# Less Mature is More Adaptable for Sentence-level Language Modeling

**Abhilasha Sancheti**[1*] **David Dale**[2] **Artyom Kozhevnikov**[2] **Maha Elbayad**[2]
[1]University of Maryland, College Park [2]FAIR at Meta
sancheti@umd.edu, {daviddale, artyomko, elbayadm}@meta.com

## Abstract

This work investigates sentence-level models (*i.e.,* models that operate at the sentence-level) to study how sentence representations from various encoders influence downstream task performance, and which syntactic, semantic, and discourse-level properties are essential for strong performance. Our experiments encompass encoders with diverse training regimes and pretraining domains, as well as various pooling strategies applied to multi-sentence input tasks (including sentence ordering, sentiment classification, and natural language inference) requiring coarse-to-fine-grained reasoning. We find that "less mature" representations (*e.g.,* mean-pooled representations from BERT's first or last layer, or representations from encoders with limited fine-tuning) exhibit greater generalizability and adaptability to downstream tasks compared to representations from extensively fine-tuned models (*e.g.,*, SBERT or SimCSE). These findings are consistent across different pretraining seed initializations for BERT. Our probing analysis reveals that syntactic and discourse-level properties are stronger indicators of downstream performance than MTEB scores or decodability. Furthermore, the data and time efficiency of sentence-level models, often outperforming token-level models, underscores their potential for future research.

## 1 Introduction

Sentence representation learning is an extensively researched area. Existing works either fine-tune encoders (such as BERT (Devlin et al., 2018)) using Siamese networks (Conneau et al., 2017; Reimers and Gurevych, 2019) or contrastive learning approaches (Gao et al., 2021) using Natural Language Inference (NLI) datasets (Bowman et al., 2015; Williams et al., 2018) or further pretraining encoders using conditional language modeling (Yang et al., 2021) to incorporate discourse information in the representations. Others use encoder-decoder models for generating (Kiros et al., 2015) or predicting surrounding sentences (Logeswaran and Lee, 2018) to learn monolingual sentence representations or train encoder-decoder models on machine translation datasets (Conneau, 2019; Artetxe and Schwenk, 2019; Duquenne et al., 2023) to learn multilingual representations. Representations thus obtained are evaluated based on their performance on downstream tasks (Muennighoff et al., 2023) such as clustering, paraphrasing, classification, and retrieval. They are also evaluated on probing tasks to assess various surface-level, syntactic, semantic, and discourse properties encoded in them (Conneau et al., 2018; Conneau and Kiela, 2018; Chen et al., 2019).

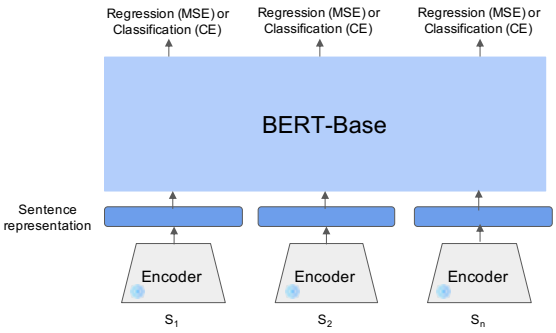



Figure 1: **Sentence-level Model:** Sentence representations from a frozen encoder are used as input to a randomly initialized BERT-like transformer model. This model is trained to predict the position (regression for sentence ordering) or category (for sequential sentence classification) for each sentence. MSE denotes mean-squared loss and CE denotes cross-entropy loss. For NLI task, we prepend a CLS token to the input and use the representation corresponding to this CLS token to predict the output class for a (premise, hypothesis) pair.

Apart from learning sentence representations, another line of research has used these sentence representations for causal language modeling or masked

[*]Work done during internship at Meta FAIR.



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11680–11695

| Model | Pretraining Dataset | Pooling Strategies | Approach/Training objective |
|---|---|---|---|
| BERT (Devlin et al., 2018) | Wikipedia, Books. | Last layer: mean, CLS, CLS before pooler, max. First layer: mean, CLS before pooler, max. | Language modeling + Next sentence prediction. |
| SciBERT (Beltagy et al., 2019) | Science and Biomedical papers. | Last layer: mean, CLS, CLS before pooler, max. First layer: mean, CLS before pooler, max | Language modeling + Next sentence prediction (Uses BERT architecture as the base model). |
| BERT-SimCSE (Gao et al., 2021) | SNLI + MNLI. | Last layer: mean, CLS, CLS before pooler, max. First layer: mean, CLS before pooler, max | Fine-tunes BERT on NLI datasets using contrastive learning. |
| SBERT (Reimers and Gurevych, 2019) | NLI + Paraphrasing + Question Answering. | Mean | Fine-tunes BERT-based Siamese network using contrastive objective. |
| SBERT-NLI (Reimers and Gurevych, 2019) | SNLI + MNLI. | Mean | Fine-tunes BERT-based Siamese network using 3-class classification loss. |
| SciBERT-NLI[1] | SNLI + MNLI. | Mean | Fine-tunes SciBERT-based Siamese network using 3-class classification loss. |
| CMLM (Yang et al., 2021) | Common Crawl. | Mean | Further pre-train BERT using Conditional Language Modeling. |

Table 1: BERT-based encoders: Pretraining datasets, pooling strategies (first and last layers), and training objectives. All models use 768-dimensional representations. See respective papers for details

language modeling to build sentence-level models resulting in improved storyline generation (Ippolito et al., 2020) and document embeddings (Czinczoll et al., 2024). Such sentence-level models represent a sequence of sentences by a sequence of sentence representations (one per sentence) as opposed to standard token-level models which take in a sequence of token representations. There also exists works that use sentence-level modeling for the task of sentence ordering (Cui et al., 2018; Basu Roy Chowdhury et al., 2021; Kumar et al., 2020; Golestani et al., 2021; Bin et al., 2023), sentence infilling (Huang et al., 2020; Mori et al., 2020), or sequential sentence classification (Cohan et al., 2019; Hillebrand et al., 2024).

However, limited research has been done in assessing: (RQ1) how does a representation learning approach (such as fine-tuning, contrastive learning, conditional language modeling, etc.) impact the performance of a downstream sentence-level modeling task? (RQ2) what properties must be encoded in sentence representations to enable sentence-level modeling tasks? And (RQ3) what are the advantages of sentence-level models over standard token-level models?

We experiment with several sentence representations, obtained from a variety of sentence encoders, to build sentence-level models for addressing (RQ1). We use existing sentence representation evaluation benchmarks (Conneau et al., 2018; Conneau and Kiela, 2018; Muennighoff et al., 2023; Chen et al., 2019) to assess the surface-level, syntactic, semantic, and discourse-level properties encoded in embeddings. We correlate these properties with their downstream task performance to answer (RQ2) in Section 4. We then compare the downstream task performance of a token-level model with that of sentence-level model in Section 4 to investigate (RQ3).

