STRUCTURED DOTS IN THE SENTENCE EMBEDDING SPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Embedding spaces allow us to manipulate texts as mathematical objects, most frequently as points or vectors paired with distance metrics as proxies of linguistic properties, mainly similarity or relatedness. Viewed through this perspective, the embedding space appears anisotropic, and thus not able to encode linguistic differences accurately enough, despite empirical proof that the embeddings are useful for a variety of tasks.

We propose that these observed shortcomings do not reflect properties of the embeddings themselves, but of the shallow measures such as cosine, which consider each dimension separately. By comparing the relative positions in the sentence embedding space of three sentence representation variations, and their performance on a variety of tasks, we show that sentence embeddings are complex objects with internal structure, which encodes linguistic structure in a systematic manner.

1 INTRODUCTION

Projecting text into an n-dimensional space maps linguistic objects into a well-defined mathematical space, with specific metrics and operations. In this space word similarity in language is equated with closeness between the corresponding vectors in the embedding space, that is, the embedding space is *smooth* Bengio et al. (2013). The smoothness of the embedding space comes with several assumptions: similar or related words or sentences will be projected to points that are close in the space, and words or sentences corresponding to points that are close in the space are similar or related.

Analysis of the embedding spaces through similarity measures (most frequently cosine) have revealed that the spaces of many pretrained transformer language models are anisotropic, with most words appearing in a narrow cone in this space, thus making similarity metrics seemingly less informative Mimno & Thompson (2017); Li et al. (2020); Timkey & van Schijndel (2021); Cai et al. (2021). Practically however, embeddings have been successfully plugged into numerous NLP tasks, and systematically evaluated using various benchmarks, e.g. the Massive [Multilingual] Text Embedding Benchmarks Muennighoff et al. (2022); Enevoldsen et al. (2025), HOLMES Waldis et al. (2024).

In this paper we present experiments that look more closely at the sentence embedding space, and show that measures, such as the cosine, that consider dimensions separately are shallow. The interplay among embedding dimensions is complex, as each can contribute to various linguistic features in different measures Bengio et al. (2013); Elhage et al. (2022). This implies that the level at which words and sentence embeddings share features is not a level that the cosine metric – or indeed, any of the commonly used metrics – can detect. Such geometric models have also been shown to be psychologically implausible, and a metric based on comparison of feature sets would be more appropriate Tversky (1977); A. Rodriguez & Merlo (2020); Doumbouya et al. (2025). Features being encoded as weighted combinations of subsets of dimensions also shed light on the apparent contradiction between the word embedding space being anisotropic, while embeddings serving as a crucial component of successful language models.

We focus here on sentence representations from the output layer of encoder transformers, as they are the ones that are used for solving downstream tasks. Briefly, our study compares three sentence representations in the same embedding space, from four transformer encoder models: the [CLS] token embedding, averaged token embeddings, and a random token embedding (Section 3.1). Cosine similarity between words and tokens in the same sentence shows high values, indicating a high level

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

of contextual information in these embeddings (Section 3.3). Cosine similarity between the three representation variations show that the [CLS] token embedding is not close – for some models almost orthogonal – to the other representation variations, despite encoding the same sentence (Section 3.4). Two types of analysis show that the conclusions that the cosine lead to are not representative of the sentence embeddings: task results for the FlashHOLMES benchmark that shows close performance for the seemingly very different embeddings (Section 4), and an analysis of the sentence embeddings using a Variational AutoEncoder-like system shows the three representation variations all encode structural sentence information the same way (Section 5).

The analyses we present provide further clues about the type of quasi-symbolic information encoded in continuous embeddings, and the nature of the embedding space, to help steer systems towards representations that are more abstract and thus generalize well across phenomena and languages.

2 Related work

Embeddings and linearity in the embedding space Projecting language units into the embedding space makes it natural to treat them as points or vectors in this space, and manipulate them as such. The linear representation hypothesis posits that semantic concepts are represented linearly in the representation spaces of language models. Mikolov et al. (2013); Elhage et al. (2021) show that vector offsets can encode female-male distinctions, grammatical number, verb tense, adjective comparatives; Hewitt & Manning (2019); Chi et al. (2020); Manning et al. (2020) find geometrical relations that mirror syntactic relations in a transformed embedding space; **?** detect linear structure in a dataset of true/false statements; Gurnee & Tegmark (2024) explore linear dimensions of space and time, and find more complex (e.g. circular) relations between specific temporal markers, such as days, months Engels et al. (2025); Park et al. (2024) search for conceptual hierarchies in the embedding space. Jiang et al. (2024) show that in generative LLMs, the next token prediction objective and the implicit bias of gradient descent together promote the linear representation of concepts.

The linear hypothesis was applied to sentences as well. Embedding dimensions of sentence embeddings – as the embedding of the special [CLS] token or the averaged token embeddings Devlin et al. (2019) – have been shown to encode some linguistic information: shallow information about sentences Nikolaev & Padó (2023c), sentence-level information (Tenney et al., 2019), including syntactic structure – reflected as relative positions in the embedding space that parallel a syntactic tree (Hewitt & Manning, 2019; Chi et al., 2020). Deeper exploration through simple classification probes (consisting of one linear NN), has shown that predicate embeddings contain information about their semantic roles structure (Conia & Navigli, 2022; Silva De Carvalho et al., 2023), embeddings of nouns encode subjecthood and objecthood (Papadimitriou et al., 2021).

Sentences contain numerous layers of information (e.g. structural, semantic, pragmatic), with local and distant connections between their words, groupings into phrases, nesting expressions. More complex approaches have worked on teasing apart syntactic and semantic information Mercatali & Freitas (2021); Bao et al. (2019); Chen et al. (2019). Nastase & Merlo (2023) have shown that it is beneficial to go beyond considering embeddings as one-dimensional arrays: embeddings reshaped as 2d arrays reveal internal structure and regularities that encode grammatical information.

Work on interpretability of transformer-based models have rarely focused exclusively on sentence embeddings. The most common type of analysis involves the internal parameters of the models, coupled with a task based testing of the output, i.e. the embeddings, using a benchmark. Variational AutoEncoders and their variations (quantized, sparse, etc.) have been useful in mining for features in the internal parameter space relevant to a variety of tasks Zhang et al. (2025b).

