2-dapt-ag_news
		IMDb	cglez/gpt2-dapt-imdb
		Ohsumed	cglez/gpt2-dapt-ohsumed
		TREC	cglez/gpt2-dapt-trec
		Wiki Toxic	cglez/gpt2-dapt-wiki_toxic

Table 6: Pre-trained models generated in this empirical study, made available in the HuggingFace Hub. More details on how to use these models are provided in the corresponding model cards.

backend. The models were trained using a single NVIDIA Tesla Volta V100 32GiB PCIe GPU. Similarly, GPT-2 experiments used the HuggingFace implementation and were computed using a single NVIDIA A100 40GiB PCIe GPU.

Text Domain Models The models used as specialized domain (TD) baselines in Fig. 2 are the following:

- BERT Review (Xu et al., 2019): Domain-adapted BERT to e-commerce reviews for sentiment analysis and option evaluation. Used as TD for *IMDb* dataset.
  activebus/BERT_Review
- BERTweet (Nguyen et al., 2020): A pre-trained BERT for English Tweets. This specialized model is used as baseline for *Sentiment 140*.
  vinai/bertweet-base
- HateBERT (Caselli et al., 2021): A domain-adapted BERT for abusive language detection in English using Reddit comments from banned communities. TD model for *Wiki Toxic*, *Wiki Toxic*$_5$, *Civil Comments*, and *Civil Comments*$_6$.
  GroNLP/hateBERT
- Sentence-BERT (Reimers and Gurevych, 2019): Different models pre-trained for semantic similarity, using various datasets containing 215M question-answer pairs. This model is based on MPNet (with similar size to BERT) and used as TD for *TREC*.
  sentence-transformers/multi-qa-mpnet-base-dot-v1
- NewsBERT (Wu et al., 2024): A BERT-base pre-trained on RealNews corpus, used as TD for *AG-News* dataset.
  uclanlp/newsbert
- BiomedBERT (Gu et al., 2021): a pre-trained BERT using abstracts and full articles from the PubMed medicine library. This model is used as TD for *Ohsumed* dataset.
  microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

A.2 Supplementary Results

A.2.1 Task Alignment Dynamics

Complementing the learning dynamics study in §2.2, Fig. 4 presents additional results showing task alignment as a function of pre-training steps, which were omitted from the main text.

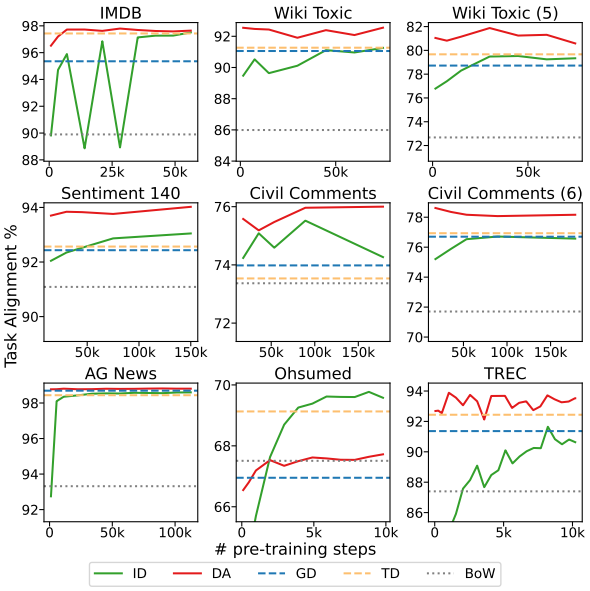


Figure 4: Task alignment (THAS%) performance as a function of PT updates in the three PT scenarios. Text domain (TD) and BoW baselines are shown for comparison.

A.2.2 Performance Differences Correlation

To complement the correlation study presented in §2.2, this section presents the scatter plot generated between the distribution distance measurements and representation quality improvements. Fig. 5 presents the correlation plot for the binary tasks, therefore it includes the DDC metric that is only defined for this type of tasks. As in the previous analysis, this comparison only considers the representations obtained in the domain-adaptive setting.