Additionally, we thoroughly study the impact of using different pooling strategies on downstream performance, investigate the robustness of our find-

| Task | Description | Domain | Datasets |
|---|---|---|---|
| **SO**<br>Sentence Ordering<br>(coarse-grained) | Input: shuffled sequence of sentences<br>Output: relative position of each sentence<br>Task Formulation: Sentence-level regression | General | **ROCStories** (Mostafazadeh et al., 2016) – human written five sentence commonsense stories capturing causal and temporal relations between everyday life events.<br>**SIND** (Huang et al., 2016) – humans write five sentence stories for a sequence of photos from Flickr album. |
| | | Scientific | **NIPS** (Logeswaran et al., 2016)- abstracts from NIPS papers from 2005-2015.<br>**AAN** (Radev et al., 2016) - abstracts from ACL anthology until 2013. |
| **SSC**<br>Sequential Sentence Classification<br>(coarse-grained) | Input: sequence of sentences in a paragraph<br>Output: role of each sentence<br>Task Formulation: Sentence-level classification | Scientific | **CSAbstruct** (Cohan et al., 2019) – manually annotated sentences from computer science abstracts (obtained from semantic scholar) for background, method, result, objective, and other. |
| **NLI**<br>Paragraph-level Natural Language Inference<br>(fine-grained) | Input: Multi-sentence premise and a hypothesis<br>Output: Whether premise entails, contradicts, or is neutral with respect to the hypothesis<br>Task Formulation: Paragraph-level classification | General | **ANLI** (Nie et al., 2020) – adversarial collected NLI dataset via human-and-machine in the loop. Premise (multi-sentence passages from Wikipedia) and target label (one of entailment, contradiction or neutral) are provided to humans to write a hypothesis. |

Table 2: Three multi-sentence input tasks, their description, and datasets used in our study. We cover tasks requiring coarse-to-fine-grained reasoning spanning two domains.

| Task | Dataset | Max/Mean | Dataset split | | |
|---|---|---|---|---|---|
| | | | Train | Dev | Test |
| SO | NIPS | 15/6 | 2448 | 409 | 402 |
| | AAN | 12/5 | 8569 | 962 | 2626 |
| | SIND | 5/5 | 40155 | 4990 | 5055 |
| | RocStories | 5/5 | 78529 | 9816 | 9817 |
| SSC | CSAbstruct | 10/7 | 1668 | 295 | 226 |
| NLI | ANLI | 18/4 | 162865 | 3200 | 3200 |

Table 3: Sentence-level dataset statistics. Max/Mean are computed over the number of input sentences.

ings across multiple pretraining seed initializations of encoders, and assess if the ability to decode a sentence from its representation is indicative of the downstream task performance of the sentence-level model in Section 4.

## 2 What is a sentence-level model?

A sentence-level model (Figure 1) takes in sentence representations as input as opposed to tokens that are used in standard token-level models (such as BERT). We consider an encoder-only Transformer architecture for the sentence-level model. The sentence-level model takes in a sequence of sentence representations from an encoder that is kept frozen during sentence-level model training. Note that for the task of sentence ordering, we remove positional embeddings since the input is a shuffled sequence of sentences. We provide details on the encoders, tasks, and datasets used in the following sections.

### 2.1 Sentence Encoders

We experiment with sentence representations obtained from several BERT-based monolingual encoders (vanilla BERT, BERT-SimCSE, SBERT, etc.) spanning different training regimes, pretraining dataset domain, and pooling strategies. We chose BERT-based encoders to facilitate controlled experiments and minimize confounding factors that can impact our findings. For a detailed list of encoders, see Table 1.

### 2.2 Downstream Tasks

We consider several multi-sentence input tasks requiring coarse-to-fine-grained reasoning from two domains, scientific and general. We select a variety of tasks to study the generalizability of the findings. We provide details on the tasks, datasets, and task formulation in Table 2.

### 2.3 Experimental Details

We provide details on the training objectives for each task, and evaluation measures below. Dataset

statistics for each task are shown in Table 3.

**Training objectives** For the sentence ordering task, we utilize the final layer representations of each sentence to predict its relative position (as per the original sequence) within a shuffled sequence. Specifically, for a given shuffled sequence of sentences ([1,3,2,5,4]), we aim to predict the relative position of each sentence in the shuffled sequence (*i.e.*, [0.2=1/5, 0.6=3/5, 0.4=2/5, 1.0=5/5, 0.8=4/5]). This sentence-level model is trained using mean-squared error (MSE) loss. In the sequence sentence classification task, we again use the final layer representations to predict the category of each sentence, training with cross-entropy loss. For NLI task, we employ the final layer representation of the [CLS] token (prepended to the input) to predict the relationship (entailment, contradiction, or neutrality) between the hypothesis and premise. This NLI model is also trained using cross-entropy loss.

**Evaluation** For each task, we report the following metrics averaged over 4 runs:
*Sentence Ordering:* Accuracy of correctly predicting a sentence's position in the original sequence.
*Sequence Classification:* Accuracy of correctly predicting the category of each sentence in a sequence.
*NLI:* Accuracy of predicting the entailment label.

To provide context for these results, we also include scores from a random baseline for each task as a lower bound performance. For the sentence ordering task, we additionally report the current state-of-the-art score (Robin et al., 2023) to establish an upper bound for performance. Please refer to Appendix A for implementation and datasets related details.

## 3 Main Findings

### 3.1 Less supervised training signals for better generalization - How BERT's mean pooling rivals specialized sentence representations.

While specialized sentence representation encoders are known to top the MTEB benchmark (Muennighoff et al., 2023) (which is a widely used benchmark for evaluating text representations on 8 types of tasks), Figure 2 demonstrates that mean-pooled BERT representations achieve performance comparable to CMLM and surpass specialized sentence representations like SBERT and BERT-SimCSE.

While SBERT representations result in the lowest downstream task accuracy for the scientific domain, their performance is relatively closer to other encoders (BERT, BERT-SimCSE, and CMLM) in the general domain. This difference may be attributed to SBERT's fine-tuning on a variety of general domain datasets. Although BERT-SimCSE is also trained on the same NLI datasets as SBERT-NLI, representations from BERT-SimCSE are more robust to the domain of downstream datasets as indicated by higher accuracy for the sequence classification task. This suggests that training using contrastive learning (BERT-SimCSE) is better than fine-tuning (SBERT-NLI) using cross-entropy loss.

