

# Understanding Domain Learning in Language Models Through Subpopulation Analysis

Zheng Zhao Yftah Ziser Shay B. Cohen

Institute for Language, Cognition and Computation

School of Informatics, University of Edinburgh

10 Crichton Street, Edinburgh, EH8 9AB

{zheng.zhao, yftah.ziser}@ed.ac.uk, scohen@inf.ed.ac.uk

## Abstract

We investigate how different domains are encoded in modern neural network architectures. We analyze the relationship between natural language domains, model size, and the amount of training data used. The primary analysis tool we develop is based on subpopulation analysis with Singular Vector Canonical Correlation Analysis (SVCCA), which we apply to Transformer-based language models (LMs). We compare the latent representations of such a language model at its different layers from a pair of models: a model trained on multiple domains (an experimental model) and a model trained on a single domain (a control model). Through our method, we find that increasing the model capacity impacts how domain information is stored in upper and lower layers differently. In addition, we show that larger experimental models simultaneously embed domain-specific information as if they were conjoined control models. These findings are confirmed qualitatively, demonstrating the validity of our method.

## 1 Introduction

Pre-trained language models (PLMs) have become an essential modeling component for state-of-the-art natural language processing (NLP) models. They process text into latent representations in such a way that allows an NLP practitioner to seamlessly use these representations for prediction problems of various degrees of difficulty (Wang et al., 2018, 2019). The opaqueness in obtaining these representations has been an important research topic in the NLP community. PLMs, and more generally, neural models, are currently studied to understand their process and behavior in obtaining their latent representations. These PLMs are often trained on large datasets, with inputs originating from different sources. In this paper, we further develop our understanding of how neural networks obtain their latent representation and study the effect of learn-

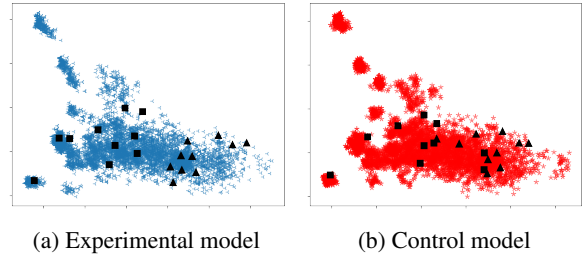


Figure 1: An example of a visualization used with our subpopulation analysis tool. The experimental model, which includes all domain data, separates in its latent representations words related to the `Books` domain ( $\blacktriangle$ ) from general words ( $\blacksquare$ ). The control model, on the other hand, mixes them together.

ing from various domains on the characteristics of the corresponding latent representations.

Texts come from various domains that differ in their writing styles, authors and topics (Plank, 2016). In this work, we follow a simple definition of a domain as *a corpus of documents sharing a common topic*. We rely on a simple tool of subpopulation analysis to compare and contrast latent representations obtained with and without a specific domain. Our analysis relies on constructing two types of models: *experimental* models, from multi-domain data, and *control* models, from single-domain data. Figure 1 describes an example in which this analysis is applied to study the way embeddings for domain-specific words cluster together in the experimental and control model representations.

We believe training in an implicit multi-domain setup is widespread and often overlooked. For example, SQuAD (Rajpurkar et al., 2016), a widely used question-answering dataset composed of Wikipedia articles from multiple domains, is often referred to as a single-domain dataset in domain adaptation works for simplicity (Hazen et al., 2019; Shakeri et al., 2020; Yue et al., 2021). This scenario is also common in text summarization, where

many datasets consist of a bundle of domains for news articles (Grusky et al., 2018), academic papers (Cohan et al., 2018; Fonseca et al., 2022), and do-it-yourself (DIY) guides (Cohen et al., 2021). While models that learn from multiple domains are frequently used, their nature and behavior have hardly been explored.

Our work sheds light on the way state-of-the-art multi-domain models encode domain-specific information. We focus on two main aspects highly relevant for many training procedures: model capacity and data size. We discover that model capacity, indicated by the number of its parameters, strongly impacts the amount of domain-specific information multi-domain models store. This property might explain the performance gains of larger models (Devlin et al., 2019; Raffel et al., 2020; Clark et al., 2020; Srivastava et al., 2022). While this paper focuses on studying the effect of domains on latent representations, the subpopulation analysis tool could be used for studying other NLP setups, such as multitask and multimodal learning.<sup>1</sup>

## 2 Methodology

For an integer  $n$ , we denote by  $[n]$  the set  $\{1, \dots, n\}$ . Our analysis tool assumes a distribution  $p(\mathbf{X})$  from which a set of examples  $\mathcal{X} = \{\mathbf{x}^{(i)} \mid i \in [n]\}$  is drawn. It also assumes a family of binary indicators  $\pi_1, \dots, \pi_d$  such that  $\pi_i(\mathbf{x})$  indicates whether the example  $\mathbf{x}$  satisfies a certain *subpopulation* attribute  $i$ . For example, in this paper we focus on domain analysis, so  $\pi_5$  could indicate if an example belongs to a BOOKS domain.

We denote by  $\mathcal{X}|_{\pi_i}$  the set  $\{\mathbf{x}^{(j)} \mid \pi_i(\mathbf{x}^{(j)}) = 1\}$ , the subset of  $\mathcal{X}$  that satisfies attribute  $i$ . Unlike standard diagnostic classifier methods (Belinkov et al., 2017a,b; Giulianelli et al., 2018), rather than building a model to *predict* the attribute, we perform subpopulation analysis by training a set of models:  $\mathbf{E}$ , trained from  $\mathcal{X}$  (the *experimental* model), and  $\mathbf{C}_i$ , trained from  $\mathcal{X}|_{\pi_i}$  (the *control* model). We borrow the terminology of “experimental” and “control” from experimental design such as in clinical trials (Hinkelmann and Kempthorne, 2007). The experimental model corresponds to the experimental (or “treatment” in the case of medical trials) group in such trials and the control model corresponds to the control group. Unlike a standard experimental design, rather than comparing a function (such as

squared difference) between the outcomes of the two groups to calculate a statistic with an underlying distribution, we instead calculate the similarity values between the representations of the two models. Our analysis is also related to Representational Similarity Analysis (Dimsdale-Zucker and Ranganath, 2018), aimed at studying similarities (across different experimental settings) between activation levels in brain neurons.

Through their latent representations, the set of models  $\mathbf{C}_i$  represent the information that is captured about  $p(\mathbf{X})$  from the relevant subpopulation of data. By comparing the different models to each other, we can learn what information is captured in the latent representations when a subset of the data is used and whether this information is different from the one captured when the whole set of data is used. With a proper control for model size and subpopulation sizes, we can determine the relationship between the different attributes  $\pi_i$  and the corresponding representations in different model components. The remaining question now is *how do we compare these representations?* Here, we follow previous work (Saphra and Lopez, 2019; Bau et al., 2019; Kudugunta et al., 2019), and apply Singular Vector Canonical Correlation Analysis (SVCCA; Raghu et al. 2017) to the latent representations of the experimental and control models.

We assume that each example  $\mathbf{x}^{(i)}$  is associated with a latent representation  $\mathbf{h}_j^{(i)}$  given by  $\mathbf{C}_j$ . For example, this could be the representation in the embedding layer for the input example, or the representation in the final pre-output layer. We define  $\mathcal{H}_j$  to be a set of latent representations  $\mathcal{H}_j = \{\mathbf{h}_j^{(k)} \mid k \in [n]\}$  for model  $\mathbf{C}_j$ . We define  $\mathcal{H}_j|_{\pi_i} = \{\mathbf{h}_j^{(k)} \mid \pi_i(\mathbf{x}^{(k)}) = 1\}$  – the latent representations of  $\mathbf{C}_j$  for which attribute  $i$  fires. Similarly, we define  $\mathcal{H}_0$  for the model  $\mathbf{E}$ . We calculate the SVCCA value between subsets of  $\mathcal{H}_0$  and subsets of  $\mathcal{H}_j$  for  $j \geq 1$ . The procedure of SVCCA in this case follows:

- Performing Singular Value Decomposition (SVD) on the matrix forms of  $\mathcal{H}_0$  and  $\mathcal{H}_j$  (matching the representations in each through the index of the example  $\mathbf{x}^{(i)}$  from which they originate). We use the lowest number of principal directions that preserve 99% of the variance in the data to project the latent representations.
- Performing Canonical Correlation Analysis (CCA; Hardoon et al. 2004) between the pro-

<sup>1</sup>Our code is available at: [https://github.com/zsquaredz/subpopulation\\_analysis](https://github.com/zsquaredz/subpopulation_analysis)

jections of the latent representations from the SVD step, and calculating the average correlation value, denoted by  $\rho_{0j}$ .

The SVD step, which may seem redundant, is actually crucial, as it had been shown that low variance directions in neural network representations are primarily noise (Raghu et al., 2017; Frankle and Carbin, 2019). The intensity of  $\rho_{0j}$  indicates the level of overlap between the latent representations of each model (Saphra and Lopez, 2019).

In the rest of this paper, we use the tool of subpopulation analysis with  $\mathbf{E}/\mathbf{C}_i$  as above for the case of domain learning in neural networks. We note that each time we use this tool, the following decisions need to be made: (a) what training set we use for each  $\mathbf{E}$  and  $\mathbf{C}_i$ ; (b) the subset of  $\mathcal{H}_j$  for  $j \geq 0$  for which we perform the similarity analysis; (c) the component in the model from which we take the latent representations. For (c), the component can be, for example, a layer. Indeed, for most of our experiments, we use the first and last layer to create the latent representation sets, as they stand in stark contrast to each other in their behavior (see § 4). We provide an illustration of our proposed pipeline in Figure 2. We are particularly interested in studying the effect of two aspects of learning: dataset size and model capacity.