The geometry of the embedding space Viewed through the lens of a similarity metric like the cosine, the embedding space of transformer models appears to be anisotropic Mimno & Thompson (2017); Li et al. (2020); Timkey & van Schijndel (2021); Cai et al. (2021). Regardless of the source of anisotropy – a few dominant dimensions, that can skew the similarity profile of the space Timkey & van Schijndel (2021), or the optimization of the cross-entropy loss Gao et al. (2019); Biś et al. (2021) – and whether it is or not inherent in transformers Godey et al. (2024); Machina & Mercer (2024), it can adversely influence model training and fine-tuning. The embedding space was determined to contain isotropic clusters and lower-dimensional manifolds that reflect word frequency properties Cai

et al. (2021). The sentence embedding space was also shown to be anisotropic Su et al. (2021), and methods to enhance the isotropy have been proposed Su et al. (2021); Huang et al. (2021).

In this paper we focus on the output layer of encoder transformer models, and investigate the sentence embedding space using first cosine, and then show evidence that the image provided by this analysis is distorted. Sentence embeddings are complex objects that encode multiple layers of structural and semantic information, which can be revealed using latent factor analyses.

3 SENTENCE REPRESENTATION COMPARISONS IN THE EMBEDDING SPACE

Human similarity judgements are subjective and vague, and existing gold-standards capture average estimates, rather than reflecting an objective truth Wang et al. (2023). We bypass this issue by using instead three representations of the same sentence in the same embedding space, and thus compare embeddings that should encode the same information. The relative positions of these sentence representation variations using the cosine is presented in Section 3.4. In the next step, we test the sentence representations on the FlashHolmes benchmark Waldis et al. (2024) and compare their relative performances with their closeness in the embedding space (Section 4). If closeness is a good proxy for the amount of information shared by the corresponding points in the space, then points that are close should correspond to representations that lead to similar performance on NLP tasks. Finally, we test whether the three sentence representation variations all encode information about linguistic structure of a sentence, and if they do, whether it is encoded in the same way (Section 5). Should this be the case, it would support the idea that linguistic information is encoded in a more complex manner in sentence embeddings, which the cosine measure is too shallow to access.

3.1 SENTENCE REPRESENTATIONS

Averaged token embeddings (mean pooling): $\mathbf{S}_{AVG}$ The representation obtained by averaging a sentence's tokens (without the special [CLS] and [SEP] tokens) is frequently used as the sentence's embedding (Reimers & Gurevych, 2019a; Nikolaev & Padó, 2023a). This representation benefits from the fact that the learning signal for transformer models is stronger at the token level, with a much weaker objective at the sentence level – next sentence prediction (Devlin et al., 2019; Liu et al., 2019), sentence order prediction (Lan et al., 2019).

The embedding of the special [CLS] token: $\mathbf{S}_{CLS}$ The [CLS] token has been introduced in BERT Devlin et al. (2018) to serve as a representation of a sentence. It is most commonly used after fine-tuning for specific tasks such as story continuation (Ippolito et al., 2020), sentence similarity (Reimers & Gurevych, 2019b), alignment to semantic features Opitz & Frank (2022). BERT's descedants all include such a special token. While previous work has shown that this variation performs worse that S_{AVG}Reimers & Gurevych (2019b) (albeit only on the sentence similarity task), it was designed as a holistic representation of the sentence. We use it "raw", i.e. without fine-tuning, because fine-tuning distorts the embeddings space through the cosine metric, which we are investigating.

The embedding of a random token: $\mathbf{S}_{T_{rand}}$ The embedding of the root token has been used as a sentence representation Nikolaev & Padó (2023b). We use a random token (randomly sampled from the sentence's tokens, excluding the special [CLS] and [SEP] tokens), to show the impact of contextual representations, and how much contextual information each token embedding contains.

A potential question arises: are these representations comparable? Are they in the same embedding space? The answer is *yes*. While the token embeddings and the special [CLS] token receive a different feedback in training at the last layer, they interact throughout the system, and more obviously and directly in the attention modules: [CLS] embeddings are projected through the same query, key and value linear layers as the token embeddings, they interact with the other token embeddings in a symmetric manner, and the result of the attention computations are added to the residual stream just as for the other tokens. Furthermore, the [CLS] token embedding has sometimes been included as part of the averaged token representation Nikolaev & Padó (2023b); Hara et al. (2026).

We investigate these three variations of representing sentences in four pretrained encoder transformer models: BERT, RoBERTa, DeBERTa, and Electra[1]. BERT is the baseline transformer model. RoBERTa is a variation of BERT with optimized training, BPE tokenization, dynamic masking and without a next sentence objective Liu et al. (2019). DeBERTA is another variation that introduces disentangled attention and an optimized mask decoder training He et al. (2021). Unlike BERT, RoBERTa and DeBERTa, Electra is not a masked language model, rather implements a replaced token recognition model, predicting whether a token in the input was produced by a generator model Clark et al. (2020). Electra also outperforms XLNet Yang et al. (2019) and MPNet Song et al. (2020)[2]. By analysing this "lineage" of models, we can quantify and visualize the impact of the differences in training regime and architecture.

While currently generative models are more popular (both within and outside the research community), generative models do not provide the variations in sentence embeddings needed for this analysis. However, encoder models share the same transformer architecture, and our analysis provides insights into how the information becomes organized after it flowed through these layers. Encoder models require much fewer resources, making them more readily accessible outside the industry and large research centers, as recent, computationally optimized models show Warner et al. (2025). They also continue to be useful, particularly in tasks that rely on having a representation of a text fragment, e.g. as reranking locums for LLMS within a web-search engine (Ye et al., 2025).

3.2 DATA

The dataset consists of 6000 sentences: 1000 for each of six languages (English, French, German, Italian, Romanian, Spanish) extracted from the ParaCrawl corpus Bañón et al. (2020) (the datasets are not parallel).

3.3 CONTEXTUAL INFORMATION IN EMBEDDINGS

For each sentence s in the dataset, we compute the cosine similarity between the embeddings of every pair of tokens and every pair of words. Considering that most tokens in a sentence are not actually similar, and that the token embeddings are contextual, this analysis shows how much contextual information these embeddings share. The density histogram plots are shown in Figure 1. For BERT, the similarities among the token representations have a wider distribution, while they become tighter and centered on a higher mean for the optimized BERT variations, RoBERTa and DeBERTa, and for Electra. The word embeddings – as averages of their token representations – follow similar trends. These results show that with the changes in the models (relative to BERT), tokens and words within the same sentence become closer in the embedding space. According to the assumption that close points correspond to similar words and vice-versa, we must conclude that the optimizations over the BERT base model lead to more sharing of contextual information among the words in a sentence.