A.3 GPT-2 Domain Representations

In this section, we present a parallel study to that described in §2 using the GPT-2 model architecture for our pre-training experiments. As before, we pre-train GPT-2 models in three pre-training (PT) scenarios: general domain (GD), in-domain (ID), and domain-adaptive (DA). For the GD scenario, we use the standard GPT-2 model pre-trained on its original WebText corpus. For the domain-specific scenarios, we use the following text corpora: $\boldsymbol{S}_{\text{IMDb}}$, $\boldsymbol{S}_{\text{WTx}}$, $\boldsymbol{S}_{\text{TREC}}$, $\boldsymbol{S}_{\text{AgN}}$, and $\boldsymbol{S}_{\text{Ohsu}}$ (see §2.1 and Table 1 for further details).

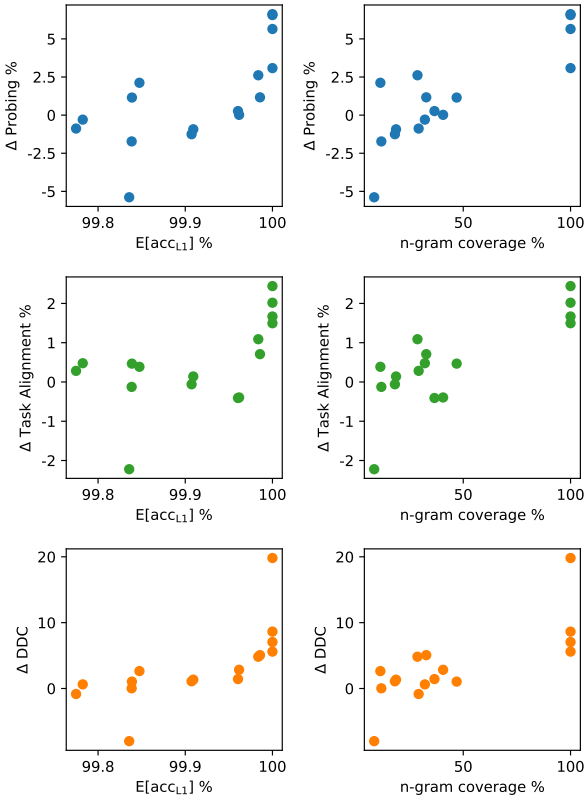


Figure 5: Correlation between distribution similarity and representation quality improvements for binary datasets and BERT-based domain-adapted representations.

Pre-training is performed using the causal language modeling objective on each corpus. Domain-specific models are pre-trained for an increasing number of epochs, with intermediate checkpoints saved to analyze learning dynamics.

To evaluate the learned representations, we use both low-annotation probing and task alignment, applying them to the following tasks: IMDb, Wiki Toxic, Wiki Toxic$_5$, TREC, AG-News, and Ohsumed.

Fig. 6 shows the performance of GPT-2 representations in the cross-domain setting. The results are qualitatively similar to those obtained with BERT representations (see Fig. 1). The best performance for each task is typically achieved when the representations are learned from the relevant domain corpus, while more dissimilar corpora tend to yield lower performance. Looking at the representations of the smallest corpora, corpus length also appears to influence the effectiveness of the learned GPT-2 representations. Once again, the domain-adaptive strategy produces the most effective representations.