**The case of domains** In this paper, we define a domain as a corpus of documents with a common topic. Since a single massive web-crawled corpus used to pre-train language models usually contains many domains, we examine to what extent domain-specific information is encoded in the pre-trained model learned on this corpus. Such domain membership is indicated by our attribute functions  $\pi_i$ . For example, we may use  $\pi_5(\mathbf{x})$  to indicate whether  $\mathbf{x}$  is an input example from the domain `Books`. Given this notion of a domain, we can readily use subpopulation analysis through experimental and control models to analyze the effect on neural representations of learning from multiple domains or a single domain.

### 3 Experimental Setup

**Data** We use the Amazon Reviews dataset (Ni et al., 2019), a dataset that facilitates research in tasks like sentiment analysis (Zhang et al., 2020), aspect-based sentiment analysis, and recommendation systems (Wang et al., 2020). The reviews in this dataset are explicitly divided into

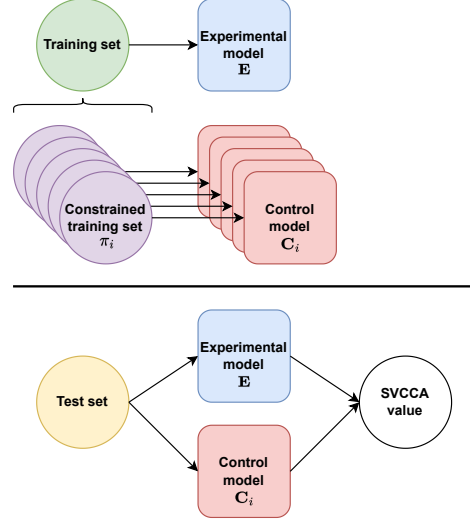


Figure 2: A diagram explaining the analysis we perform. At the top, during training, we create two sets of models from constrained datasets (based on different  $\pi_i$ ) and a dataset that is not constrained. The result of this training is two set of models, the experimental model ( $\mathbf{E}$ ) and control models ( $\mathbf{C}_i$ ). To perform the similarity analysis, we compute latent representation from a common test set for both models, and then run SVCCA (bottom).

different product categories that serve as domains, which makes it a natural testbed for many multi-domain studies. A noteworthy example of a research field that heavily relies on this dataset is domain adaptation (Blitzer et al., 2007; Ziser and Reichart, 2018; Du et al., 2020; Lekhtman et al., 2021; Long et al., 2022), which is the task of learning robust models across different domains, closely related to our research.<sup>2</sup> We sort the domains by their review counts and pick the top five, which results in: `Books`, `Clothing Shoes and Jewelry`, `Electronics`, `Home and Kitchen`, and `Movies and TV` domains. To further validate our data quality, we use the 5-core subset of the data, ensuring that all reviewed items have at least five reviews authored by reviewers who wrote at least five reviews.

A representative dataset sample is presented in Table 1. We consider the different domains within the Amazon review dataset as *lexical domains*, i.e., domains that share a similar textual structure and functionality but differ with respect to their vocabulary. For example, we see the review snippet from the `Books` domain contains an aspect (“ending”) for which a negative sentiment is conveyed (“didn’t have a proper”). Similarly, we find an aspect (“han-

<sup>2</sup>We use the latest version of the dataset, consisting reviews from 1996 up to 2019.

---

**Books:** ... the book didn't have a proper ending but rather a rushed attempt to conclude the story and put everyone away neatly ...

**Clothing:** ... clearly of awful quality, the design and paint was totally wrong, the mask was short and stumpy as well as slightly deformed and bent to the left ...

**Home:** ... there are no handles, and the plastic gets too hot to hold, so you have to awkwardly pour by the top ...

---

Table 1: A representative sample of review snippets.

dle”) with a corresponding conveyed sentiment (“too hot”) for the `Home` domain. We can see this shared pattern across all domains, with different aspects and sentiment terms. We would not expect this to be the case for other datasets, which might have different differentiators for domains. For example, Amazon reviews and Wikipedia pages on `Books` domain may have a similar vocabulary, however, a review is more likely to convey sentiment toward a particular book, and a Wikipedia article is more likely to focus on describing the book. Thus, the Amazon Reviews dataset is an ideal testbed for our analysis.

In addition to the Amazon Reviews dataset, we experimented on the WikiSum dataset (Cohen et al., 2021) to further validate our findings. The WikiSum dataset is a coherent paragraph summarization dataset based on the WikiHow website.<sup>3</sup> WikiHow consists of do-it-yourself (DIY) guides for the general public, thus is written using simple English and ranges over many domains. Similar to Amazon Reviews, we also pick the top five domains for our experiments: `Education`, `Food`, `Health`, `Home`, and `Pets`. Since the dataset is designed for summarization, we concatenate the document and summary together for our MLM task. We present the results for this dataset at the end of § 4.

**Task** We study the language modeling task to understand the nature of multi-domain learning better. More precisely, we experiment with the masked language modeling (MLM) task, which randomly masks some of the tokens from the input, then predicts the masked word based on the context as the training objective. We focus on the MLM task as it is a prevalent pre-training task for many standard models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) that serve as building blocks for many downstream tasks. Using examples from a set of pre-defined domains, we train a BERT model from scratch to fully control our ex-

periment and isolate the effect of different domains. This is crucial since a pre-trained BERT model is already trained on multiple domains, hence hard to drive correct conclusions through our analysis from such a model. Moreover, recent studies (Magar and Schwartz, 2022; Brown et al., 2020) showed the risk of exposure of large language models to test data in the pre-training phase, also known as *data contamination*.

**Model** We use the BERT<sub>BASE</sub> (Devlin et al., 2019) architecture for all of our experiments. We train two types of models: the experimental model **E**, trained on all five domains with the MLM objective, and the control model **C<sub>i</sub>** for  $i \in [5]$  trained on the  $i$ th domain. We are particularly interested in the effect of two aspects on the model representation: model capacity and data size. We use the capacity of 100% for BERT<sub>BASE</sub> size. BERT<sub>BASE</sub> has 768-dimensional vectors for each layer, adding up to a total of 110M parameters. We also experiment with a reduced model capacity of 75%, 50%, 25%, and 10% by reducing the dimension of the hidden layers. We follow Devlin et al. (2019) design choices, e.g., 12 layers with 12 attention heads per layer. We set the base training data size (100%) for **E** to be 50K, composed of 10K reviews per domain. Each **C<sub>i</sub>** is trained on single domain data containing 10K reviews. **E** and **C<sub>i</sub>** *share all the examples of domain i*. To study the effect of data size on model representation, we take subsets from the data split and create smaller datasets: a 10% split and a 50% split. We also create a 200% split to simulate the case with abundant data. We provide additional details about our training procedure in Appendix A.

## 4 Experiments and Results

Our research questions (RQs) examine how domain-specific information is encoded in **E** by calculating its SVCCA score with **C<sub>i</sub>** for a specific  $i$ . For a given domain, we use a held-out test set for getting the experimental and control model representations as an input for the SVCCA method. Intuitively, a high SVCCA score between **E** and **C<sub>i</sub>** indicates **E** stores domain-specific information for domain  $i$ , as **C<sub>i</sub>** was trained solely on data from domain  $i$ . A low SVCCA score between **E** and **C<sub>i</sub>** could mean one of two things: a) **E** can generalize to data from  $d_i$  without explicitly storing domain-specific information about it, or b) **E** can not store information about **C<sub>i</sub>**, as a result of, for