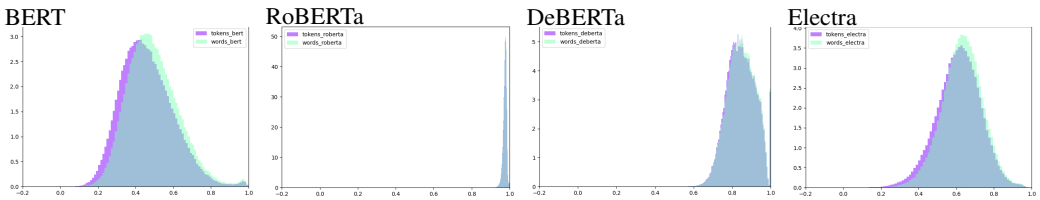


Figure 1: Histograms of cosine similarities for words and tokens in 1000 English sentences (results for French, German, Italian, Romanian and Spanish in appendix). The y-scales are different for each subplot for better visualization.

The next step is an analysis of the similarities between the three types of embeddings for several models from the BERT family, shown in Figure 2. For each sentence s in the dataset, we compute the cosine between every pair of representations $\mathrm{S}_{AVG}(s)$, $\mathrm{S}_{CLS}(s)$ and $\mathrm{S}_{T_{rand}}(s)$.

[1] `google-bert/bert-base-multilingual-cased`, `FacebookAI/xlm-roberta-base`, `microsoft/deberta-v3-base`, `google/electra-base-discriminator`

[2] Table 3 in Clark et al. (2020) and MPNet (manual comparison between bottom of Table 3 in Song et al. (2020) and Table 3 in Clark et al. (2020).

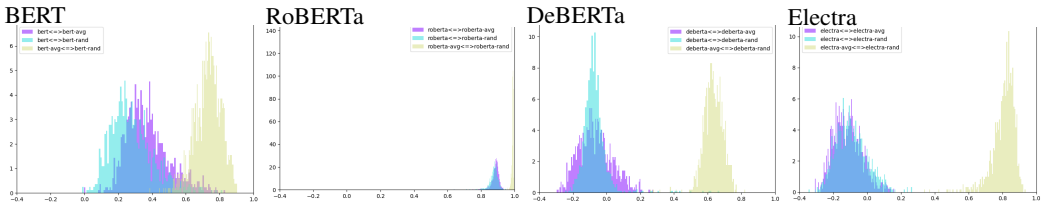


Figure 2: Histograms of cosine similarities for 1000 English sentences (results for French, German, Italian, Romanian and Spanish in appendix). In yellow are the similarities between S_{AVG} and $S_{T_{rand}}$, in blue the similarities between S_{CLS} and $S_{T_{rand}}$, and in purple the similarities between S_{CLS} and S_{AVG}. The y-scales are different for each subplot for better visualization.

3.4 Sentence representation variations in the embedding space

For all models, S_{AVG} are very close to $S_{T_{rand}}$, adding support to the observation from Section 3.3, that tokens encode much contextual information. The holistic sentence embeddings S_{CLS} are quite dissimilar from both S_{AVG} and $S_{T_{rand}}$ for all models except for RoBERTa. The optimized training of RoBERTa has the effect of bringing all variations of the sentence embeddings closer together.[3] Note that S_{CLS} are almost orthogonal to S_{AVG} and $S_{T_{rand}}$ for both DeBERTa and Electra, seemingly indicating that the holistic S_{CLS} embeddings encode different information than the contextual token embeddings. This will be tested in the next sections.

4 Sentence representation comparisons on linguistics tasks

This section shows experiments that investigate whether the similarities or differences noted in the embedding space analysis are reflected in their relative performances.

4.1 Dataset and code

We use the FlashHolmes benchmark Waldis et al. (2024) to test the three embedding types. The benchmark contains: (i) 216 tasks grouped into five categories – morphology, syntax, semantics, discourse and reasoning, (ii) code to test new models, and (iii) a leaderboard for comparisons. The input of a task is obtained from the specified model, and the output is predicted using a classifier probe implemented as a linear NN layer. The results on these tasks will produce a more informative picture of the type of information encoded in the three representation variations.

4.2 Performance comparisons

Figure 3 presents a summary of the performance of the different sentence representation methods for each task, and on the task averages.[4] The first four bar groups in each panel correspond to each model, and show the number of times each of the three sentence representation methods (represented as different shades of the model colour) has achieved the best performance for that model. The fifth group shows the number of times each sentence variation for each model has had the best overall performance. The sixth panel shows these statistics as averages over all tasks.

On morphology, syntax and semantics the S_{AVG} has most frequently the highest performance for all models, while for reasoning and discourse $S_{T_{rand}}$ is often the best. However, the results in Figure 4 show little variation in the results on the reasoning and discourse tasks, indicating that probably all results are close to the tasks baseline. For the morphology, syntax and semantics tasks, the performance of Electra's S_{AVG} and S_{CLS} sentence representations are very close in terms of average performance, while $S_{T_{rand}}$ is much lower. This seemingly contradicts the analysis in Section 3.4 which shows that S_{AVG} and S_{CLS} representations are almost orthogonal, while S_{AVG} and $S_{T_{rand}}$ are very close.

[3] Additional results are presented in the appendix: analysis of the cosine measure within the [CLS] and averaged subspaces, the Euclid measure in Section B.4 and an analysis of the correlation between the cosine similarity and sentence lengths in Section B.2.

[4] For detailed (task-level) results see figures 11-12 and tables 2–6 in the appendix.

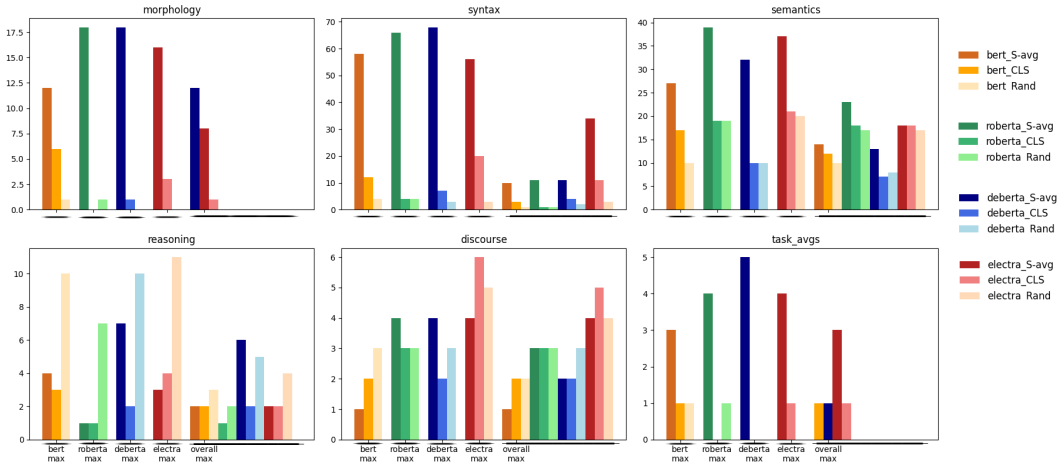


Figure 3: Performance as the counts of tasks for which each sentence representation performs best out of a model's variations (first 4 sets of columns) and overall (the 5th set of columns).