Fig. 7 illustrates how representation performance

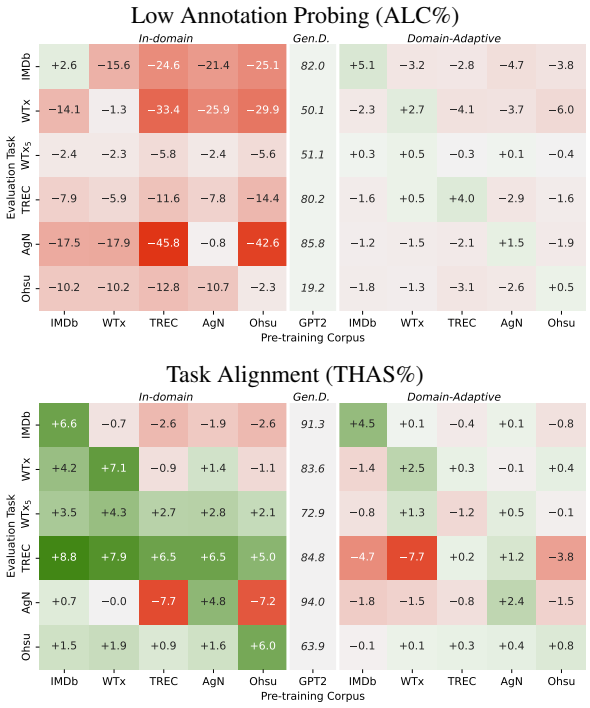


Figure 6: Performance of cross-domain GPT-2 pre-trained representations.

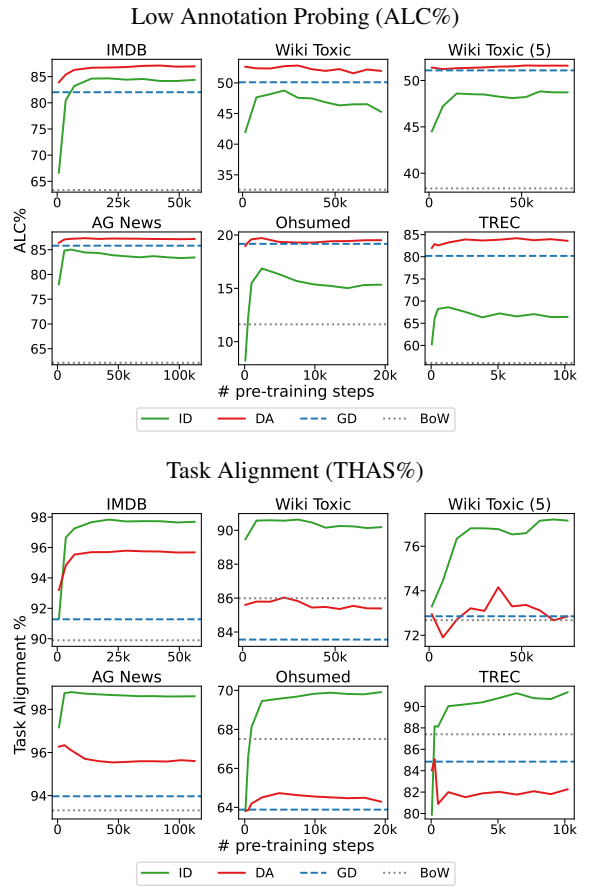


Figure 7: GPT-2 representation performance as a function of PT steps.

evolves with continued pre-training. The learning dynamics for GPT-2 mirror those observed with BERT. However, ID representations generally remain below the performance of GD representations. This may be due to the extensive pre-training of the standard GPT-2 model on WebText, which establishes a high-performance baseline that is difficult to surpass.

Metric	ncvg	E[acc$_{L1}$]
ΔProbe	68.84	48.78
ΔTHAS	27.26	49.43

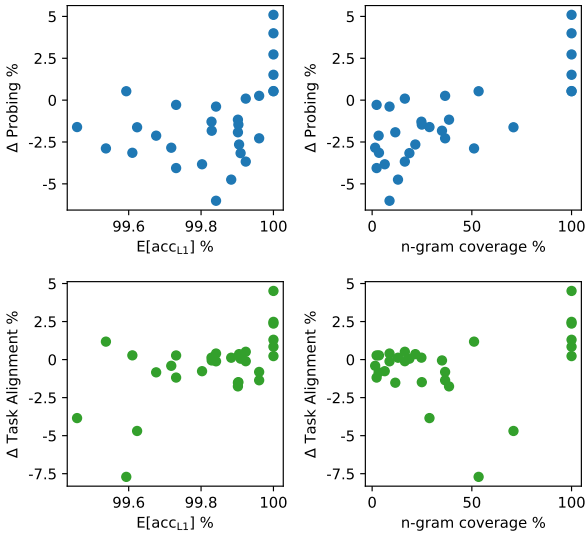


Figure 8: Correlation between distribution similarity and representation quality improvements for GPT-2 domain-adapted representations. We report Spearman's coefficients in the table above and the corresponding scatter plots bellow.