<sup>3</sup><https://www.wikihow.com>



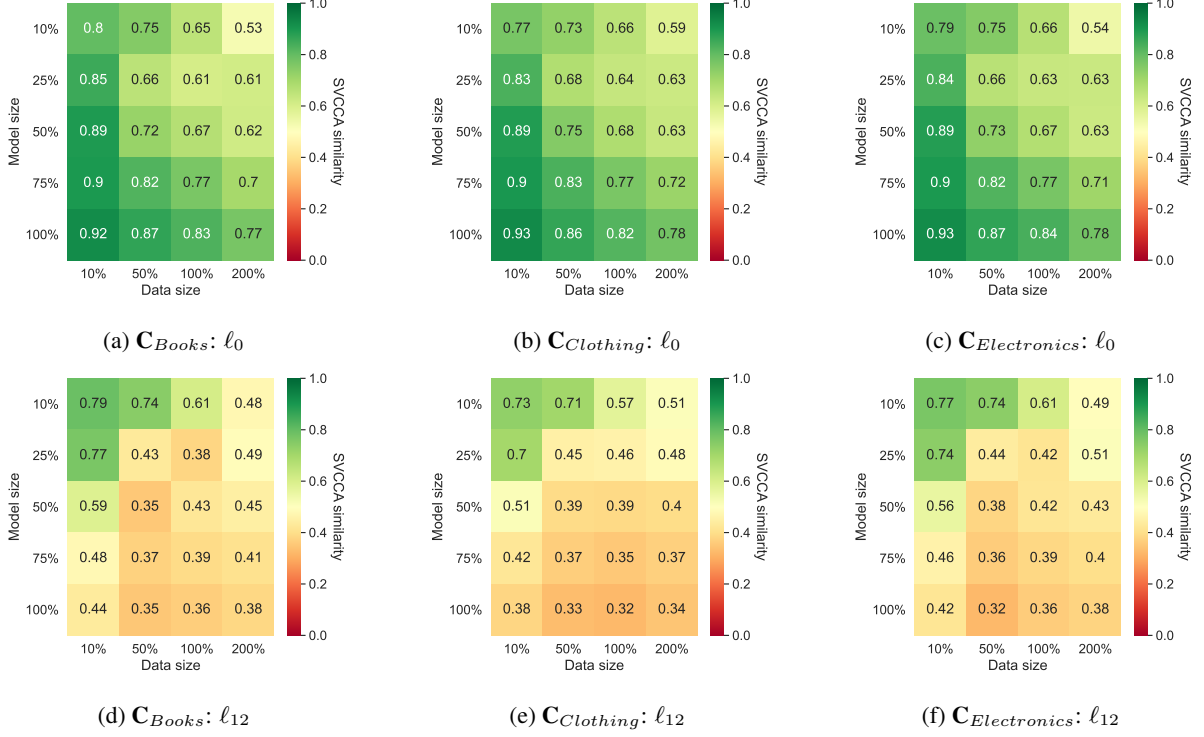


Figure 3: The SVCCA scores between  $\mathbf{E}$  and different  $\mathbf{C}_i$ s for different data sizes and model capacities. We only display for three domains here, and we provide the rest in Appendix B.2. The top row presents the results for the embedding layer  $\ell_0$ , and the bottom row presents them for the last layer  $\ell_{12}$ .

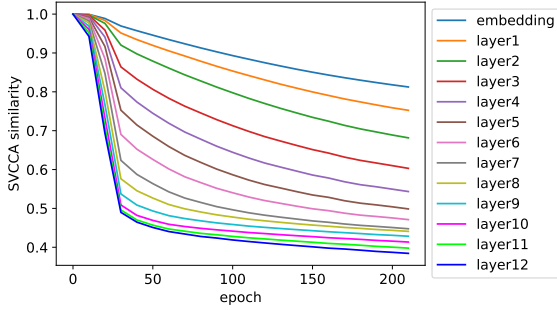


Figure 4: Training dynamics for all layers between  $\mathbf{E}$  and  $\mathbf{C}_{Books}$ . Here both model and data size are 100%.

example, lack of model capacity. The way to distinguish between the two is subjective and depends on whether one finds  $\mathbf{E}$  performance when applied to data from  $d_i$  to be satisfactory. This paper analyzes how information is stored at the model layers. As we inspect highly complex models consisting of multiple layers, it is challenging to determine to what extent a certain layer contributes to a model’s overall performance. For those reasons, when comparing equivalent layers of different models, we focus on the amount of domain-specific information encoded in  $\mathbf{E}$  for a given layer. With these preliminaries in mind, we are now ready to ask the

following research questions:

#### RQ1: How does the similarity between the corresponding layers in $\mathbf{E}$ and $\mathbf{C}$ evolve over training?

We perform an iterative comparison between the  $\mathbf{E}$  and  $\mathbf{C}_i$  for each  $i \in [5]$ . After each epoch, we calculate the SVCCA score between corresponding layers of the models, i.e., layer  $j$  of  $\mathbf{E}$  is compared to layer  $j$  of  $\mathbf{C}_i$ . As  $\mathbf{E}$  is trained on more data points than  $\mathbf{C}_i$ , and both use the same batch size, for any given epoch,  $\mathbf{E}$  had more weights’ updates than  $\mathbf{C}_i$ . More precisely, after the  $k$ th epoch,  $\mathbf{C}_i$  and  $\mathbf{E}$  had completed  $k$  passes on data points from  $d_i$ , but  $\mathbf{E}$  used additional data points from the rest of the domains. We choose this alignment to examine the effect of the additional training data drawn from other domains.

Figure 4 presents the training dynamics analysis for the **Books** domain (we denote the **Books** control model as  $\mathbf{C}_{Books}$ ). We include training dynamics analyses of other control models and domains in Appendix B.1, as they demonstrate similar trends. Since both  $\mathbf{C}_{Books}$  and  $\mathbf{E}$  are initialized with the same weights, the initial SVCCA score is 1 for all layers before training. We observe that as training progresses, the SVCCA values of higher layers

(closer to the output) consistently become lower compared to the first layer. The order of SVCCA values is almost perfectly preserved with respect to the order of the layers in the network. The separation is higher for lower layers, with higher layers receiving similar SVCCA values. This is evidence that *E stores more domain-specific information in lower layers than in deeper layers throughout the training procedure*. Singh et al. (2019), who researched the nature of multilingual models, observed a similar pattern of dissimilarity in deeper layers for multilingual model representations of parallel sentences in different languages.

The alignment between the similarity of the layer pairs (**E** and **C**) and their depth also exists for models with random weights. It can be partially attributed to the mathematical artifact of decreasing correlation values for layers that are deeper because of the use of nonlinear activation units. To see to what extent this artifact plays a role in this alignment, we created ten models with random weights (no training, so there is no longer an experimental/control distinction) and calculated SVCCA between all 45 pairs for the first and last layers. We discovered that the mean difference between SVCCA scores of the first layer comparison and the last layer comparison is 0.139 (with a standard deviation of 0.001 over 45 pairs). In Figure 4, the difference is much larger when comparing the control model to the experimental model (0.428), indicating that the difference in layer SVCCA score cannot be only attributed to the mathematical artifact of increasing depth with more nonlinear activation. We still note that one should exercise caution when using linear methods, such as SVCCA, to analyze nonlinear processes.

The observed training dynamics motivates us to focus on the embedding layer ( $\ell_0$ ) and final layer ( $\ell_{12}$ ) for the rest of our analysis, as they serve as a lower bound ( $\ell_0$ ) and an upper bound ( $\ell_{12}$ ) with respect to the SVCCA scores of **C**<sub>*i*</sub> and **E** throughout the training process. In addition, those layers have interesting attributes that we would like to explore.  $\ell_0$ , a non-contextualized word embeddings layer, is known for encoding mainly lexical information (de Vries et al., 2020; Vulić et al., 2020). The highly contextualized  $\ell_{12}$  is fed directly to the masked word classifier, thus playing a significant role in the MLM task. Our interest in the fully-trained models leads us to the following question:

**RQ2: How do data size and model capacity affect domain encoding in  $\ell_0$  and  $\ell_{12}$ ?** To answer this question, we measure the SVCCA score between variants of **E** and their corresponding **C**<sub>*i*</sub> for different domains. The variants differ with respect to two parameters, data size and model capacity.

Figure 3 presents our results. We observe training the model on larger datasets decreases the SVCCA scores across all model capacities and domains for both  $\ell_0$  and  $\ell_{12}$ . For each data point we add to the control model, we add  $d$  data points to the general model, where  $d - 1$  out of them belong to other domains. This means while we keep a constant ratio between the number of datapoints for the domains, the absolute gap between a given domain and the rest of the domains is growing for larger data sizes. This might explain why adding more data points increase **E** and **C** divergence.

A possible explanation for these trends might be how we define domains. The Amazon reviews dataset is divided by product categories which can be seen as lexical domains (see § 3). More precisely, all the domains share a similar structure and writing style of Amazon product reviews. The differences lie in the vocabulary of each domain. We hypothesize that the *E uses the increased capacity to keep more domain-specific information in  $\ell_0$ , where the lexical information is kept and diverges from C in  $\ell_{12}$ , where the highly contextualized representations are stored*. As we hypothesize that our domains differ mostly with respect to their vocabularies, we refine the mentioned above experiment by raising the following research question:

**RQ3: To what extent does E encode domain-specific information for domain-specific words?**

To shed light on the domains’ lexical nature, we inspect the patterns of domain-specific and general words. Domain-specific words need to appear with at least 20 reviews in the domain in hand and no more than 10 reviews in total for the rest of the other domains. General words must appear in at least 20 reviews in each domain. Those definitions are often used in domain adaptation works to describe domain discrepancy and find adaptable features (Blitzer et al., 2007; Ziser and Reichart, 2017). We provide some examples of domain-specific and general words in Appendix B.3. It is noteworthy that the union of the domain-specific and general words is not the complete vocabulary. To calculate the SVCCA scores for a subset of words, we first apply SVD to all inputs. Then we use the corre-