S_{AVG} and S_{CLS} embeddings have high and close performance for most task types. Compared with the analysis of the cosine similarity among them, this result is unexpected, as for Electra and DeBERTa in particular, the S_{CLS} and the S_{AVG} embeddings are almost orthogonal. Not only these embeddings have similar performance, but even for variations of the same task S_{AVG} gives best results for one task, and S_{CLS} gives best results for the other. For RoBERTa, where the cosine similarity between these two variations is very high, their relative performance is very different[5].

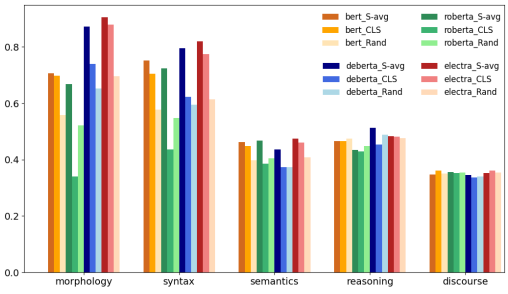


Figure 4: Average accuracy performance of embedding variations on the FlashHolmes benchmark

This supports the hypothesis that the cosine measure does not reflect the shared information in these representations. There can be an alternative explanation for the close performance of S_{AVG} and S_{CLS}: they encode the same information, but in different ways. We explore this in the next section.

5 Probing for Structure

In this section we test whether the same information is encoded in different ways in the three representation types, by focusing on sentence structure. Sentence structure is complex, relying on clues about phrase boundaries and phrase properties. We test whether information about sentence structure can be detected in the three sentence representation variations, and whether it is encoded in a similar manner, i.e. a system that is trained on one type of representations can identify the same structures in the other two representations without additional training. For this we use the experimental set-up and data of Nastase & Merlo (2024)[6], who have shown that some types of structural information – noun, prepositional, or verb phrase (chunks) structure – is recoverable from sentence representations.

[5] e.g. blimp_determiner_noun_agreement_with_adj_irregular_(1 and 2), blimp_irregular_plural_ subject_verb_agreement_(1 and 2), blimp_principle_A_case_(1 and 2), blimp_principle_A_domain_(1 and 2) – in appendix.

[6] https://github.com/CLCL-Geneva/BLM-SNFDisentangling

5.1 DATA AND CODE

The chunk-number dataset (CN) is a syntactic structure dataset consisting of English sentences with the syntactic pattern np (pp$_1$ (pp$_2$)) vp[7], where each np, pp$_1$, pp$_2$, vp can be in the singular or plural form, and the subject (np) always agrees with the verb (vp). This dataset is extracted from the BLM datasets used to explore the capability of language models to perform subject-verb agreement An et al. (2023); Nastase et al. (2024b).

The syntax-semantic roles dataset (SSR) consists of sentences whose syntactic structure and semantic roles highlight the variations associated with the spray/load alternation Samo et al. (2023):

syntactic-semantic pattern				example
NP-Agent	Verb	NP-Loc	PP-Theme	I pumped the pipes with regular air
NP-Theme	VerbPass	PP-Agent		The grain was sown by the monks
NP-Theme	VerbPass	PP-Loc	PP-Agent	Regular air was pumped into the stove by me
NP-Theme	VerbPass	PP-Loc		The contents were stuck in a box
NP-Loc	VerbPass	PP-Agent		The suitcase was loaded by the buyer
NP-Loc	VerbPass	PP-Theme	PP-Agent	The field was scattered with seeds by the workers
NP-Loc	VerbPass	PP-Theme		The surface was splattered with the stuffing

An instance is built for each sentence as a triple (in, out^+, Out^-), where in is an input sentence with chunk structure p, out^+ is a sentence different than in but with the same chunk structure p, and Out^- is a set of $N_{negs} = 7$ randomly selected sentences each of which has a chunk pattern different from p (and different from each other). Both datasets consist of 4004 instances, uniformly split over the patterns of the input sentence. For each run, the 4004 instances are split randomly into train:dev:test – 2576:630:798 for CN / 2569:637:798 for SSR – while maintaining balance between input sentence patterns on all three subsets.

The system is a variational encoder-decoder, building on the idea that variational auto-encoders are particularly appropriate for finding latent structure Kingma & Welling (2019); Zhang et al. (2025a). The encoder consists of a CNN layer that splits the input sentence embedding into layers of information, and a linear layer to compress the output of the CNN into a latent representation. The decoder is a mirror of the encoder. Sentence embeddings have 768 dimensions, and are presented to the system as a 2d-array of size 32x24. The encoder compresses the input to a latent layer of size 5. To encourage the targeted structure to emerge on the latent layer, each input in will be decoded into out^+ – a different sentence but with the same chunk structure – using the $N_{negs} = 7$ sentences with different structure than the input as contrastive examples. While the system receives a supervision signal – the correct output – it does not receive explicit information about a sentence's structure.

5.2 PERFORMANCE COMPARISON

We present two perspectives of the performance: (i) in terms of F1 averages over three runs (how well does the system perform in building a sentence representation closest to the correct one) and (ii) an analysis of the latent layer of the system, shown in Figure 5.

BERT and Electra in particular show very high results, with results on $S_{T_{rand}}$ even higher than S_{CLS}. DeBERTa is a bit of an outlier, both with respect to detecting the chunk structure and on the correlation between similarity and sentence length (Section B.2). These differences may come from DeBERTa's optimized training, with disentangled attention matrices and token embeddings with separate position and content sections, which leads to differently organized embeddings than BERT, RoBERTa and Electra. The encoder-decoder system was used "as is", without tuning any parameters for a specific model. For DeBERTa, fine-tuning or changing the architecture to properly exploit its working assumptions, seems to be necessary. We will investigate this in future work. While disappointing at a first glance, this is a positive negative result: it strengthens the conclusion that the results are indeed the effect of discovering structural information encoded in the representations, and are not an effect of the "power" of the probe.