Finally, Fig. 8 presents the correlation between domain similarity metrics and representation gains achieved through GPT-2 pre-training. As in the BERT case, we observe a correlation between these factors, although the coefficients are generally lower (particularly for n-gram coverage). The scatter plots also reveal a greater number of outliers in the GPT-2 results.

Overall, these findings suggest that the observed phenomena are consistent across transformer-based models, reinforcing the generality of our conclusions.

A.4 Clustering Quality

To analyze the changes in the embedding structures produced by pre-training from the clustering quality perspective, we computed DBI scores at

various granularity levels. We evaluated the hierarchical clustering partitions employed to compute task alignment. Due to the computational cost of DBI, we only computed scores every 25 partitions.

In Fig. 9 (right), we report DBI curves as a function of the number of clusters in the partitions for the same representations considered in the task alignment curves. Lower DBI scores indicate better clustering partitions. Table 9 shows the ADBI aggregate scores, i.e. the area under the DBI curves, corresponding to the eight feature extraction strategies presented in A.6.

Domain-adaptive PT has a limited impact on clustering quality. Interestingly, for coarse partitions with fewer clusters (early points in the curves), clustering quality often declines relative to the GD baseline. In contrast, in-domain PT alters the cluster structures leading to larger differences in DBI scores compared to the baseline. This is expected, as the clustering structures, starting from a random spatial distribution, evolve as PT progresses.

In general, PT decreases the quality of clusters. For a given representation, whether using in-domain PT or domain-adaptive PT, DBI scores tend to worsen as the number of training epochs increases. This illustrates how the spatial encoding that embeddings undergo during PT does not translate into compactness and separability of the induced structures. Instead, the induced structures do not lend themselves to straightforward assumptions about the global structure of the embedding space.

The behavior of this embedding space property is uncorrelated with task alignment and probing, as the relative ranking of representations is not preserved. This observation is consistent with the findings in Gonzalez-Gutierrez et al. (2023).

A.5 Curves for Aggregate Evaluation Metrics

Fig. 9 illustrates the curves used to calculate the aggregate metrics presented in §2 and A.4. The left of the figure shows the first 500 points of the task alignment curve used to compute THAS. The center of the figure presents the low-annotation probing learning curves used to compute ALC. To the right, DBI curves as a function of cluster granularity define the clustering quality aggregate metric.

For these examples, we used pre-trained text representations using a single feature extraction strategy (the last layer with token average pooling) of our four binary classification tasks. We denote BERT pre-trained with the standard corpus as $\text{BERT}_{\text{BASE}}$. Models pre-trained in-domain are denoted by the dataset as a subscript (e.g., $\text{BERT}_{\text{IMDb}}$ for a model pre-trained solely on IMDb). Domain-adapted models indicate both corpora in subscripts, as in $\text{BERT}_{\text{BASE+S140}}$, which represents $\text{BERT}_{\text{BASE}}$ continued with Sentiment140. Epoch counts are included in the model notation (e.g., $\text{BERT}_{\text{IMDb-80}}$ for the 80-epoch IMDb model).