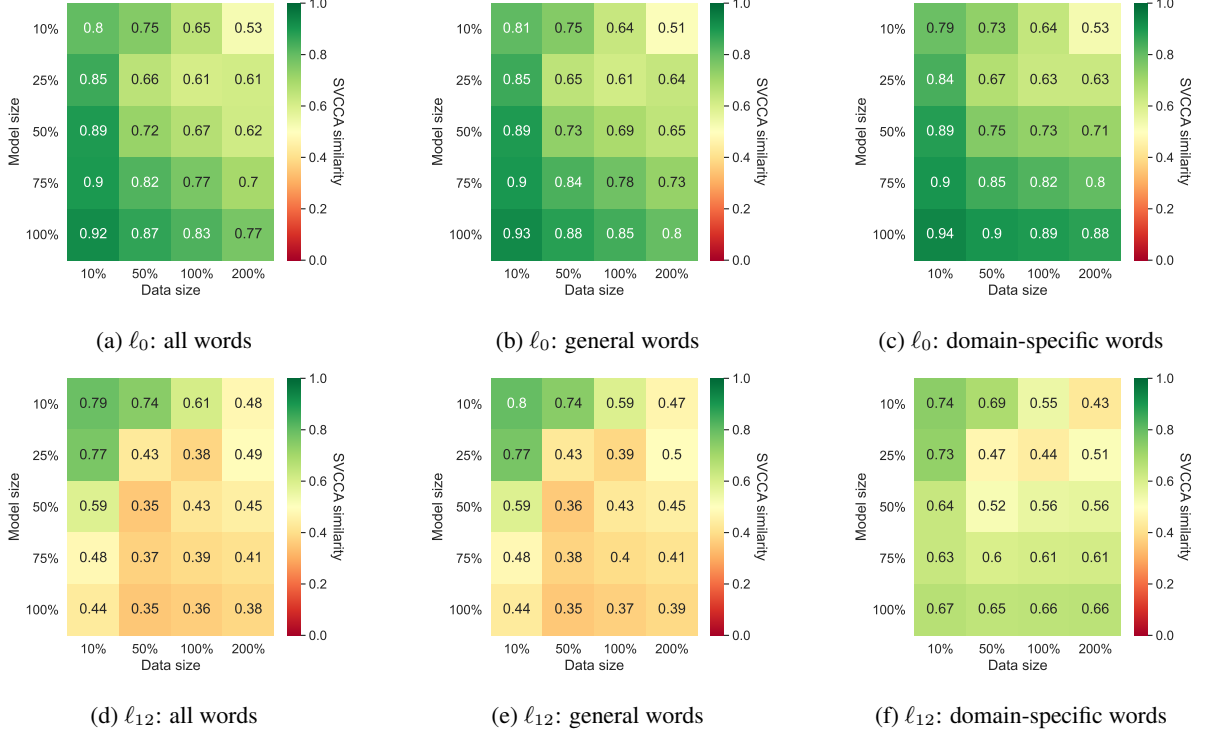


Figure 5: The SVCCA score between  $\mathbf{E}$  and  $\mathbf{C}_{Books}$  for different subsets of tokens. The top row presents the results for the embedding layer  $\ell_0$ , and the bottom row presents them for the last layer  $\ell_{12}$ .

sponding representations of the subset tokens to calculate the CCA similarity.

Figure 5 presents our results for the Books domain.<sup>4</sup> We present the Books domain analysis for all the words taken from RQ2 for reference (on the left-hand side of the figure). We observe high SVCCA scores for domain-specific words for  $\ell_{12}$ . For large data sizes (100% and 200%), the trends of domain-specific words are opposite to the ones of RQ2, i.e.,  $\mathbf{E}$  uses the additional capacity to encode more domain-specific information. This indicates that as model capacity increases,  $\mathbf{E}$  can capture similar information to  $\mathbf{C}_{Books}$  for domain-specific words. This justifies the construction of large language models, mixing multiple subpopulations, as it demonstrates that *if the  $\mathbf{E}$  model has large enough capacity, it separately creates representations for the different subpopulations that are similar to  $\mathbf{C}_i$  model, which is a specialized model for a given domain.* Domain-specific words and their representations are crucial for the success of many NLP tasks, for example, Named Entity Recognition (Rocktäschel et al., 2013; Shang et al., 2018; Gu et al., 2021). We can see that the SVCCA scores for all the words and general words are al-

most identical. These findings make us suspect that word frequency and domain specificity are strongly connected. Indeed, we find out that the average frequency for Books domain-specific words is 75 with a median of 43. For general words, the average is 7696, and the median is 1440, making general words the main factor in the SVCCA scores for all words.

Finally, we would like to ensure the patterns we observe throughout this paper affect the behavior of the model:

**RQ4: Do the observed trends manifest in the models’ behavior?** We conducted two qualitative analyses to understand better if the models’ behavior expresses our findings. For the first analysis, we compare MLM predictions of  $\mathbf{E}$  and  $\mathbf{C}$  to check whether higher SVCCA values are associated with similar word predictions. For  $\ell_0$ , we calculate the k-nearest neighbors of the word embeddings for a given word as a proxy to make predictions. For  $\ell_{12}$ , we follow the standard procedure by feeding the last layer representation to the final MLM classifier in BERT. Table 2 presents our analyses. We can see that for  $\ell_0$ , as we increase the model capacity, we get more similar predictions for both domain-specific and general words. This finding agrees

<sup>4</sup>The rest of the domains exhibit similar patterns. We provide all results in Appendix B.4

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
blackberry	proxy	linux	mac
linux	linux	mac	linux
biologist	peer	blackberry	computers
viking	windows	vista	windows
samsung	servers	xp	xp

(a) 5-nearest neighbors for the domain-specific word **Macintosh** with  $i=\text{Electronics}$ .

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
networks	connections	routers	router
phones	networks	products	networks
devices	ports	systems	connections
problems	computers	mice	computers
models	cables	connections	products

(c) *Other wired and wireless [MASK] I had never had this problem.* The masked word is a domain-specific word **routers** with  $i=\text{Electronics}$ .

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
functioning	riding	functioning	functioning
work	running	work	repair
worked	work	worked	work
playing	walking	looking	riding
responding	cleaning	works	looking

(b) 5-nearest neighbors for the general word **working** with  $i=\text{Home}$  and  $\text{Kitchen}$ .

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
away	apart	apart	aside
apart	off	flat	apart
aside	away	short	down
downhill	downhill	out	back
asleep	asleep	off	along

(d) *Sadly, those hopes began to fall [MASK] shortly after I finished the Prologue.* The masked word is a general word **apart** with  $i=\text{Books}$ .

Table 2: (a) and (b) are the 5-nearest neighbors using the embedding layer weights. (c) and (d) are model predictions using last layer representations. m denotes model capacity. All models here use a data size of 100%.

with the trend in Figure 3 that higher model capacity is associated with higher SVCCA similarity for  $\ell_0$ . For  $\ell_{12}$ , we can see that as model capacity increases, predictions for the general word becomes inconsistent, whereas, for domain-specific words, it is the opposite. This finding also agrees with our findings in RQ2 and RQ3, in which we observe the  $\ell_{12}$  SVCCA values are decreasing for general words as we increase the model capacity and decrease for domain-specific words. We provide additional examples in Appendix B.5.

For the second analysis, we employ principal component analysis (PCA) to reduce the dimension of general and domain-specific representations for  $\ell_0$  and  $\ell_{12}$  for both **E** and **C<sub>Books</sub>**. We provide visualizations in Figure 6. We can see that as model capacity increases,  $\ell_0$  representations of both general and domain-specific words from **E** and **C<sub>Books</sub>** are aligned to a similar subspace. Additionally,  $\ell_{12}$  representations of general words and domain-specific words for both models exhibit opposite behavior: domain-specific words are more aligned with increasing model capacity while general words start to detach. All of these agree with our findings in corresponding SVCCA scores trends in Figure 5. Even though we did not explicitly examine the relations between general and specific words in our work, we can observe that general and domain-specific word representations form different clusters in both models. Those clusters are more separated in  $\ell_0$  than in  $\ell_{12}$ , suggesting that models use their increased capacity to keep more domain-

specific information in  $\ell_0$ .

**WikiSum results** Due to the lack of computational resources required, we only validate our main findings, namely, RQ2 and RQ3, using WikiSum. We present the results in Appendix B.6. We choose **Health** domain as it is the largest domain of this dataset. We observe that the trend in SVCCA scores across different scenarios on WikiSum is generally the same as those on Amazon Reviews, demonstrating that our findings are consistent.

## 5 Related Work

**Analyzing neural representations** Raghu et al. (2017) proposed SVCCA for comparing representations for the same data points from different layers and networks invariant to an affine transform. They also discovered that lower layers in a multi-layer neural network converge more quickly to their final representations in contrast to higher layers. Building off of SVCCA, Morcos et al. (2018) developed projection weighted CCA (PWCCA) using an aggregation technique. Using the SVCCA tool, Saphra and Lopez (2019) studied the learning dynamics of neural language models by probing the evolution of syntactic, semantic, and topic representations across time and models. Kudugunta et al. (2019) used SVCCA to understand massively multilingual neural machine translation representations over 100 languages. Their major findings are that encoder representations of different languages form clusters based on their linguistic similarities.