While both BERT-SimCSE and SBERT use contrastive learning and SBERT has been fine-tuned on significantly larger amounts of data than BERT-SimCSE, representations from SBERT result in lower performance on scientific domain datasets than those from BERT-SimCSE. This shows that representations from less fine-tuned encoders are more generalizable and adaptable.

CMLM's best performance across all the tasks (except for sentence ordering in the scientific domain) indicates that continued pretraining of BERT using conditional language modeling (semi-supervised) results in more generalized representations as compared to contrastive learning (BERT-SimCSE, SBERT) or fine-tuning (SBERT-NLI).

As expected, representations from SciBERT outperform those from BERT for scientific domain datasets and perform lower than BERT for the general domain. Similar trends are observed between SBERT-NLI and SciBERT-NLI which is a scientific domain variant of SBERT-NLI. SciBERT-NLI which is a specifically fine-tuned (from SciBERT) Scientific analogue of SBERT-NLI also performs lower than SciBERT across all the tasks however it performs better on MTEB tasks (see Figure 10). Further analysis in Section 4 suggests that better performance of an encoder over another is due to having more syntactic and discourse-level properties encoded in their representations. All of these findings suggest that less mature representations are more adaptable to learning downstream task-specific properties than those that have already been tuned to strongly encode the semantics.

### 3.2 CLS token representations right before the pooling layer are surprisingly better.

We study the impact of using different pooling strategies to obtain the representations from en-

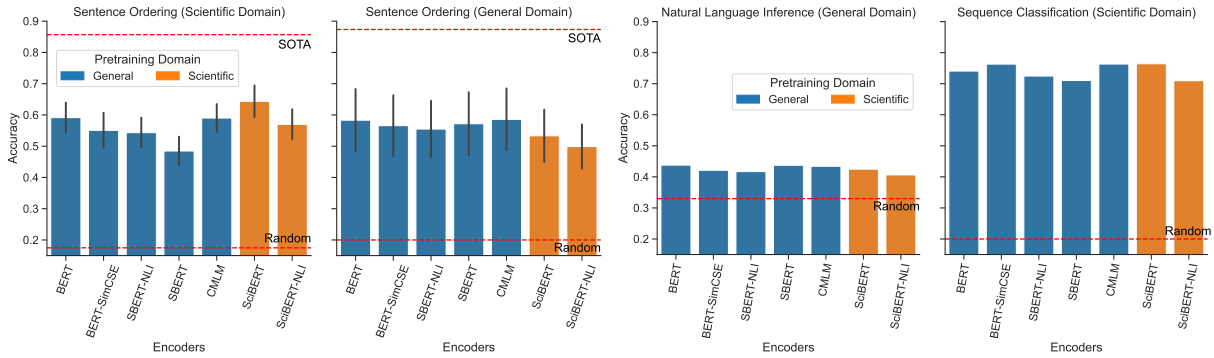



Figure 2: Sentence-level model accuracy using representations from encoders with varying training regimes and pretraining domains. Mean pooling is used for all encoders except BERT-SimCSE (CLS pooling). Variance shown is across sentence ordering datasets.

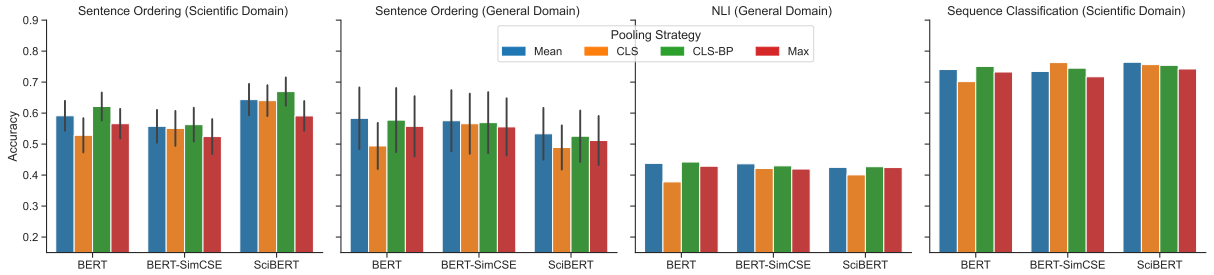



Figure 3: Accuracy for sentence-level models trained using representations from BERT, BERT-SimCSE, and SciBERT obtained using different pooling strategies from the final transformer layer.

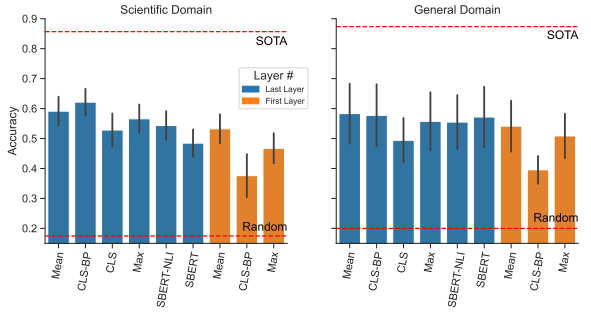



Figure 4: Sentence-level model accuracy: Comparing first/last layer pooling for BERT against SBERT and SBERT-NLI. Variance across sentence ordering datasets is shown.

coders on downstream sentence-level modeling task performance. We experiment with BERT, BERT-SimCSE, and SciBERT encoders, using representations pooled from the final layer. Specifically, we compare representations derived from the CLS token before the pooling layer (CLS-BP) with those derived after the pooling layer (CLS)[2].

Figure 3 shows that using CLS-BP yields surprisingly better downstream task accuracies than using CLS. However, the performance difference between CLS and CLS-BP is smaller for BERT-SimCSE than for other models, possibly because BERT-SimCSE uses the CLS representation during fine-tuning with contrastive learning.

CLS-BP also results in improved downstream task performance over mean pooled representations in the scientific domain. However, in the general domain, the performance is comparable. These observations align with our previous finding that less mature representations have more capacity to adapt and learn properties required for the downstream task. Further analysis in Section 4 reveals that CLS-BP encodes higher syntactic, semantic, and discourse-level properties than CLS, which leads to its improved downstream task performance.

### 3.3 First-layer representations can be more effective than representations from specifically trained encoders.

Given that less mature representations are generally more adaptable and that representations from the last layer are more task-specific than those from the first layer (Rogers et al., 2020), we investigated whether similar results are observed when pooling representations from the first layer. Focusing on

[2]The standard CLS representation from a BERT encoder uses an additional pooling layer during pretraining.