A.6 Comparison of Feature Extraction Strategies

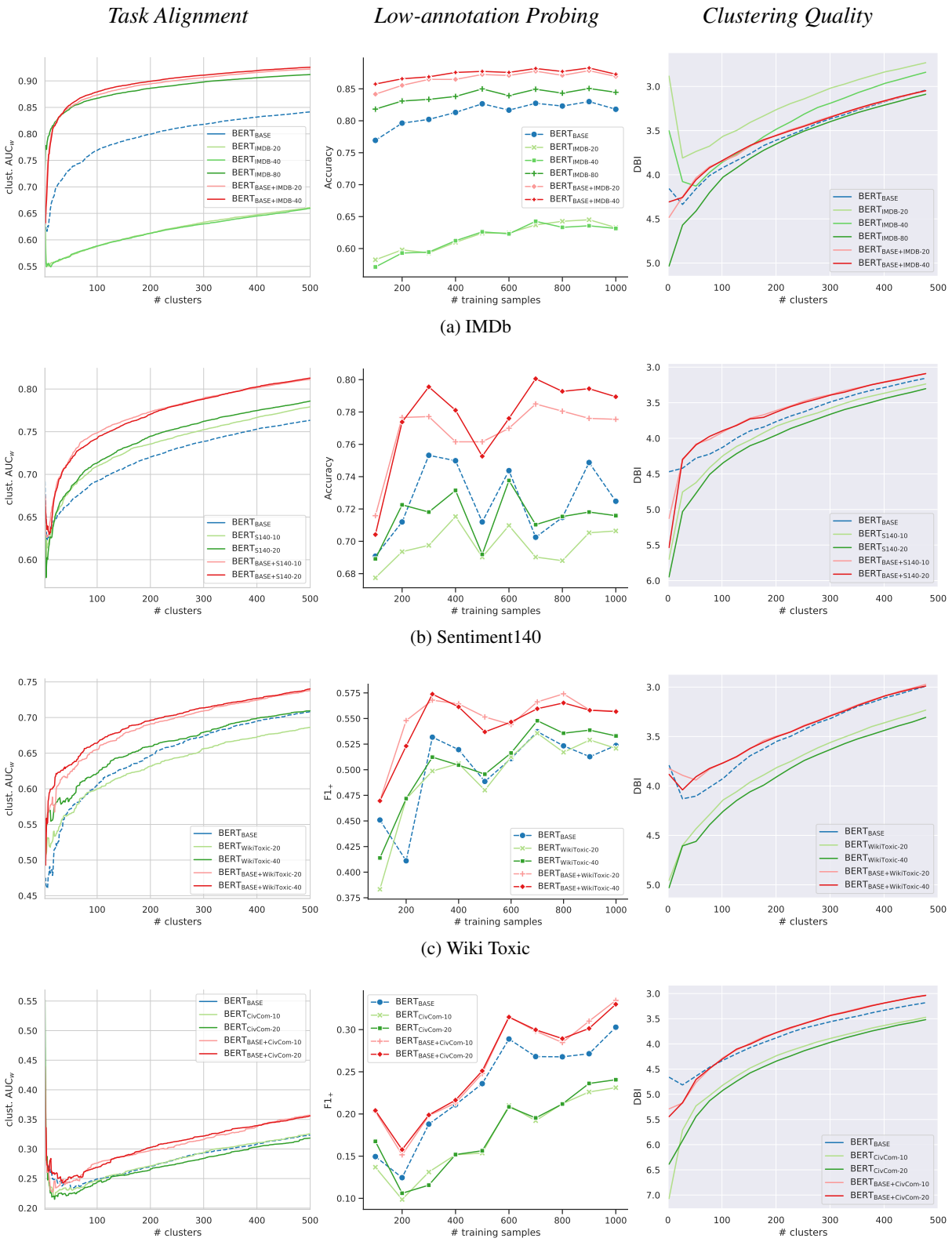
To better understand the role of model's layers and tokens during PT, Tables 7, 8 and 9 present aggregate metrics for the same datasets and models as in §2, evaluated using different feature extraction strategies.

We consider four layer extraction methods: last layer ($_1$), second-to-last ($_2$), concatenating the last four ($_{\text{cat1:4}}$) and averaging all twelve ($_{\mu 1:12}$); in combination to two token pooling strategies: average of all tokens ($^{\mu}$), and taking the `[CLS]` token alone ($^{\text{CLS}}$).

Interestingly, the conclusions in §2 remain consistent across the various representation functions, indicating that embedding improvement during PT affects all the layer and token representations in the model.

Analyzing BERT's token representations reveals that the general domain BERT achieves its strongest performance when extracting embeddings from all the tokens, outperforming the representations derived from the `[CLS]` token, regardless of the layer. This pattern remains consistent for ID representations, with the exception of models pre-trained on Wiki Toxic. This outcome is expected, as the NSP objective, which primarily leverages the `[CLS]` token, was not computed during PT. Interestingly, DA models generate relatively strong `[CLS]` token representations. This result aligns with the findings of Liu et al. (2019b), who suggests that MLM is sufficient for effective PT.