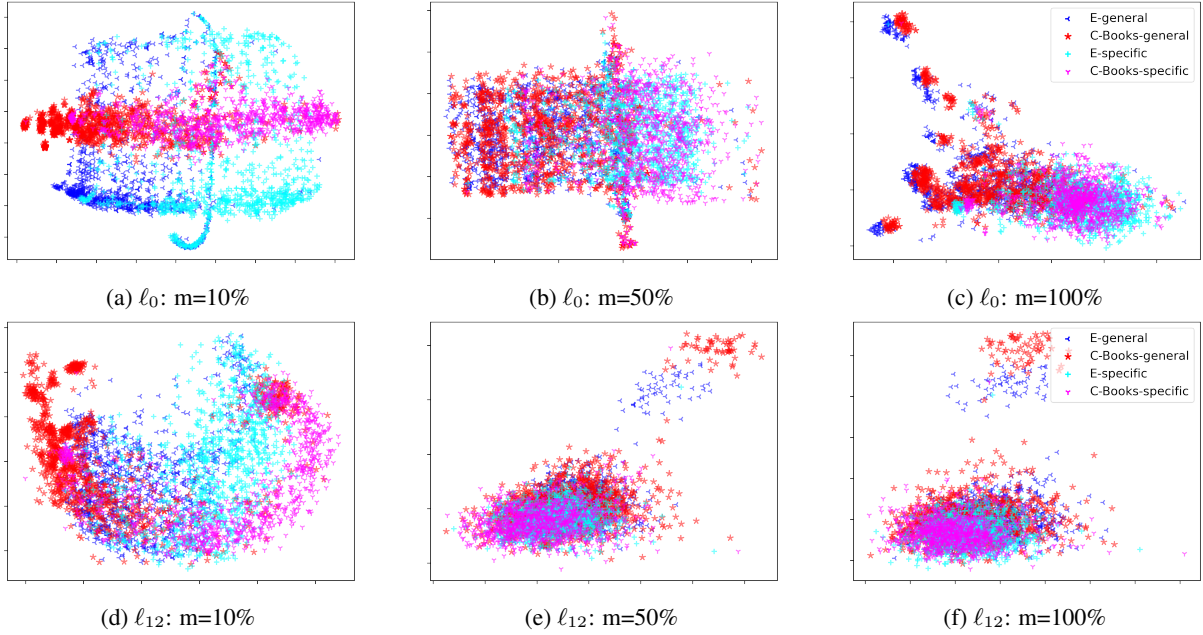


Figure 6: Visualization for  $\ell_0$  and  $\ell_{12}$  representations for **E** and **C<sub>Books</sub>**. We use colors (blue/cyan for **E** and red/magenta for **C<sub>Books</sub>**) to separate representations for generals and domain-specific words.  $m$  denotes model capacity. All models here use a data size of 100%.

**Diagnostic Classifiers** Another prominent tool for analyzing learned representations is diagnostic classifiers (DCs; Belinkov et al., 2017a,b; Giullianelli et al., 2018). DCs measure the amount of information encoded in representations about a particular task by using them as input to a classifier, which is trained on the task in a supervised manner. DC users assume that the higher their performance for this task, the more task-specific information is encoded in the representations. While widely adopted, DCs have several pitfalls. For example, Zhang and Bowman (2018) showed that learning a classifier on top of random embeddings is often competitive and, in some cases, even better than doing so with representations taken from a pre-trained model when trained on enough data. Saphra and Lopez (2019) demonstrated that, unlike SVCCA, DCs showed a stable correlation between language models and target labels throughout training epochs, in contrast to the language models’ immense improvement over time.

## 6 Conclusions and Future Work

We present a novel methodology based on subpopulation analysis which helps understand how subdomains are represented in a multi-domain model. Our findings show that neural models encode domain information differently in lower and upper layers and that larger models (in our case, **E**) tend

to “preserve a copy” of small, more specialized models (**C**). Generally, we observe rapid model improvements in NLP tasks when model capacity and dataset size, the two dimensions we study, increase. We encourage the research community to study the cause for these improvements from a multi-domain angle (i.e., the ability to encode specific information about many domains at once using the increased capacity). In future work, we would like to apply our methodology to examine the behavior of multilingual, multitask, and multi-modal models.

## Acknowledgments

This work was supported by the UKRI Centre for Doctoral Training (CDT) in Natural Language Processing through UKRI grant EP/S022481/1 and CDT funding from Huawei. We would like to thank Bonnie Webber, Ivan Titov and the anonymous reviewers for their helpful feedback. We appreciate the use of computing resources through the CSD3 cluster at the University of Cambridge.

## Limitations

Throughout this work, we use the BERT<sub>BASE</sub> model. While it is widely adopted in the NLP community, there are other more advanced models (such as BERT<sub>LARGE</sub>, RoBERTa and GPT3) that we do not experiment with due to a lack of

resources. Given that the differences between models of the BERT family are mostly irrelevant to the way we conduct our experiments, we believe our results would generalize, at the very least, to this family of models.

In addition, we do not experiment with a large amount of training data for two reasons: a) We want to control for the domains from which we draw examples, and those have a size limitation, and b) Training many models on a large dataset is computationally expensive. Our multi-domain setup is comprised of five domains. We believe a higher number of domains should be considered for real-world scenarios.

To control our experiments, we train all models from scratch. For real-world scenarios, it would be harder to divide the training data into homogeneous and natural domains. While our proposed methodology can be easily adapted to different similarity measurement methods, we focus on SVCCA, which restricts us to linear correlations. In future work, we plan to investigate the nature of domains using non-linear techniques.

We identify domains through a common topic, and as a result, the shared lexical choices within the domain. This is the most common case for classifying domains, but we acknowledge that there are additional valuable ways to define domains. For example, domains could be separated based on writing style while still having a significant shared vocabulary (Amazon book reviews and Wikipedia articles about books).

## References

- Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2019. [Identifying and controlling important neurons in neural machine translation](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017a. [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.
- Yonatan Belinkov, Lluís Màrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2017b. [Evaluating layers of representation in neural machine translation on part-of-speech and semantic tagging tasks](#). In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1–10, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- John Blitzer, Mark Dredze, and Fernando Pereira. 2007. [Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification](#). In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 440–447, Prague, Czech Republic. Association for Computational Linguistics.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA. Curran Associates Inc.
- Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. [ELECTRA: Pre-training text encoders as discriminators rather than generators](#). In *ICLR*.
- Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. [A discourse-aware attention model for abstractive summarization of long documents](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.
- Nachshon Cohen, Oren Kalinsky, Yftah Ziser, and Alessandro Moschitti. 2021. [WikiSum: Coherent summarization dataset for efficient human-evaluation](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 212–219, Online. Association for Computational Linguistics.
- Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. [What’s so special about BERT’s layers? a closer look at the NLP pipeline in monolingual and multilingual models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

- Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Halle R Dimsdale-Zucker and Charan Ranganath. 2018. Representational similarity analyses: a practical guide for functional mri applications. In *Handbook of behavioral neuroscience*, volume 28, pages 509–525. Elsevier.
- Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. [Adversarial and domain-aware BERT for cross-domain sentiment analysis](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.
- Marcio Fonseca, Yftah Ziser, and Shay B. Cohen. 2022. [Factorizing content and budget decisions in abstractive summarization of long documents by sampling summary views](#). In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Jonathan Frankle and Michael Carbin. 2019. [The lottery ticket hypothesis: Finding sparse, trainable neural networks](#). In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.
- Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. [Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.
- Max Grusky, Mor Naaman, and Yoav Artzi. 2018. [Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies](#). 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 708–719, New Orleans, Louisiana. 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).
- David R. Hardoon, Sándor Szedmak, and John Shawe-Taylor. 2004. [Canonical correlation analysis: An overview with application to learning methods](#). *Neural Comput.*, 16(12):2639–2664.
- Timothy J Hazen, Shehzaad Dhuliawala, and Daniel Boies. 2019. Towards domain adaptation from limited data for question answering using deep neural networks. *arXiv preprint arXiv:1911.02655*.
- Klaus Hinkelmann and Oscar Kempthorne. 2007. *Design and analysis of experiments, volume 1: Introduction to experimental design*, volume 1. John Wiley & Sons.
- Sneha Kudugunta, Ankur Bapna, Isaac Caswell, and Orhan Firat. 2019. [Investigating multilingual NMT representations at scale](#). 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 1565–1575, Hong Kong, China. Association for Computational Linguistics.
- Entony Lekhtman, Yftah Ziser, and Roi Reichart. 2021. [DILBERT: Customized pre-training for domain adaptation with category shift, with an application to aspect extraction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 219–230, Online and Punta Cana, Dominican Republic. 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. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Quanyu Long, Tianze Luo, Wenya Wang, and Sinno Pan. 2022. [Domain confused contrastive learning for unsupervised domain adaptation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2982–2995, Seattle, United States. Association for Computational Linguistics.
- Inbal Magar and Roy Schwartz. 2022. [Data contamination: From memorization to exploitation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165, Dublin, Ireland. Association for Computational Linguistics.
- Ari S. Morcos, Maithra Raghu, and Samy Bengio. 2018. [Insights on representational similarity in neural networks with canonical correlation](#). In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5732–5741.
- Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. [Justifying recommendations using distantly-labeled reviews and fine-grained aspects](#). 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 188–197, Hong Kong, China. Association for Computational Linguistics.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor



- Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [Pytorch: An imperative style, high-performance deep learning library](#). In *NeurIPS*.
- Barbara Plank. 2016. [What to do about non-standard \(or non-canonical\) language in NLP](#). In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, Bochum, Germany, September 19-21, 2016*, volume 16 of *Bochumer Linguistische Arbeitsberichte*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *JMLR*.
- Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. [SVCCA: singular vector canonical correlation analysis for deep learning dynamics and interpretability](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6076–6085.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Tim Rocktäschel, Torsten Huber, Michael Weidlich, and Ulf Leser. 2013. [WBI-NER: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs](#). In *Second Joint Conference on Lexical and Computational Semantics (\*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 356–363, Atlanta, Georgia, USA. Association for Computational Linguistics.
- 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.
- Siamak Shakeri, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Feng Nan, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. [End-to-end synthetic data generation for domain adaptation of question answering systems](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5445–5460, Online. Association for Computational Linguistics.
- Jingbo Shang, Liyuan Liu, Xiaotao Gu, Xiang Ren, Teng Ren, and Jiawei Han. 2018. [Learning named entity tagger using domain-specific dictionary](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2054–2064, Brussels, Belgium. Association for Computational Linguistics.
- Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. [BERT is not an interlingua and the bias of tokenization](#). In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. [Probing pretrained language models for lexical semantics](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275.
- 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.
- Jianling Wang, Kaize Ding, Liangjie Hong, Huan Liu, and James Caverlee. 2020. Next-item recommendation with sequential hypergraphs. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pages 1101–1110.
- 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 *Proc. of EMNLP*.