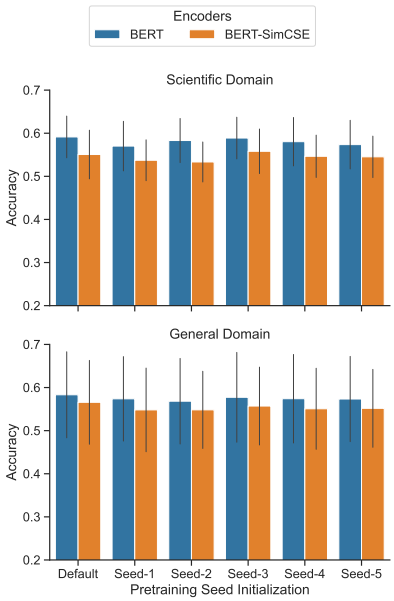



Figure 5: Sentence-level model accuracy (BERT and BERT-SimCSE) across different pretraining initializations. "Default" denotes Hugging Face models. Variance is across sentence ordering datasets.

the sentence ordering task, we report results for the BERT encoder in Figure 4, with results for BERT-SimCSE and SciBERT provided in the appendix (Figure 15). For the scientific domain datasets, mean-pooled representations from the first layer of BERT outperform SBERT. However, the reverse is true for the general domain, potentially due to SBERT's fine-tuning on general domain datasets. Across both domains, mean representations from the first layer are either better than or comparable to CLS representations from the last layer of BERT. This is likely because the first-layer representations encode higher syntactic, semantic, and discourse-level properties (see Figure 11).

## 4 Analysis and Discussion

### 4.1 Robustness of Findings Across BERT Initializations.

To ensure the robustness of our findings across different pretraining seed initializations of BERT encoder, we rerun all the sentence ordering experiments for BERT and BERT-SimCSE[3]. Using five randomly chosen off-the-shelf multiBERT models (Sellam et al., 2021), we train different versions of BERT-SimCSE starting from these models

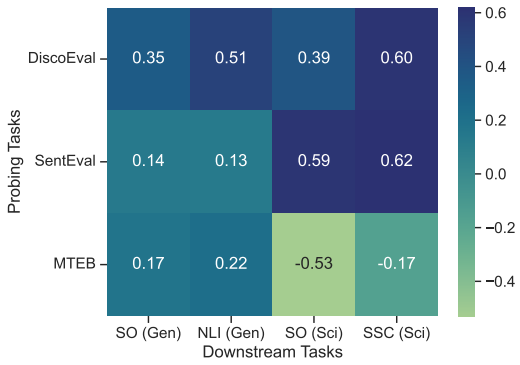



Figure 6: Correlations between probing and downstream task performance for encoders with different training regimes and domains. Gen: General, Sci: Scientific.

using the official implementation[4]. We then obtain representations from these trained encoders to train sentence-level models, reporting the results in Figure 5. The similar performance trends observed for BERT and BERT-SimCSE indicate the robustness of our findings. Similar trends were also observed for multiBERT models using different pooling strategies (see Figure 14 in appendix).

### 4.2 Syntactic and Discourse Properties Drive Downstream Performance, Not MTEB.

As established in Section 3, even mean-pooled representations from vanilla BERT encoder achieve comparable or better performance than representations from specifically trained encoders. To understand why this occurs and identify the properties crucial for strong performance on sentence-level tasks, we evaluate the syntactic, semantic, and discourse-level properties encoded in the representations using SentEval (Conneau et al., 2018) and DiscoEval (Chen et al., 2019) probing tasks[5]. We also investigate whether high performance on MTEB (Muennighoff et al., 2023), a widely used representation evaluation benchmark, correlates with high downstream performance. Figure 6 shows the Spearman's correlation between performance on the evaluation benchmarks and the downstream tasks. Our findings indicate that syntactic and discourse-level properties are more strongly associated with downstream task success. Surprisingly, higher MTEB performance does not guarantee better downstream performance. In fact, we even observe a negative correlation for the sen-

[3]We only compare performance for these encoders as we were unable to reproduce SBERT results and CMLM code is not publicly available

[4]https://github.com/princeton-nlp/SimCSE

[5]We focused on the probing tasks available in these benchmarks, as the downstream tasks are also included in MTEB.

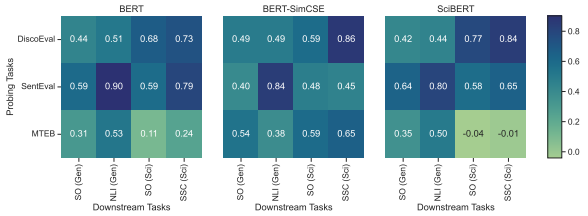



Figure 7: Correlations between probing task and downstream task performance for encoders with different pooling strategies. Gen: General, Sci: Scientific.

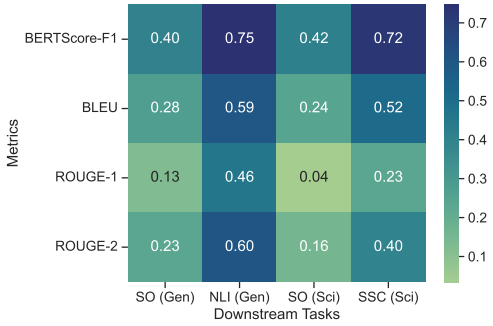



Figure 8: Correlations between decodability (measured via BLEU, ROUGE, and BERTScore) and downstream task performance for various encoders. Gen: General, Sci: Scientific.

tence ordering task. Further analysis of individual encoder scores (see Figure 10 in appendix) reveals that BERT representations exhibit the strongest discourse-level properties and comparable syntactic properties, contributing to their strong downstream performance. A potential explanation is that less mature representations, having undergone less task-specific fine-tuning, retain greater flexibility to adapt to the specific requirements of the downstream task.

A similar analysis for different pooling strategies across BERT, BERT-SimCSE, and SciBERT yielded similar results (Figure 7) . However, in this case, MTEB performance showed a stronger correlation with downstream performance, suggesting that semantic properties become more influential when the base encoder is held constant. Further analysis (see Figures 11 to 13 in appendix) indicates that CLS-BP outperforms CLS across all probing benchmarks, explaining its superior downstream performance.

### 4.3 Decodability Does not Always Equal Downstream Success.

We hypothesize that if, for a given sentence encoder, sentences can be decoded with high fidelity

| Task | Dataset | Accuracy | |
|---|---|---|---|
| | | Token | Sentence |
| SO | NIPS | 36.31* | **54.29** |
| | AAN | 48.55* | **63.98** |
| | SIND | 48.30 | **48.33** |
| | RocStories | 61.15 | **68.30** |
| SSC | CSAbstruct | 73.76 | **74.06** |
| NLI | ANLI | **47.38** | 43.78 |

Table 4: Accuracy (%) for different tasks from a token-level and corresponding sentence-level model. $*$ denotes that the model was trained with 8 permutations per training example.

from their encoded representations, then this encoder will have strong downstream task performance in a sentence-level modeling paradigm. This hypothesis is based on recent findings that sentence representations are a bottleneck for sentence-level models (Kamath et al., 2023). To test this hypothesis, we train a T5 decoder (Raffel et al., 2020) for each encoder to reconstruct 2 million training instances from the BookCorpus dataset (Zhu et al., 2015). The decoder is conditioned on the fixed-size representation yielded by the encoder *i.e.,* a single vector input to T5's cross-attention mechanism.