To determine whether the variations in sentence representation encode the same information in the same manner, we look at the cross-testing results – training on one representation, and testing on the

[7]The BNF notation: pp$_1$ and pp$_2$ may be included or not, pp$_2$ may be included only if pp1 is included

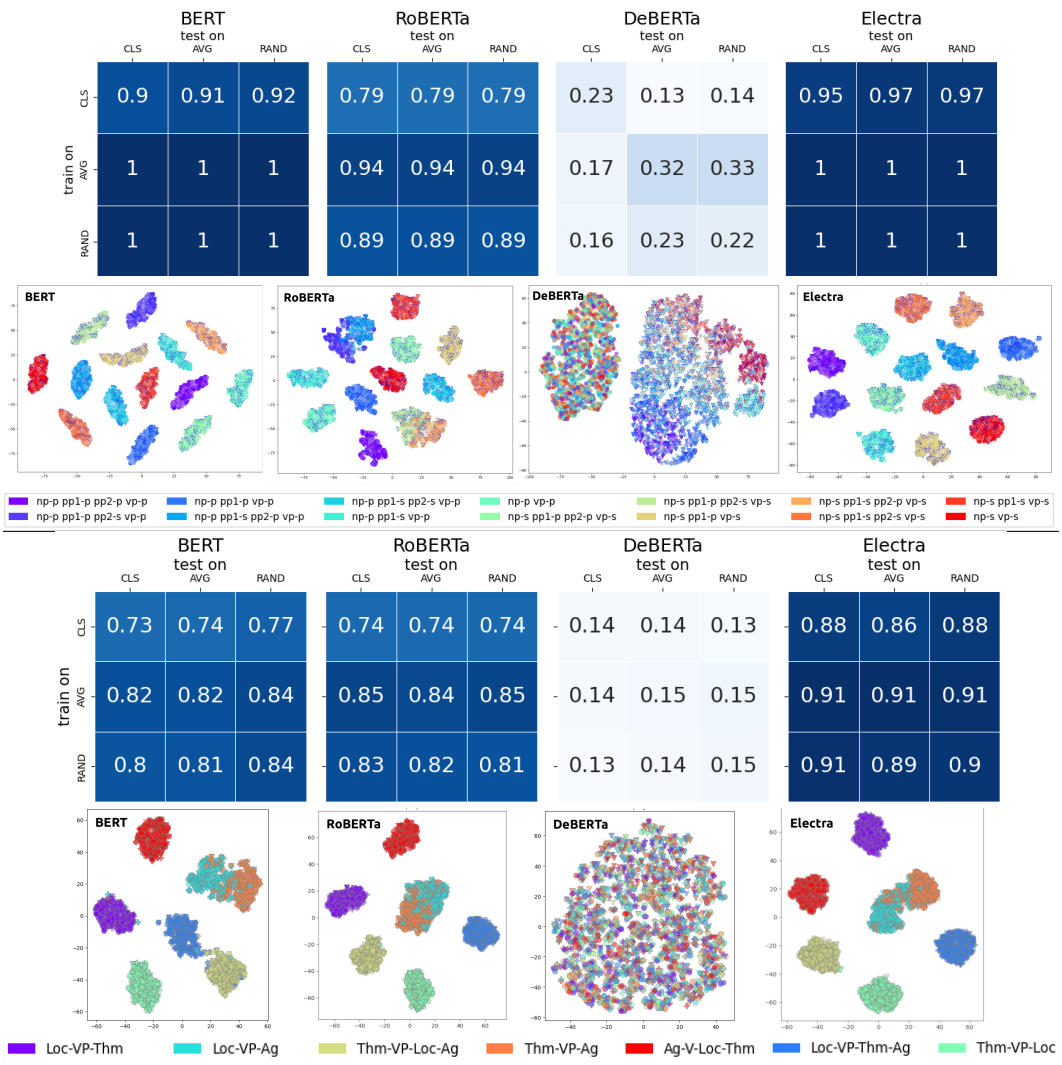


Figure 5: Evidence for similar localization of linguistic information: (i) training and testing on all variations (average F1 scores over three runs); (ii) tSNE plots of the projection on the latent layer of all representation variations (S_{AVG} ◯, S_{CLS} ▽, $S_{T_{rand}}$ +), for a model trained on the S_{AVG} . Results on the CN dataset on top, and the SSR dataset on the bottom.

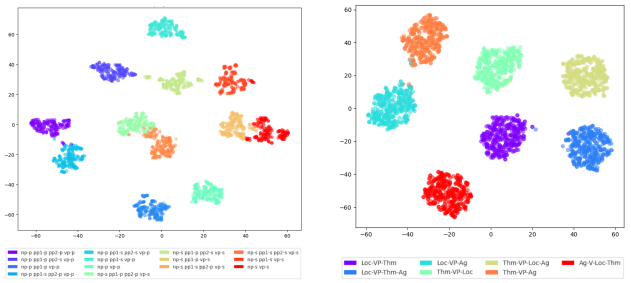
others (Figure 5). Despite the differences revealed by the cosine similarity analysis, where for Electra the S_{CLS} representations are almost orthogonal to S_{AVG} and $S_{T_{rand}}$, these experiments show that all three representations encode information about the chunk patterns and the syntactic-semantic structure in a sentence, and moreover, this information is encoded in the same manner. Additional support for this hypothesis comes from the analysis of the latent layer. Figure 5 show the tSNE projection[8] of the latent representations obtained from a model trained using the S_{AVG} sentence representations. After training on the training data represented with S_{AVG}, we pass through the encoder the sentences in the dataset encoded with all sentence representation variations, collect all latents and project them in 2D using tSNE.

The results on the task and the analysis of the latent vectors provide complementary views about whether structural information is encoded in the sentence embeddings. High results on the task (choosing the sentence with the same structure) may come from each sentence representation type

[8]The tSNE projection preserves the neighbourhoods. We use the tSNE implementation from scikit-learn, with 2 components and default parameters.

being mapped into a different area in the latent space. The tSNE plots show that for sentences with the same structure, all three sentence representations are mapped onto shared regions. This is highly significant: it indicates that the clues based on which the structure is detected is encoded similarly in the three sentence representations, regardless of the differences among them revealed by the previous experiments. For both datasets, clusters that are minimally different – have the same number of chunks/phrases, but differ in the grammatical number of one of the phrases, or the semantic roles – are well separated. The two clusters that are very close in the SSR latent projections correspond to the LOC-VP-AG and THM-VP-AG, which differ only in the semantic role of the subject of the passive verb. This shows that the latent analysis has indeed discovered linguistic patterns.

We tested whether structure was detectable in the same manner in a generative model, in particular a version of openai[9] that uses the same embedding dimensions (768), such that the same system architecture with the same hyperparameters can be used. Figure 6 shows the tSNE projection of the latent vectors obtained when using as sentence representation the last token embedding, and F1 averages over 3 runs for choosing the correct answer. We note the same type of clustering as for the encoder models.



avg F1 = 0.9003 (std = 0.0058) avg F1 = 0.9807 (std = 0.0021)

Figure 6: tSNE projection of latent vectors of the data represented with last token embeddings of an openai model, and score averages over 3 runs.