Zhenrui Yue, Bernhard Kratzwald, and Stefan Feuerriegel. 2021. [Contrastive domain adaptation for question answering using limited text corpora](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9575–9593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. [Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Shaozhong Zhang, Dingkai Zhang, Haidong Zhong, and Guorong Wang. 2020. A multiclassification model of sentiment for e-commerce reviews. *IEEE Access*, 8:189513–189526.

Yftah Ziser and Roi Reichart. 2017. [Neural structural correspondence learning for domain adaptation](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 400–410, Vancouver, Canada. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018. [Pivot based language modeling for improved neural domain adaptation](#). 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 1241–1251, New Orleans, Louisiana. Association for Computational Linguistics.

## A Additional Details for Experiments

Here we provide some additional details for our experiments.

**Training** We set the validation data size for **E** to be 10K, which is composed of 2K reviews from each domain. For validation set of each  $C_i$ , we use the same 2K reviews used for **E** from each domain. For consistency, we use the same validation set for all data sizes. We use a test set with 2.5K reviews for each domain. The same test set is fed to both **E** and  $C_i$  across all model capacities and data sizes to obtain representations for subpopulation analysis. When it is clear from the context which  $C_i$  for  $i \in [5]$  we are referring to (and under which training regime), we will use the simplification **C**.

All models use the validation set cross-entropy loss to perform early stopping, and we train a model for a maximum of 500 epochs. We provide the validation loss (cross-entropy) for the **E** model in Table 3. From the results, we can see that for fixed data size, model performance saturates when reaching model capacity of 100%. Thus, unlike data size, we do not perform further experiments with model capacity larger than 100%.

	10%d	50%d	100%d	200%d
10%m	6.052	5.541	4.788	3.886
25%m	5.764	3.257	2.745	2.354
50%m	4.366	2.758	2.451	2.144
75%m	4.017	2.781	2.435	2.149
100%m	4.012	2.786	2.436	2.16

Table 3: Validation cross-entropy loss on the experimental model for different model capacities and data sizes where m refers to model capacity and d refers to data size used to train the model.

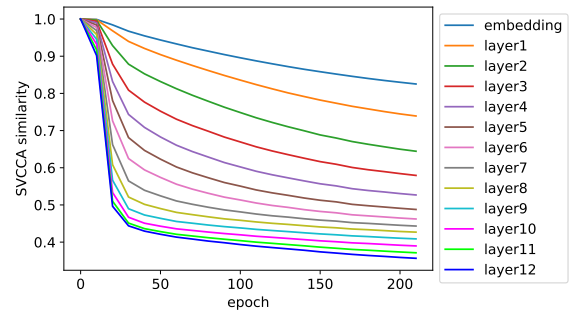


Figure 7: Training dynamics for all layers between **E** and  $C_{Clothing}$ . Here both model and data size are 100%.

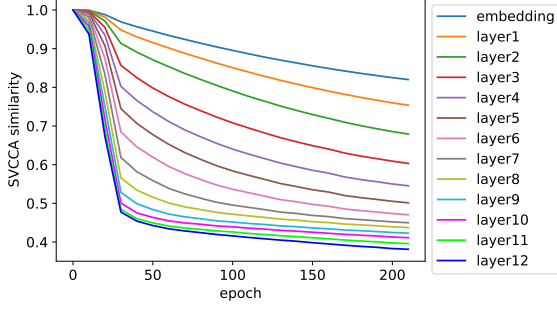


Figure 8: Training dynamics for all layers between  $\mathbf{E}$  and  $\mathbf{C}_{Electronics}$ . Here both model and data size are 100%.

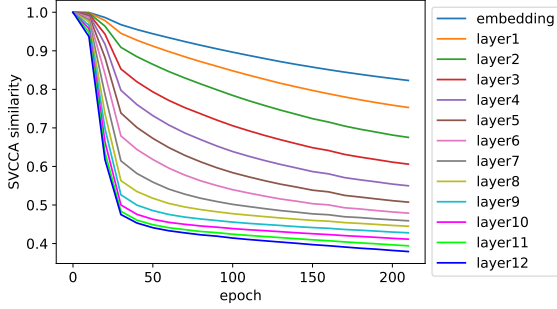


Figure 9: Training dynamics for all layers between  $\mathbf{E}$  and  $\mathbf{C}_{Home}$ . Here both model and data size are 100%.

All models are trained on 4 NVIDIA A100 GPUs with a batch size of 32 per GPU. We use PyTorch (Paszke et al., 2019) and the HuggingFace library (Wolf et al., 2020) for all model implementation.

## B Additional Details for Results

### B.1 Additional Results for RQ1

We provide additional experimental results for training dynamics on Clothing Shoes and Jewelry (Figure 7), Electronics (Figure 8), Home and Kitchen (Figure 9), and Movies and TV (Figure 10).

### B.2 Additional Results for RQ2

In § 4, we provided SVCCA results between  $\mathbf{E}$  and different  $\mathbf{C}_i$ s for three domains. Here we present the results for the rest of the two domains in Figure 13a, 13d, 14a, and 14d.

### B.3 Example of General and Domain-specific Words

We provide a sample of general words and domain specific words for each domain in Table 4. Note that list of general words are domain independent,

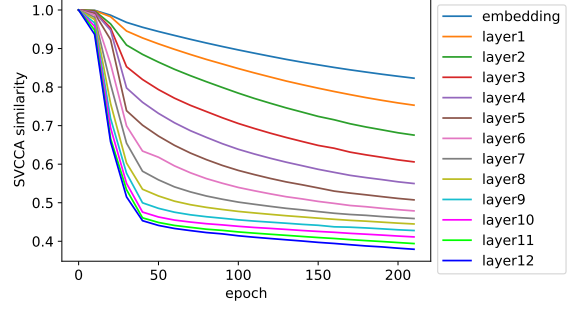


Figure 10: Training dynamics for all layers between  $\mathbf{E}$  and  $\mathbf{C}_{Movies}$ . Here both model and data size are 100%.

i.e., the general word list is the same for all domains.

### B.4 Additional Results for RQ3

Here we present additional results for SVCCA score between  $\mathbf{E}$  and  $\mathbf{C}_i$  for different subsets of tokens. Figure 11 illustrates for  $\mathbf{C}_{Clothing}$ , Figure 12 illustrates for  $\mathbf{C}_{Electronics}$ , Figure 13 illustrates for  $\mathbf{C}_{Home}$ , and Figure 14 illustrates for  $\mathbf{C}_{Movies}$ .

### B.5 Additional Results for RQ4

Here we provide more example MLM predictions of  $\mathbf{E}$  and  $\mathbf{C}_i$ . Table 5 presents predictions using k-nearest neighbors of the word embeddings. Table 6 presents predictions using the final layer representation.

### B.6 Additional Results on WikiSum

Here we provide additional results on WikiSum Health domain in Figure 15, including SVCCA results between  $\mathbf{E}$  and  $\mathbf{C}_{Health}$ , as well as results for different subsets of tokens.

<b>General words:</b> totally, preference, cost, mistake, hello, noticeable, play, factor, common, friend, previously, upon, explain, future, everyone
<b>Books:</b> gutenber, appendix, autobiographical, grammatically, bookshelves, democrat, asides, arabic, stagnant, curriculum, minutiae, gripped, publishers, referencing, socialism
<b>Clothing:</b> marten, docker, florsheim, rockports, skechers, buckles, 38d, fleece, nylons, insoles, tees, pantyhose, puckered, slippers, footwear
<b>Electronics:</b> printable, wifi, 105mm, aux, energizer, recordable, directories, reinstall, gigabit, reboots, portability, vga, hitachi, configurations, wirelessly
<b>Home:</b> cupcakes, kitchenaid, undercooked, ikea, chopper, mugs, steamers, juices, fiesta, kettles, aroma, toasted, rinsed ovens, airtight
<b>Movie:</b> scenic, 16x9, nightclub, cheesiest, filmmakers, supernova, serials, weepy, purists, incarnations, lionsgate, reportedly, suggestive, 1931, choreography

Table 4: A representative sample of general words (top row) and domain specific words (bottom rows) taken from different categories (domains) of the dataset.

m=50%		m=100%	
E	$C_i$	E	$C_i$
editors	volumns	editors	editors
publisher	buyer	publisher	publisher
heirs	listing	editor	editor
libraries	edit	writers	authors
universities	hardcover	authors	reviewers

(a) 5-nearest neighbors for the domain-specific word **publishers** with  $i=Books$ .

m=50%		m=100%	
E	$C_i$	E	$C_i$
comics	jokes	comics	comics
jokes	joke	comedian	joke
comedian	accolades	laughs	comedian
directors	critics	comedies	critics
commentators	reviewers	jokes	laughs

(c) 5-nearest neighbors for the domain-specific word **comedians** with  $i=Movies$  and  $TV$ .

m=50%		m=100%	
E	$C_i$	E	$C_i$
towards	towards	towards	towards
beside	settled	against	at
surrounding	at	onto	onto
beneath	concerning	at	against
against	behind	beside	near

(b) 5-nearest neighbors for the general word **toward** with  $i=Books$ .

m=50%		m=100%	
E	$C_i$	E	$C_i$
print	vinyl	plastic	plastic
plastic	bonded	print	vinyl
cloth	plastic	materials	cardboard
cardboard	junk	paperback	print
printed	cardboard	cardboard	tissue

(d) 5-nearest neighbors for the general word **paper** with  $i=Clothing$  Shoes and Jewelry.