Using beam search decoding (beam size = 4), we reconstruct sentences from the representations produced by each encoder. We then calculate the Spearman's correlation coefficient between decodability (measured using BLEU (Papineni et al., 2002) ROUGE (Lin and Hovy, 2003), and BERTScore (Zhang et al., 2019)) and downstream task accuracy across the encoders. Results are shown in Figure 8. The strong correlation observed for NLI and SSC tasks, as measured by BERTScore, supports our hypothesis. However, the weak to moderate correlation observed for the sentence ordering task suggests that higher decodability does not always translate to improved downstream performance.

For completeness, the individual BLEU, ROUGE, and BERTScore metrics for each encoder are available in the appendix (Figure 9).

### 4.4 The Efficiency Advantage of Sentence-Level Models.

We hypothesize that token-level models excel at fine-grained downstream tasks, while sentence-level models are better suited for coarse-grained tasks We define fine-grained tasks as those requir-

ing detailed analysis of individual words or phrases (*e.g.,* Named Entity Recognition or NLI), while coarse-grained tasks operate at the sentence or document level (*e.g.,* Sentence Ordering, Sentiment Classification). Token-level models can access all token representations across sentences in the input, potentially capturing more fine-grained semantic and discourse information.

To test this, we fine-tune a standard pretrained BERT model for all the tasks (see Appendix B in appendix) and compare its accuracy to that of corresponding sentence-level models trained using mean-pooled representations from the BERT encoder. For the coarse-grained tasks (SO and SSC), sentence-level models achieve comparable or better accuracy than token-level models (Table 4). Conversely, token-level models outperform sentence-level models on the fine-grained NLI task.

For the NIPS and AAN datasets, where the number of sentences per input is significantly higher than in other datasets, sentence-level models achieved $65 - 75\%$ higher accuracy than token-level models, despite using eight times fewer training examples. This suggests that sentence-level models are more effective at handling longer contexts. as opposed to a limit of 512 tokens in the token-level models. Furthermore, the sentence-level models converged faster (10 epochs) than the token-level models (15 epochs), demonstrating their training-time efficiency.

## 5 Related Work

**Sentence representations for language modeling.** Sentence-level modeling has been applied to various tasks, including sentence ordering (Cui et al., 2018; Basu Roy Chowdhury et al., 2021; Kumar et al., 2020; Golestani et al., 2021; Bin et al., 2023), sentence infilling (Huang et al., 2020; Mori et al., 2020), sequential sentence classification (Cohan et al., 2019; Hillebrand et al., 2024) and story continuation (Ippolito et al., 2020). More recently, sentence-level (or concept-level) modeling has been explored to model language at a higher abstraction level (LCM team et al., 2024). Related to our work, Czinczoll et al. (2024) predicts representations of masked chunks instead of tokens. They show that the base encoder effectiveness doesn't always translate to chunk-level performance. However, their work doesn't explore the diverse encoders, training regimes, and pooling strategies that are the focus of our work.

**Probing sentence representations** Probing sentence representations is a common technique to assess how linguistic properties are captured in learned representations (Ettinger et al., 2016; Veldhoen et al., 2016; Adi et al., 2016; Conneau et al., 2018; Hupkes et al., 2018; Conneau and Kiela, 2018; Chen et al., 2019). Shi et al. (2016) study syntactic knowledge in machine translation, and Vanmassenhove et al. (2017) investigate aspects of machine translation systems, showing that tense information can be extracted but is often lost during decoding. Conneau et al. (2018) introduces probing tasks to examine surface-level, syntactic, and semantic properties and correlate them with downstream tasks such as NLI and MT. Hupkes et al. (2018) train diagnostic classifiers to extract information from a sequence of hidden representations, hypothesizing that high classifier accuracy indicates the network's ability to track specific information. Giulianelli (2018), on the other hand, use these classifiers to predict numbers from the internal states of a language model. Kim et al. (2019) study what different NLP tasks teach models about function word comprehension. See Belinkov and Glass (2019) for a survey of related work.

SemEval (Agirre et al., 2012) is a common benchmark for evaluating sentence representations. Conneau and Kiela (2018) aggregate multiple STS datasets to address the limited expressivity of individual SemEval datasets and focus on fine-tuning classifiers on top of representations. However, it lacks retrieval or clustering tasks where representations could be directly compared without additional classifiers. MTEB Muennighoff et al. (2023) unifies datasets from different representation tasks into one evaluation framework to provide a holistic performance review of sentence encoders. While these benchmarks focus on task-specific and linguistic properties, Chen et al. (2019) (DiscoEval) propose tasks to assess the discourse-level information.

## 6 Conclusion

This work provides a comprehensive analysis of sentence encoders for sentence-level modeling, establishing the importance of representation maturity, the crucial role of syntactic and discourse-level properties, and the efficiency of sentence-level models. Our findings challenge views regarding the benefits of extensive fine-tuning and highlight the potential of less mature representations for greater adaptability (*e.g.,* exploring architectural modifi-

cations to preserve representation flexibility). The observation that syntactic and discourse-level properties are key drivers of downstream performance suggests fruitful avenues for future research such as, developing novel pretraining objectives that explicitly target these properties and move beyond MTEB evaluation. Further, the data and time efficiency of sentence-level models coupled with their competitive performance, makes them a promising direction for future work in language modeling. We believe these insights will contribute to the development of more effective and efficient sentence representation learning techniques.

## Limitations

We acknowledge that our findings are limited to classification and regression tasks. Future works may consider extending this study to downstream generation tasks. Our study focused on encoder-only BERT-based sentence encoders in English to ensure a controlled setting. However, further investigation is required to determine whether our findings are generalizable to multilingual encoders. While we show that a sentence-level model is better or comparable to a token-level model, a promising future direction is to consider a hierarchical model that first operates on token-level followed by a sentence-level model to investigate if it results in a best-of-both-worlds model wherein both fine-grained and discourse-level information could be learned by the model for performing any downstream tasks. We acknowledge that 'maturity' in this work encompasses two key aspects (1) how advanced we are in the finetuning stage as well as (2) the location of the features in the network (early vs. late layer) and the primary focus of this work was on identifying a common thread between them and examining the high-level relationship with downstream performance and property probes, such as syntactic and discourse properties. However, future research may aim to disentangle these two factors and provide formal definitions.