Nastase et al. (2024a) have shown, through experiments on several languages, that sentence embeddings do not encode chunk structure as an abstraction, but rather as linguistic clues – such as phrase boundaries and number information – that can be assembled into the chunk structure. This observation and the results and plots in Figure 5 indicate that the S_{AVG} and S_{CLS} encode the information about phrase boundaries, number and even semantic roles in the same manner and in the same location for BERT and Electra in particular.

We were not concerned here with fine-tuned embeddings, specifically because fine-tuning – most commonly using textual similarity datasets and the cosine metric – distorts the embedding space. Furthermore, Nastase & Merlo (2024) have compared the [CLS] sentence representation from Electra and BERT with several fine-tuned models, and have shown that fine-tuning for semantic similarity leads to a loss of syntactic structural information.

6 DISCUSSION AND CONCLUSIONS

We have shown that the perspective on the embedding spaces of several encoder transformer models obtained through the cosine measure is distorted, as the usual expectations of an embedding space are not met. What we found instead is a complex picture, in which orthogonal representations lead to very similar results on many tasks. Experiments in detecting a sentence's chunk structure – the sequence of NP/VP/PP phrases and their grammatical number attributes, or the semantic role structure in a sentence – have shown that information relevant for reconstructing these types of structure is encoded in the same manner in the superficially very different representations, as a system trained on one sentence representation has a very similar performance when tested on the others.

The experiments presented in this paper add to the complex picture of embedding spaces. To build better models – that generalize well, including across languages; that represent entities, their properties and relationships; that are smaller and more efficient – we need to understand the representations they induce, and where the strengths and weaknesses are. Quantifying the type of information encoded in embeddings is also useful for finetuning, where understanding what us gained or lost ca help assess finetuned representations beyond the tested tasks. Detecting structure in sentence representations can contribute towards bridging the gap between connectionist models and symbolic approaches to language modelling.

[9] https://huggingface.co/openai-community/openai-gpt

AI USE STATEMENT

AI was not used in any stage of the reported work or in writing this paper.

ETHICS STATEMENT

The current study uses selections from previously published data that highlights linguistic phenomena, and does not pose ethical issues in terms of cultural biases of any sort. The work is concerned with exploring linguistic issues in sentence embeddings in pretrained language models. This raises no ethical issues from a topic or result point of view. Ecologically, our studies are not big energy consumers. The experiments were run on an HP PAIR Workstation Z4 G5 MT, with 12 x Intel Xeon(R) w3-2425 processors, 64G RAM, and an Nvidia GeForce RTX 4090 24G G6X GPU.

REFERENCES

Maria A. Rodriguez and Paola Merlo. Word associations and the distance properties of context-aware word embeddings. In Raquel Fernández and Tal Linzen (eds.), *Proceedings of the 24th Conference on Computational Natural Language Learning*, pp. 376–385, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.30. URL https://aclanthology.org/2020.conll-1.30/.

Aixiu An, Chunyang Jiang, Maria A. Rodriguez, Vivi Nastase, and Paola Merlo. BLM-AgrF: A new French benchmark to investigate generalization of agreement in neural networks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1363–1374, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.99.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. ParaCrawl: Web-scale acquisition of parallel corpora. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4555–4567, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.417. URL https://aclanthology.org/2020.acl-main.417/.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. Generating sentences from disentangled syntactic and semantic spaces. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6008–6019, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1602. URL https://aclanthology.org/P19-1602/.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. Too much in common: Shifting of embeddings in transformer language models and its implications. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5117–5130, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.403. URL https://aclanthology.org/2021.naacl-main.403/.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=xYGNO86OWDH.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. A multi-task approach for disentangling syntax and semantics in sentence representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 2453–2464, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1254. URL https://aclanthology.org/N19-1254/.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. Finding universal grammatical relations in multilingual BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5564–5577, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.493. URL https://aclanthology.org/2020.acl-main.493.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*, 2020. URL https://openreview.net/pdf?id=r1xMH1BtvB.

Simone Conia and Roberto Navigli. Probing for predicate argument structures in pretrained language models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4622–4632, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.316. URL https://aclanthology.org/2022.acl-long.316/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL www.aclweb.org/anthology/N19-1423.

Moussa Koulako Bala Doumbouya, Dan Jurafsky, and Christopher D. Manning. Tversky neural networks: Psychologically plausible deep learning with differentiable tversky similarity, 2025. URL https://arxiv.org/abs/2506.11035.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition, 2022. URL https://arxiv.org/abs/2209.10652.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide,

Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. doi: 10.48550/arXiv.2502.13595. URL https://arxiv.org/abs/2502.13595.

Josh Engels, Eric Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, volume 2025, pp. 84591–84622, 2025.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. Representation degeneration problem in training natural language generation models. In *7th International Conference on Learning Represen- tations (ICLR 2019)*, 2019. doi: 10.48550/arXiv.1907.12009.

Nathan Godey, Éric Clergerie, and Benoît Sagot. Anisotropy is inherent to self-attention in transformers. In Yvette Graham and Matthew Purver (eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35–48, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.3. URL https://aclanthology.org/2024.eacl-long.3/.

Wes Gurnee and Max Tegmark. Language models represent space and time. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 2483–2503, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0a6059857ae5c82ea9726ee9282a7145-Paper-Conference.pdf.

Tomomasa Hara, Hiroto Kurita, Masaaki Imaizumi, Kentaro Inui, and Sho Yokoi. Why mean pooling works: Quantifying second-order collapse in text embeddings. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 47180–47201, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2183. URL https://aclanthology.org/2026.acl-long.2183/.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced BERT with disentangled attention. 2021. URL https://openreview.net/forum?id=XPZIaotutsD.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. WhiteningBERT: An easy unsupervised sentence embedding approach. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 238–244, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.23. URL https://aclanthology.org/2021.findings-emnlp.23/.

Daphne Ippolito, David Grangier, Douglas Eck, and Chris Callison-Burch. Toward better storylines with sentence-level language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7472–7478, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.666. URL https://aclanthology.org/2020.acl-main.666/.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, November 2019. ISSN 1935-8237. doi: 10.1561/2200000056. URL https://doi.org/10.1561/2200000056.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019. URL http://arxiv.org/abs/1909.11942.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9119–9130, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.733. URL https://aclanthology.org/2020.emnlp-main.733/.