Table 5: Example predictions of  $E$  and  $C_i$  using 5-nearest neighbors from embedding layer weights.  $m$  denotes model capacity. All models here use data size of 100%.

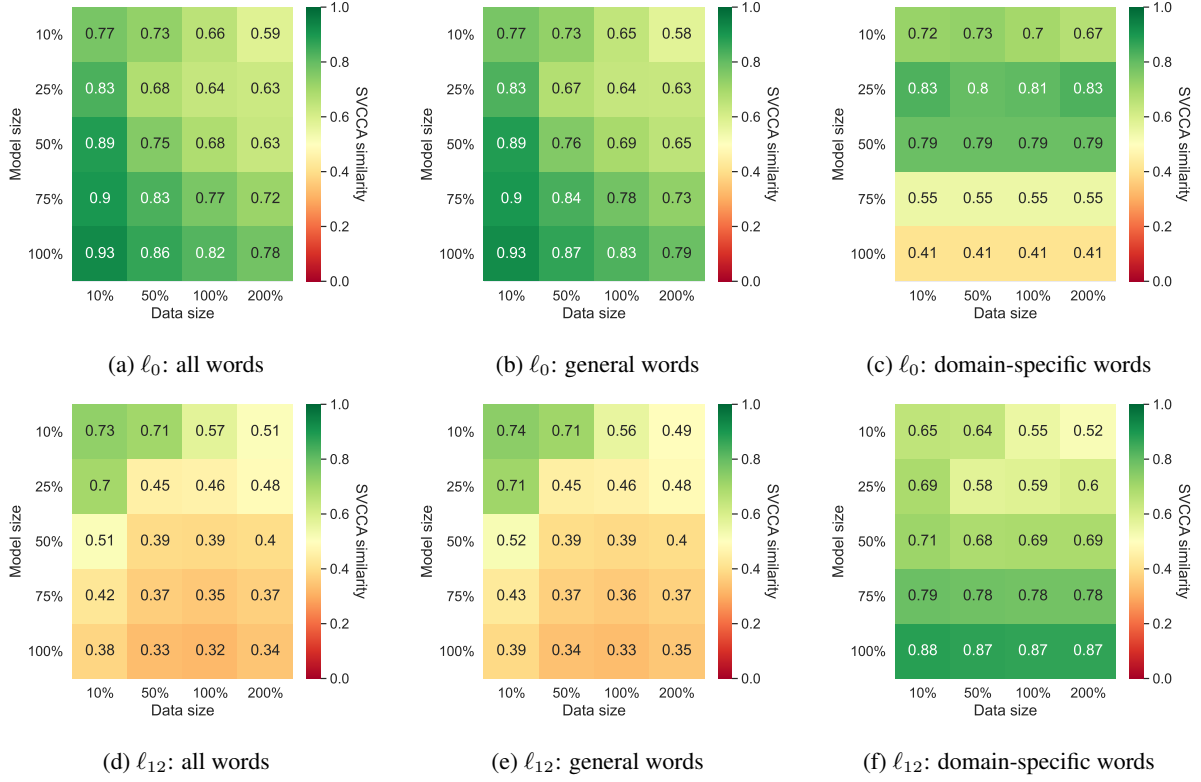


Figure 11: The SVCCA score between  $\mathbf{E}$  and  $\mathbf{C}_{Clothing}$  for different subsets of tokens. The top row presents the results for the embedding layer  $\ell_0$ , and the bottom row presents them for the last layer  $\ell_{12}$ .

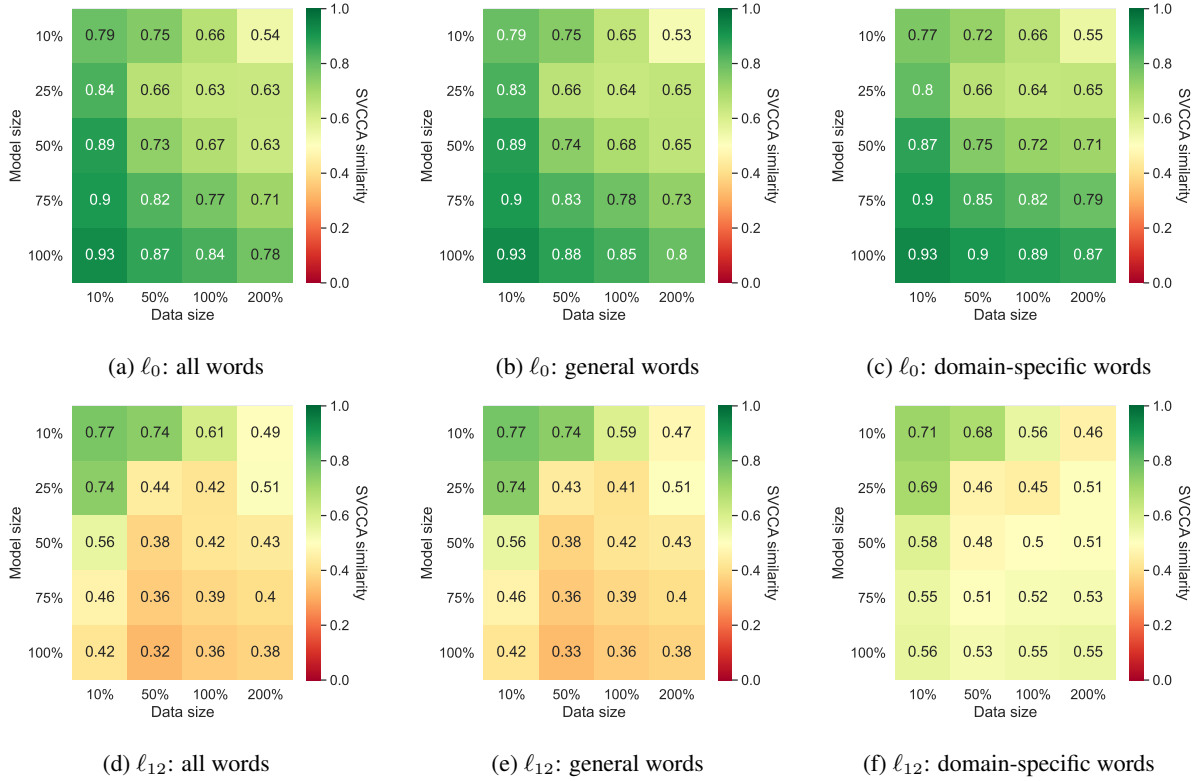


Figure 12: The SVCCA score between  $\mathbf{E}$  and  $\mathbf{C}_{Electronics}$  for different subsets of tokens. The top row presents the results for the embedding layer  $\ell_0$ , and the bottom row presents them for the last layer  $\ell_{12}$ .



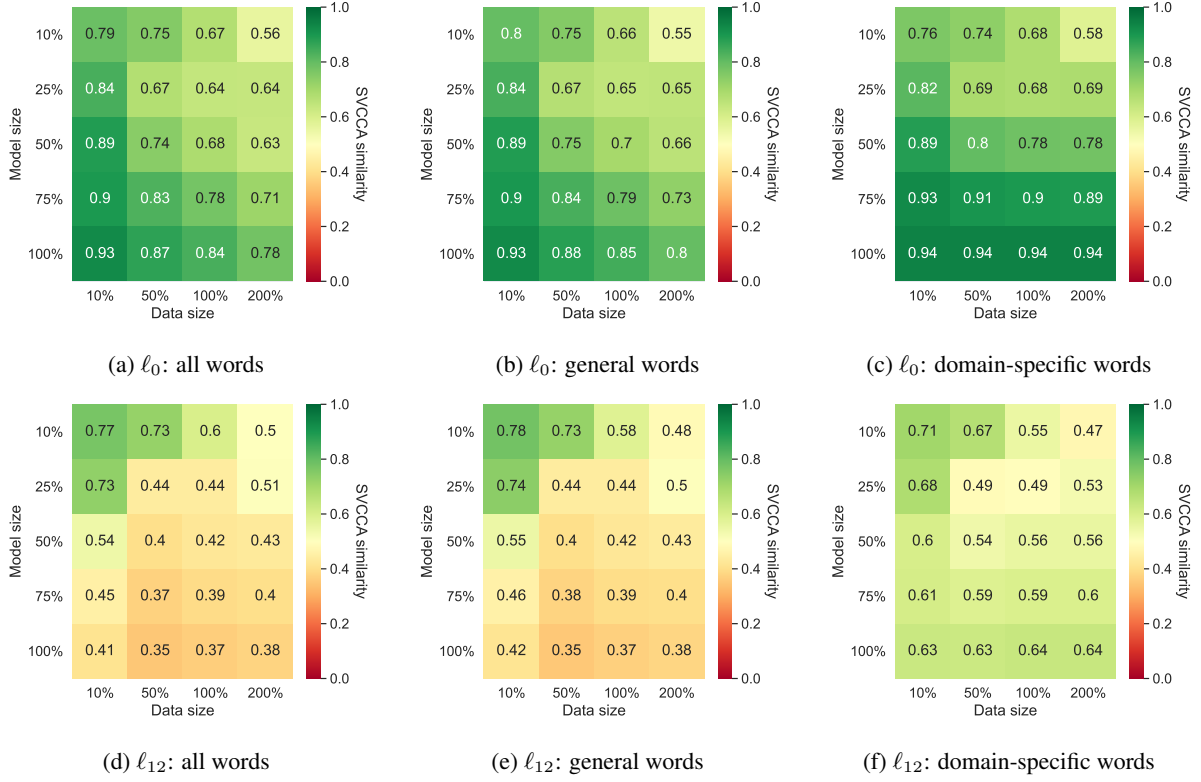


Figure 13: The SVCCA score between  $\mathbf{E}$  and  $\mathbf{C}_{Home}$  for different subsets of tokens. The top row presents the results for the embedding layer  $\ell_0$ , and the bottom row presents them for the last layer  $\ell_{12}$ .