## Acknowledgements

We would like to thank all the anonymous reviewers and researchers at FAIR for their useful and constructive feedback.

## References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. in* sem 2012: The first joint conference on lexical and computational semantics–volume 1: Proceedings of the main conference and the shared task, and volume 2: Proceedings of the sixth international workshop on semantic evaluation (semeval 2012). *Association for Computational Linguistics. URL http://www. aclweb. org/anthology/S12-1051*.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.

Somnath Basu Roy Chowdhury, Faeze Brahman, and Snigdha Chaturvedi. 2021. Is everything in order? a simple way to order sentences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10769–10779, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

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.

Yi Bin, Wenhao Shi, Bin Ji, Jipeng Zhang, Yujuan Ding, and Yang Yang. 2023. Non-autoregressive sentence ordering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4198–4214, Singapore. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. 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 649–662, Hong Kong, China. Association for Computational Linguistics.

Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, and Dan Weld. 2019. Pretrained language models for sequential sentence classification. 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 3693–3699, Hong Kong, China. Association for Computational Linguistics.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. 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.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2018. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349.

Tamara Czinczoll, Christoph Hönes, Maximilian Schall, and Gerard De Melo. 2024. NextLevelBERT: Masked language modeling with higher-level representations for long documents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4656–4666, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. 2023. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv e-prints*, pages arXiv–2308.

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.

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

Mario Giulianelli. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*.

Melika Golestani, Seyedeh Zahra Razavi, Zeinab Borhanifard, Farnaz Tahmasebian, and Hesham Faili. 2021. Using bert encoding and sentence-level language model for sentence ordering. In *International Conference on Text, Speech, and Dialogue*, pages 318–330. Springer.

Lars Hillebrand, Prabhupad Pradhan, Christian Bauckhage, and Rafet Sifa. 2024. Pointer-guided pretraining: Infusing large language models with paragraph-level contextual awareness. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 386–402. Springer.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Yichen Huang, Yizhe Zhang, Oussama Elachqar, and Yu Cheng. 2020. Inset: Sentence infilling with intersentential transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2502–2515.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch. 2020. Toward better storylines with sentence-level language models. *arXiv preprint arXiv:2005.05255*.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, Singapore. Association for Computational Linguistics.

Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, et al. 2019. Probing what different nlp tasks teach machines about function word comprehension. *arXiv preprint arXiv:1904.11544*.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. 2020. Deep attentive ranking networks for learning to order sentences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8115–8122.

LCM team, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, Joao Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. 2024. Large Concept Models: Language modeling in a sentence representation space.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 150–157.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *International Conference on Learning Representations*.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2016. Sentence ordering using recurrent neural networks.

Yusuke Mori, Hiroaki Yamane, Yusuke Mukuta, and Tatsuya Harada. 2020. Finding and generating a missing part for story completion. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 156–166.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Dragomir R Radev, Mark Thomas Joseph, Bryan Gibson, and Pradeep Muthukrishnan. 2016. A bibliometric and network analysis of the field of computational linguistics. *Journal of the Association for Information Science and Technology*, 67(3):683–706.

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. *Journal of machine learning research*, 21(140):1–67.

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.

Cécile Robin, Atharva Kulkarni, and Paul Buitelaar. 2023. Identifying FrameNet lexical semantic structures for knowledge graph extraction from financial customer interactions. In *Proceedings of the 12th Global Wordnet Conference*, pages 91–100, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

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.

Thibault Sellam, Steve Yadlowsky, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Turc, Jacob Eisenstein, Dipanjan Das, et al. 2021. The multiberts: Bert reproductions for robustness analysis. *arXiv preprint arXiv:2106.16163*.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural mt learn source syntax? In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1526–1534.

Eva Vanmassenhove, Jinhua Du, and Andy Way. 2017. Investigating'aspect'in nmt and smt: translating the english simple past and present perfect. *Computational Linguistics in the Netherlands Journal (CLIN)*, 7:109–128.

Sara Veldhoen, Dieuwke Hupkes, Willem H Zuidema, et al. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@ NIPS*, pages 69–77. Barcelona.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Ziyi Yang, Yinfei Yang, Daniel Cer, Jax Law, and Eric Darve. 2021. Universal sentence representation learning with conditional masked language model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6216–6228, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

## A Implmentation Details

**Experimental Settings** Each model has $85,664,257$ parameters and is trained for $10$ epochs. For each task, we use hyperparameters (batch size $\in \{4, 8, 16, 32, 64\}$, learning rate $\in \{1e-5, 2e-5, 3e-5\}$, warmup ratio $\in \{0.06, 0.03, 0.3, 0.6\}$) that give the best validation accuracy.

**Dataset** We use the datasets made available publicly by the respective authors. We use all three rounds of data for ANLI dataset.

## B Token-level Model Details

We fine-tune a vanilla BERT-base model on the three tasks as follows: (1) for the sentence ordering task, we prepend each sentence with a [SEP] token and use the final layer representation corresponding to this token to train the model on the regression task of predicting the relative position of each sentence using mean-squared loss; (2) for the sequence classifications task, we take the representations corresponding to the [SEP] token and train on a 5-class classification task using the cross-entropy loss; and (3) for the NLI task, we take the representation corresponding to the [CLS] token to train the model on the 3-class classification task using cross-entropy loss. Each model has $108,929,281$ parameters and is trained for $15$ epochs. For AAN and NIPS datasets, the model was trained using $1$ transformer layer ($30,962,689$ parameters) to avoid underfitting. For each task, we use hyperparameters (batch size $\in \{4, 8, 16, 32, 64\}$, learning rate $\in \{1^{e-5}, 2^{e-5}, 3^{e-5}\}$, warmup ratio $\in \{0.06, 0.03, 0.3, 0.6\}$) that give the best validation accuracy.

## C Decoder Training Details and Results

We train the decoder of T5-large (Raffel et al., 2020) conditioned on a fixed sentence representation obtained from each encoder we study in this work. To match the dimension of the representations from each encoder and the T5-large decoder, we add a linear layer followed by layer normalization. We keep the sentence encoder frozen and train the decoder for 1 epoch on randomly sampled 2M examples from the Bookcorpus dataset (Zhu et al., 2015) available on huggingface[6]. Sentences in this dataset are $14 \pm 9$ words long and the model has $436,416,000$ parameters which took $19hrs$ to train on 1 Tesla V100-SXM2-32GB GPU. We use a batch size of 32, Adam optimizer with default parameters, and a learning rate of $1e^{-5}$. Decoding results are shown in Figure 9.