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

Anemily Machina and Robert Mercer. Anisotropy is not inherent to transformers. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4892–4907, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.274. URL https://aclanthology.org/2024.naacl-long.274/.

Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117:30046 – 30054, 2020.

Giangiacomo Mercatali and André Freitas. Disentangling generative factors in natural language with discrete variational autoencoders. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 3547–3556, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.301. URL https://aclanthology.org/2021.findings-emnlp.301/.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff (eds.), *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics. URL https://aclanthology.org/N13-1090/.

David Mimno and Laure Thompson. The strange geometry of skip-gram with negative sampling. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2873–2878, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1308. URL https://aclanthology.org/D17-1308/.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL https://arxiv.org/abs/2210.07316.

Vivi Nastase and Paola Merlo. Grammatical information in BERT sentence embeddings as two-dimensional arrays. In Burcu Can, Maximilian Mozes, Samuel Cahyawijaya, Naomi Saphra, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Chen Zhao, Isabelle Augenstein, Anna Rogers, Kyunghyun Cho, Edward Grefenstette, and Lena Voita (eds.), *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pp. 22–39, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.repl4nlp-1.3. URL https://aclanthology.org/2023.repl4nlp-1.3.

Vivi Nastase and Paola Merlo. Are there identifiable structural parts in the sentence embedding whole? In Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (eds.), *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 23–42, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.blackboxnlp-1.3. URL https://aclanthology.org/2024.blackboxnlp-1.3/.

Vivi Nastase, Chunyang Jiang, Giuseppe Samo, and Paola Merlo. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In *Tenth Italian Conference on Computational Linguistics*, 2024a.

Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, and Rachele Sprugnoli (eds.), *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pp. 631–643, Pisa, Italy, December 2024b. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL https://aclanthology.org/2024.clicit-1.71/.

Dmitry Nikolaev and Sebastian Padó. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. In Yonatan Belinkov, Sophie Hao, Jaap Jumelet, Najoung Kim, Arya McCarthy, and Hosein Mohebbi (eds.), *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 142–154, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.blackboxnlp-1.11. URL https://aclanthology.org/2023.blackboxnlp-1.11.

Dmitry Nikolaev and Sebastian Padó. Representation biases in sentence transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3701–3716, Dubrovnik, Croatia, May 2023b. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.268.

Dmitry Nikolaev and Sebastian Padó. The universe of utterances according to BERT. In Maxime Amblard and Ellen Breitholtz (eds.), *Proceedings of the 15th International Conference on Computational Semantics*, pp. 99–105, Nancy, France, June 2023c. Association for Computational Linguistics. URL https://aclanthology.org/2023.iwcs-1.12.

Juri Opitz and Anette Frank. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang (eds.), *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 625–638, Online only, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.aacl-main.48.

Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. Deep subjecthood: Higher-order grammatical features in multilingual BERT. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2522–2532, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.215. URL https://aclanthology.org/2021.eacl-main.215/.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *Workshop on Theoretical Foundations of Foundation Models (TF2M) at ICML 2024*, 2024.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019a. URL https://arxiv.org/abs/1908.10084.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong,

China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Giuseppe Samo, Vivi Nastase, Chunyang Jiang, and Paola Merlo. BLM-s/lE: A structured dataset of English spray-load verb alternations for testing generalization in LLMs. In *Findings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Danilo Silva De Carvalho, Giangiacomo Mercatali, Yingji Zhang, and André Freitas. Learning disentangled representations for natural language definitions. In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1371–1384, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.101. URL https://aclanthology.org/2023.findings-eacl.101.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33: 16857–16867, 2020.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval, 2021. URL https://arxiv.org/abs/2103.15316.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. What do you learn from context? probing for sentence structure in contextualized word representations. In *The Seventh International Conference on Learning Representations (ICLR)*, pp. 235–249, 2019.

William Timkey and Marten van Schijndel. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4527–4546, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.372. URL https://aclanthology.org/2021.emnlp-main.372/.

Amos Tversky. Features of similarity. *Psychological Review*, 84:327–352, 1977. URL https://api.semanticscholar.org/CorpusID:9173202.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. Holmes: A benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647, 2024. doi: 10.1162/tacl_a_00718. URL https://aclanthology.org/2024.tacl-1.88/.

Yuxia Wang, Shimin Tao, Ning Xie, Hao Yang, Timothy Baldwin, and Karin Verspoor. Collective human opinions in semantic textual similarity. *Transactions of the Association for Computational Linguistics*, 11:997–1013, 2023. doi: 10.1162/tacl_a_00584. URL https://aclanthology.org/2023.tacl-1.56/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Dezhi Ye, Junwei Hu, Jiabin Fan, Bowen Tian, Jie Liu, Haijin Liang, and Jin Ma. Best practices for distilling large language models into BERT for web search ranking. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Kareem Darwish,

and Apoorv Agarwal (eds.), *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pp. 128–135, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-industry.11/.

Yujie Zhang, Weikang Yuan, and Zhuoren Jiang. Bridging intuitive associations and deliberate recall: Empowering LLM personal assistant with graph-structured long-term memory. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 17533–17547, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.901. URL https://aclanthology.org/2025.findings-acl.901/.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. A survey of generative information extraction. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4840–4870, Abu Dhabi, UAE, January 2025b. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.324/.

A WORDS VS. TOKEN EMBEDDING SIMILARITIES DISTRIBUTION

Figure 7 shows a comparison between the distribution of token and word similarities within the same sentence. A tighter distribution – as displayed by RoBERTa embeddings – indicates that all contextual embeddings are closer to each other, and thus encode more contextual information. BERT and Electra embeddings display distributions with larger standard deviation, indicating that there is more variation in the information encoded in the individual tokens and words. Electra token/word similarities have a higher mean, indicating that these embeddings encode more contextual information than BERT ones. All distributions have a high spike close to 0. These pairs include punctuation and "suffix" tokens.