Regarding layer selection, feature extraction typically favors upper layers (Devlin et al., 2019; Reimers and Gurevych, 2019), as they are more specialized than lower layers (Ethayarajh, 2019). From the perspective of our metrics, this holds true in these PT scenarios, although the differences between layers are not very pronounced. Strategies that incorporate lower layers, such as averaging all 12, often produce strong representations and even outperform upper-layer strategies for datasets like Civil Comments or Wiki Toxic.



(a) IMDb

(b) Sentiment140

(c) Wiki Toxic

(d) Civil Comments

Figure 9: Representation performance curves for different datasets using BERT embeddings generated from the last layer with token average pooling. Representations are produced under three PT scenarios: general domain (dashed blue), in-domain (green), and domain-adaptive (red). Multiple pre-training stages are shown with the number of epochs indicated in the representation name.

Dataset	PT	Ep.	Task Alignment %							
			μ_1	CLS_1	μ_2	CLS_2	$\mu_{\text{cat1:4}}$	$\text{CLS}_{\text{cat1:4}}$	$\mu_{\mu 1:12}$	$\text{CLS}_{\mu 1:12}$
IMDb	ID	20	88.86	87.72	88.82	87.34	88.86	87.45	88.71	87.49
		40	88.93	87.38	88.72	86.99	88.83	87.06	88.78	87.21
		80	97.51	95.74	97.50	95.20	97.10	94.94	97.04	94.84
	GD		95.36	94.56	95.31	93.45	95.25	93.43	94.58	92.49
	DA	20	97.72	98.70	97.78	98.95	97.71	98.89	96.37	98.51
		40	97.80	99.04	97.90	99.20	97.79	99.15	96.55	98.87
Sentiment140	ID	10	93.05	92.79	93.01	92.54	92.87	92.73	92.63	92.60
		20	93.20	92.66	93.23	92.72	93.12	92.71	92.83	92.52
	GD		92.43	92.36	92.40	92.10	92.68	91.86	92.20	91.34
	DA	10	94.02	94.87	94.04	94.92	94.18	94.80	93.40	94.18
		20	94.10	94.84	94.13	95.00	94.32	94.93	93.45	94.25
Wiki Toxic	ID	20	90.11	90.60	90.12	90.69	90.50	90.59	90.76	89.59
		40	90.96	91.86	90.99	92.11	91.00	92.01	91.07	91.19
	GD		91.06	86.54	90.49	86.64	90.75	86.72	91.07	87.26
	DA	20	91.91	91.09	91.90	92.33	91.96	91.80	91.85	91.93
		40	92.08	91.12	91.91	92.63	92.18	91.62	92.02	91.75
Civil Comments	ID	10	74.27	74.25	74.33	72.25	73.89	74.55	74.29	73.68
		20	73.96	74.20	73.77	72.68	74.22	74.23	74.31	74.05
	GD		73.98	71.33	73.23	71.46	73.69	71.28	74.21	71.87
	DA	10	76.00	76.69	75.93	75.63	75.61	76.75	75.43	74.92
		20	75.93	76.86	75.85	75.26	75.49	77.21	75.51	75.32

Table 7: Task alignment of pre-trained BERT embeddings across different datasets and feature extraction strategies.