## D Probing Results

We perform the probing results using the official implementation for SentEval[7], DiscoEval [8], and MTEB [9] tasks. We only use the English language tasks in the MTEB benchmark excluding $15$ tasks for which we got a significant ($<$ $0.8$ Spearman's correlation) correlation between performance for encoders trained using different pretraining initialization seeds. These tasks include: FEVER, ArxivClusteringP2P, ImdbClassification, SICK-R, MedrxivClusteringS2S, STS13, ToxicConversationsClassification, RedditClusteringP2P, ArguAna, Touche2020, MedrxivClusteringP2P, BIOSSES, STS22, SummEval, BiorxivClusteringP2P.

We present the results related to specific scores for each encoder on the probing tasks in Figures 10, 11, 12, and 13.

## E Additional Results

We provide additional pooling layer results in Figure 15 for SciBERT, where representations from the last layer are better than the first layer for the scientific domain indicating that pretraining dataset domain is more important than other factors. However, for general domain datasets, mean representations from the first layer perform better than CLS representations from the last layer and SciBERT-NLI or comparable to Max representations from the last layer.

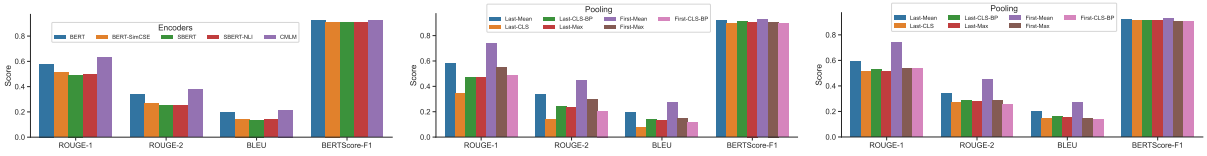
We provide results (Figure 14) for different pooling strategies when encoders with different pertaining initialization seeds were used.

[6] https://huggingface.co/datasets/bookcorpus/bookcorpus

[7] https://github.com/facebookresearch/SentEval

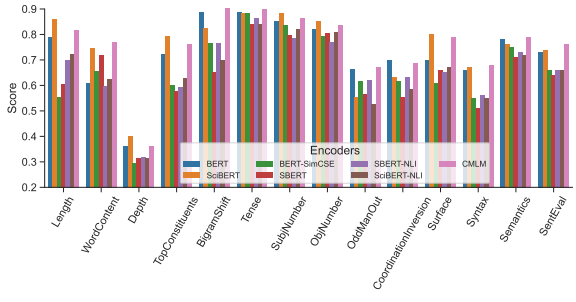
[8] https://github.com/ZeweiChu/DiscoEval

[9] https://github.com/embeddings-benchmark/mteb

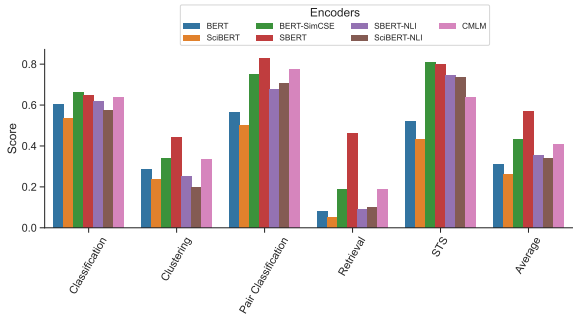



(a) Encoders (different training regimes) (b) BERT (different pooling strategies) (c) BERT-SimCSE (different pooling strategies)

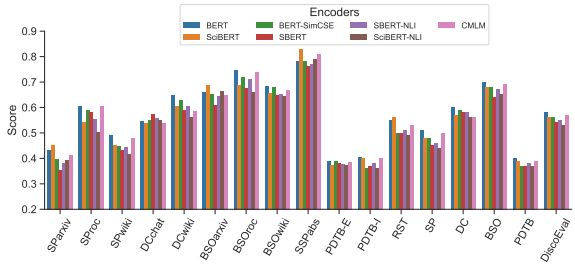
Figure 9: Metric scores for sentences decoded given their representations obtained from encoders with different training regimes (left), and using different pooling strategies from BERT (middle) and BERT-SimCSE (right) encoders.




(a) SentEval

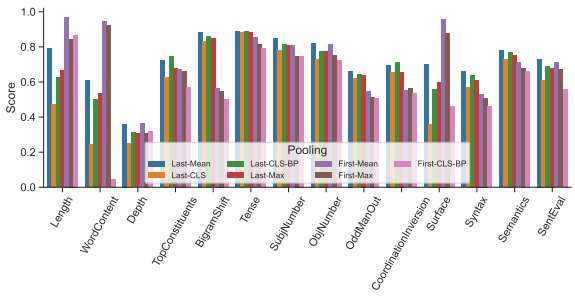



(b) MTEB

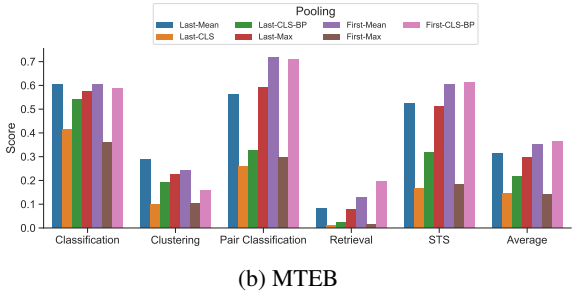



(c) DiscoEval

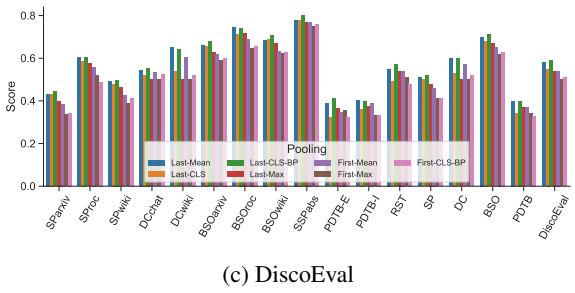
Figure 10: Probing task scores for encoders with different training regimes. **SentEval**, **Average** (MTEB), and **DiscoEval** show the averaged scores across all the tasks in these evaluation benchmarks.