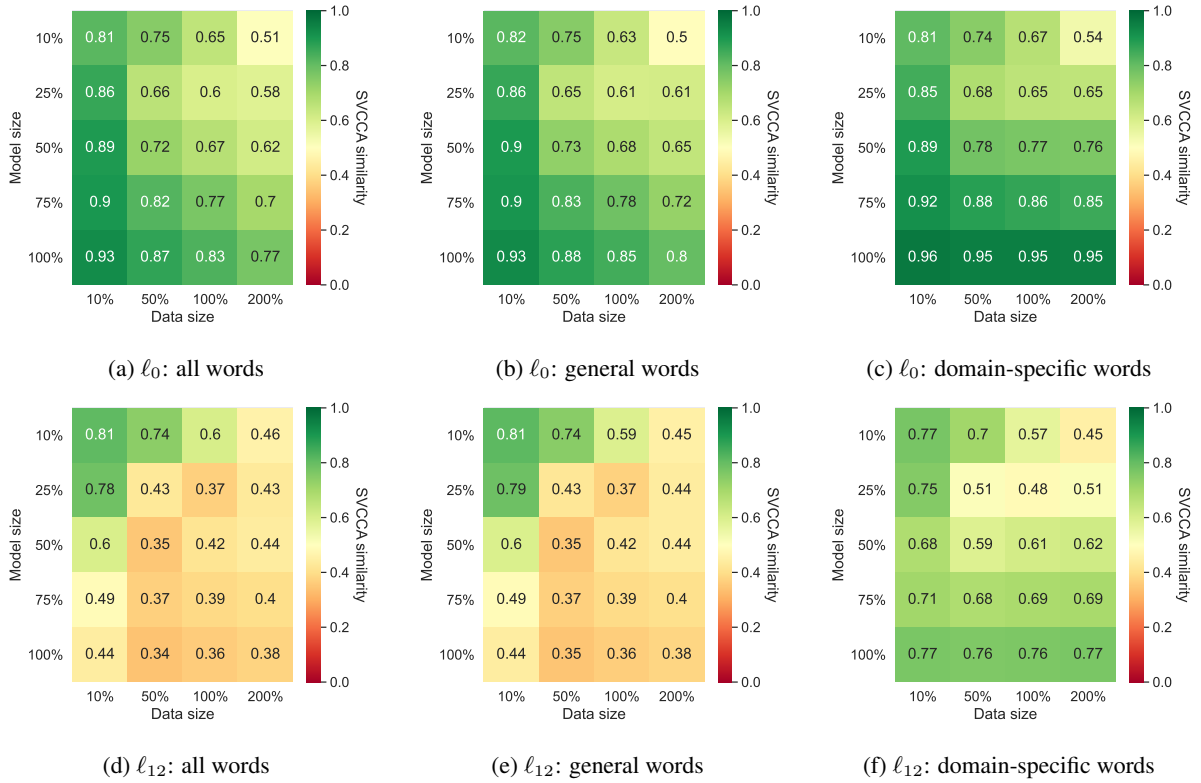


Figure 14: The SVCCA score between  $\mathbf{E}$  and  $\mathbf{C}_{Movies}$  for different subsets of tokens. The top row presents the results for the embedding layer  $\ell_0$ , and the bottom row presents them for the last layer  $\ell_{12}$ .

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
food	counter	bottle	counter
counter	hands	refrigerator	bottle
wine	oil	wine	hands
oil	food	food	sink
salad	salad	fridge	stove

(a) *I realize the point of my purchase was to reduce the amount of olive oil I sprayed on my [MASK] but I do end up having to pump it up and mist twice.* The masked word is a domain-specific word **salad** with  $i=Home$  and Kitchen.

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
say	have	worry	worry
think	say	complain	say
complain	know	wonder	know
know	care	know	think
worry	understand	say	complain

(c) *Amazon replaced it with no hassle, but I always have to [MASK] about these drives.* The masked word is a general word **worry** with  $i=Electronics$ .

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
guy	guy	girl	guy
musician	woman	guy	woman
dude	man	killer	hero
kid	kid	gal	cop
vampire	person	dude	man

(b) *There had to be the four friends-a hypochondriac, a smoothing-talking [MASK] who gets everyone in trouble, the joker's friend who's a bit of a ham but has slightly more brains, and a girl.* The masked word is a domain-specific word **joker** with  $i=Movies$  and TV.

m=50%		m=100%	
E	C <sub>i</sub>	E	C <sub>i</sub>
instructed	expected	suggested	suggested
suggested	instructed	stated	instructed
well	stated	instructed	expected
usual	advertised	advertised	well
indicated	normal	well	stated

(d) *I ordered a half size down as [MASK] and the size 11 eclipses my foot.* The masked word is a general word **suggested** with  $i=Clothing$  Shoes and Jewelry.

Table 6: Example MLM predictions of **E** and **C<sub>i</sub>** using last layer representation.  $m$  denotes model capacity. All models here use a data size of 100%.

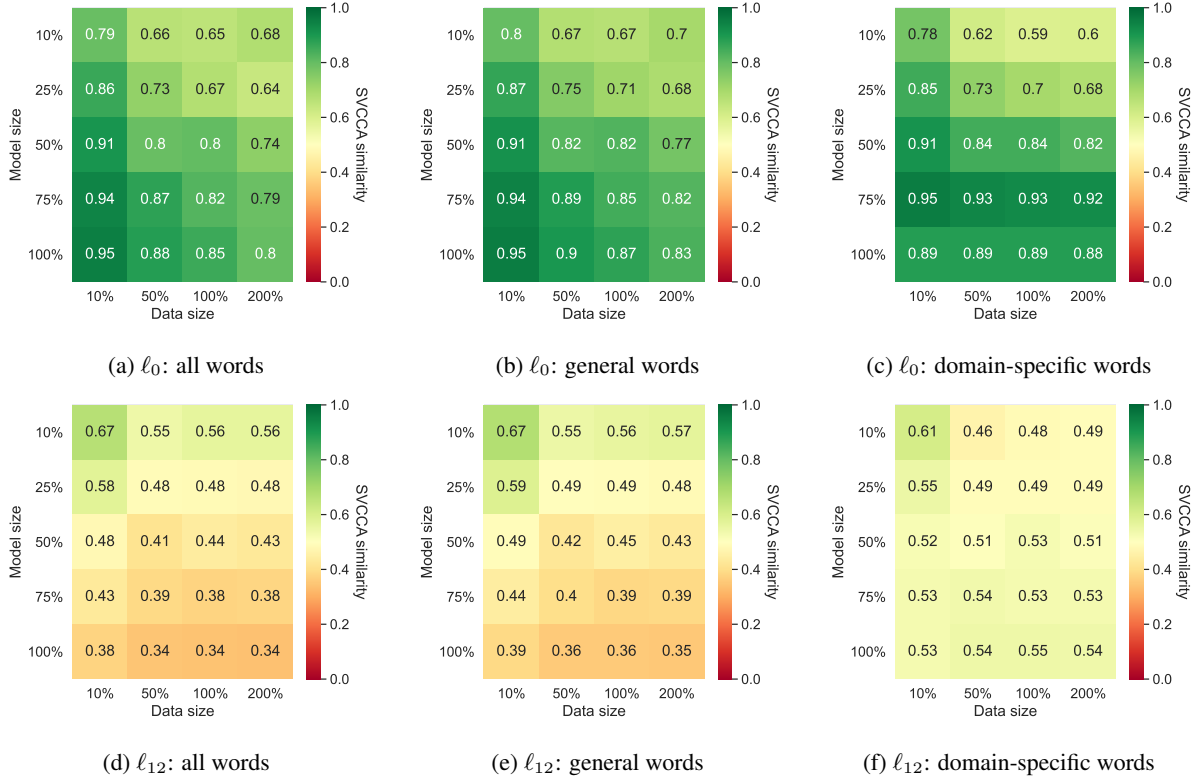


Figure 15: The SVCCA score between **E** and **C<sub>Health</sub>** for different subsets of tokens. The top row presents the results for the embedding layer  $\ell_0$ , and the bottom row presents them for the last layer  $\ell_{12}$ .