Dataset	PT	Ep.	ALC %							
			μ_1	CLS_1	μ_2	CLS_2	$\mu_{\text{cat1:4}}$	$\text{CLS}_{\text{cat1:4}}$	$\mu_{\mu 1:12}$	$\text{CLS}_{\mu 1:12}$
IMDb (acc)	ID	20	61.60	55.34	61.89	54.27	62.17	55.41	63.15	56.01
		40	61.70	55.11	61.63	54.12	62.62	55.10	62.99	55.91
		80	84.11	81.33	83.98	80.55	83.86	80.99	83.04	80.53
	GD		81.39	78.52	81.23	77.09	82.14	78.69	80.15	76.61
	DA	20	86.72	88.03	86.67	88.51	87.04	89.03	84.92	87.87
		40	87.12	89.40	87.35	89.30	87.57	89.73	85.42	89.10
Sentiment140 (acc)	ID	10	70.19	68.73	69.74	68.59	70.90	70.25	69.63	68.49
		20	71.92	70.41	71.50	70.04	72.35	71.97	71.44	71.05
	GD		71.84	70.97	72.52	69.38	73.58	71.86	71.00	69.18
	DA	10	76.71	77.09	76.80	75.89	78.47	78.31	76.29	77.24
		20	77.48	77.16	77.60	76.53	78.86	79.06	77.30	77.52
Wiki Toxic ($F1_+$)	ID	20	49.96	52.96	49.54	53.21	49.63	53.11	49.17	52.81
		40	50.54	55.75	50.70	56.42	50.97	56.13	50.37	56.31
	GD		50.09	48.18	50.10	46.58	51.41	48.11	52.60	48.49
	DA	20	54.55	54.00	55.01	54.98	54.81	55.72	55.43	54.50
		40	54.68	53.75	54.51	54.76	55.05	55.51	55.39	54.34
Civil Comments ($F1_+$)	ID	10	18.78	20.44	17.44	20.29	19.01	20.26	19.84	20.67
		20	18.38	20.23	17.90	20.40	19.22	19.76	19.79	20.40
	GD		23.63	19.63	23.08	17.52	23.92	17.88	24.61	19.43
	DA	10	25.98	26.32	25.56	26.04	26.99	26.30	26.96	27.36
		20	25.94	26.16	25.64	26.18	26.88	26.15	26.12	27.69

Table 8: Area under the learning curve (ALC) of low-annotation probes for different feature extraction strategies. We report accuracy for balanced datasets and F1 of the target class for imbalanced datasets.

Dataset	PT	Ep.	ADBI							
			μ_1	CLS$_1$	μ_2	CLS$_2$	$\mu_{\text{cat1:4}}$	CLS$_{\text{cat1:4}}$	$\mu_{\mu 1:12}$	CLS$_{\mu 1:12}$
IMDb	ID	20	1.18	1.07	1.18	0.97	1.16	0.99	1.12	1.01
		40	1.21	0.90	1.21	0.83	1.20	0.84	1.15	0.87
		80	1.39	0.92	1.38	0.88	1.30	0.89	1.25	0.93
	GD		1.35	1.36	1.32	1.32	1.31	1.35	1.29	1.33
	DA	20	1.34	1.33	1.34	1.33	1.32	1.37	1.29	1.35
		40	1.34	1.33	1.34	1.32	1.32	1.36	1.28	1.35
Sentiment 140	ID	10	1.39	1.26	1.41	1.28	1.38	1.28	1.37	1.33
		20	1.41	1.27	1.43	1.29	1.40	1.29	1.39	1.33
	GD		1.36	1.27	1.35	1.23	1.36	1.26	1.36	1.25
	DA	10	1.36	1.24	1.36	1.25	1.36	1.24	1.35	1.19
		20	1.37	1.24	1.36	1.26	1.37	1.25	1.35	1.20
Wiki Toxic	ID	20	1.36	1.27	1.38	1.27	1.18	1.28	1.14	1.23
		40	1.40	1.29	1.40	1.28	1.20	1.30	1.15	1.23
	GD		1.20	1.24	1.23	1.26	1.21	1.29	1.16	1.25
	DA	20	1.26	1.23	1.24	1.26	1.23	1.26	1.16	1.24
		40	1.27	1.25	1.25	1.26	1.23	1.27	1.16	1.25
Civil Comments	ID	10	1.46	1.36	1.49	1.32	1.42	1.36	1.44	1.40
		20	1.48	1.39	1.50	1.34	1.44	1.39	1.45	1.40
	GD		1.28	1.32	1.31	1.32	1.30	1.36	1.24	1.31
	DA	10	1.32	1.33	1.26	1.36	1.32	1.35	1.25	1.33
		20	1.32	1.35	1.26	1.37	1.32	1.37	1.25	1.34

Table 9: Clustering quality measured by the area under the DBI scores for pre-trained embeddings using diverse feature extraction strategies. Lower scores mean better clustering.