(a) SentEval

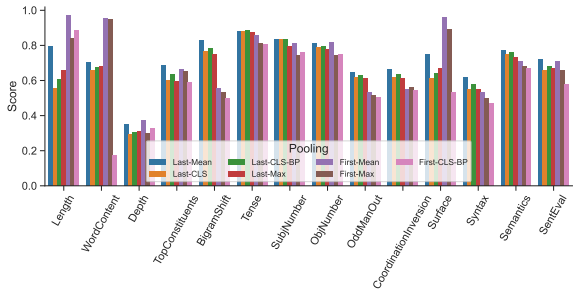



(b) MTEB

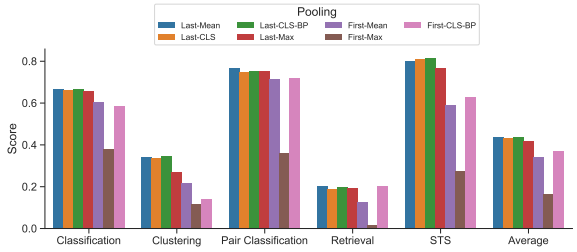



(c) DiscoEval

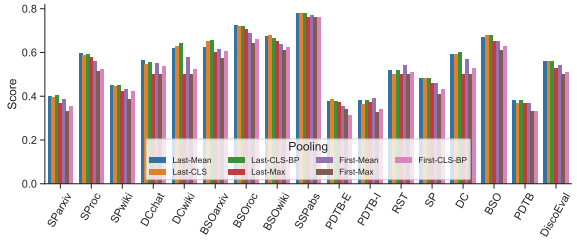
Figure 11: Probing task scores for **BERT** encoder with different pooling strategies. **SentEval**, **Average** (MTEB), and **DiscoEval** show the averaged scores across all the tasks in these evaluation benchmarks.




(a) SentEval

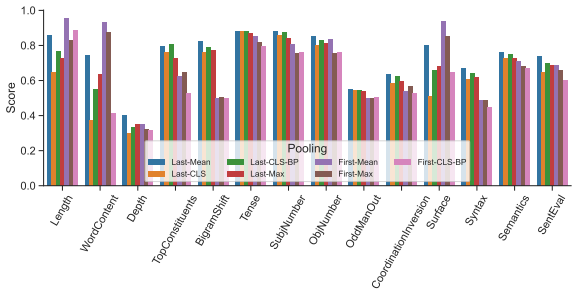



(b) MTEB

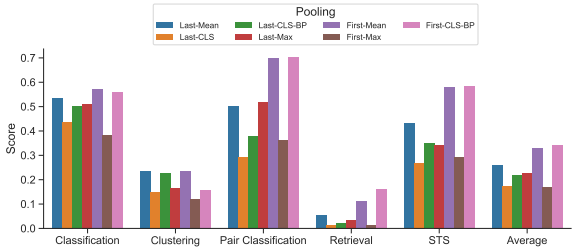



(c) DiscoEval

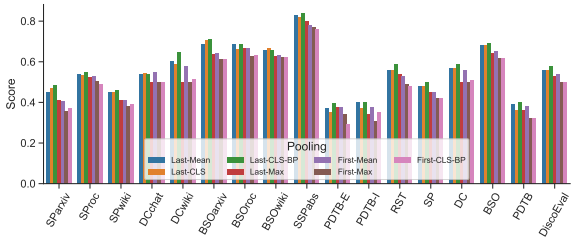
Figure 12: Probing task scores for **BERT-SimCSE** encoder with different pooling strategies. **SentEval**, **Average** (MTEB), and **DiscoEval** show the averaged scores across all the tasks in these evaluation benchmarks.




(a) SentEval




(b) MTEB




(c) DiscoEval

Figure 13: Probing task scores for **SciBERT** encoder with different pooling strategies. **SentEval**, **Average** (MTEB), and **DiscoEval** show the averaged scores across all the tasks in these evaluation benchmarks.

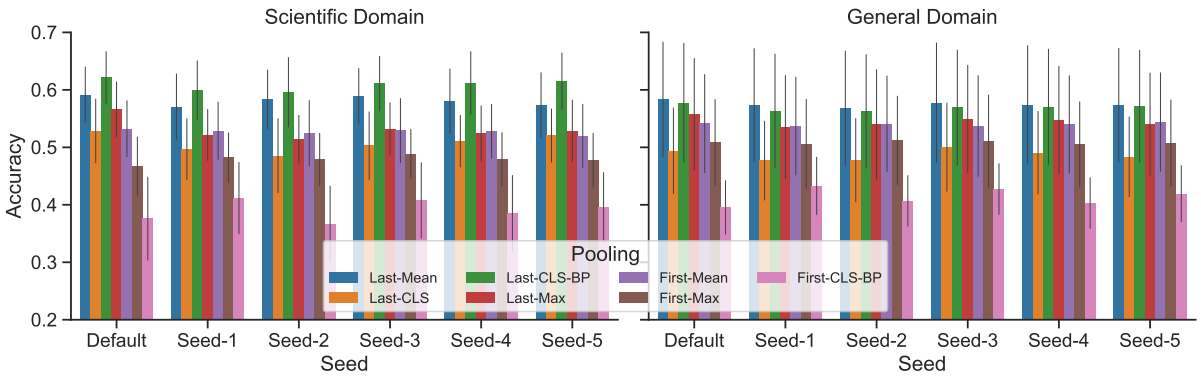
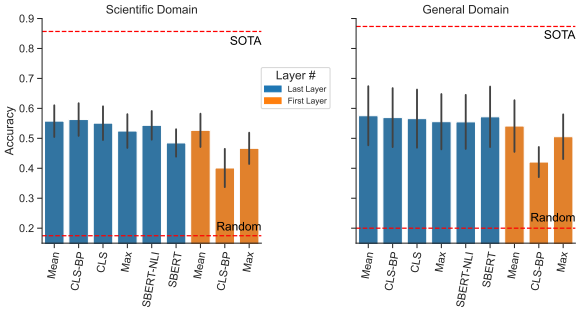
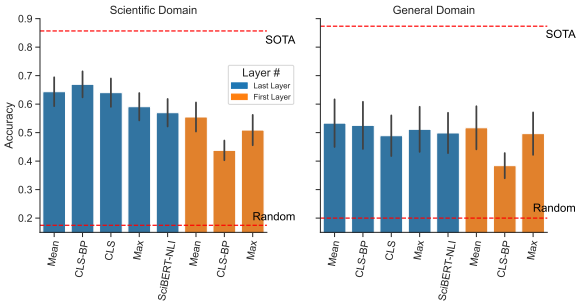



Figure 14: Accuracy for sentence-level models trained using representations (with different pooling strategies) from BERT encoders using different pretraining initialization seeds. Default denotes the scores obtained with default encoders available on huggingface. Variance is across datasets for the sentence ordering task.




(a) Different poolings from BERT-SimCSE encoder




(b) Different poolings from SciBERT encoder

Figure 15: Accuracy for sentence-level models trained using representations obtained by pooling from first and last layers of BERT-SimCSE (a) and SciBERT (b) encoders, and compared to other encoders. Variance is across datasets for the sentence ordering task.