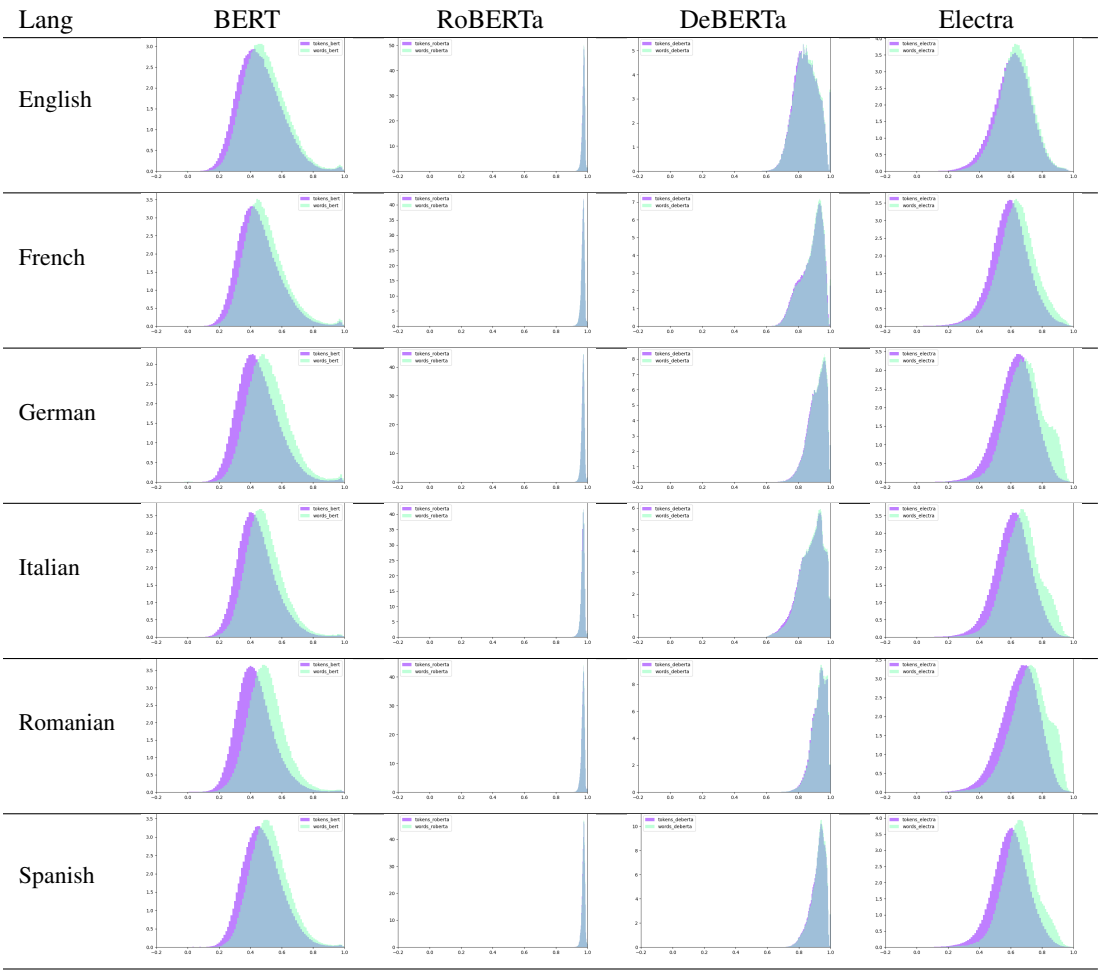


Figure 7: Cosine similarities histograms computed for words and tokens from 1000 English/French/German/Italian/Romanian/Spanish. **The y axes have different scales across the plots, but the x-axis has the same range [0,1] across the plots.**

B MULTILINGUAL SENTENCE REPRESENTATIONS COMPARISON

B.1 COSINE SIMILARITY HISTOGRAMS

The plots show similar tendencies across the 6 covered languages.

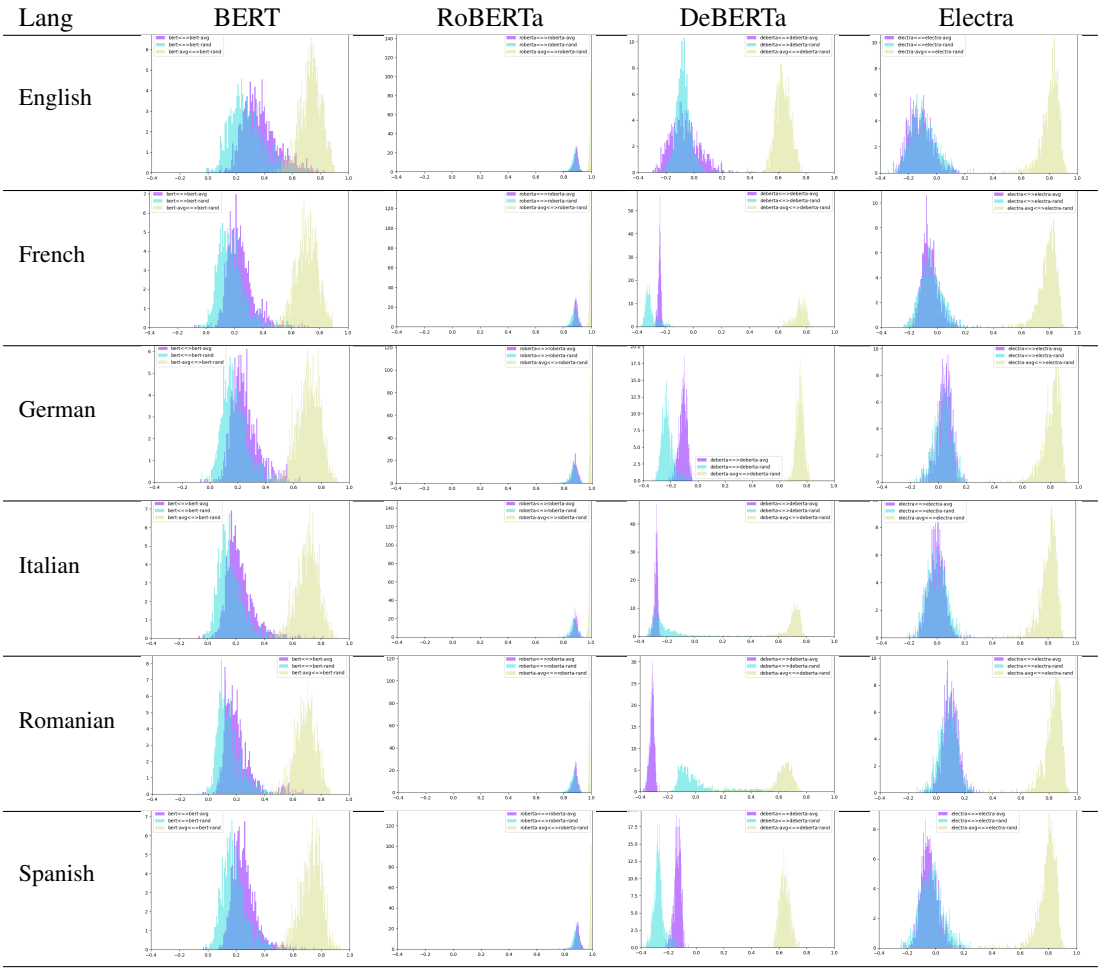


Figure 8: **Cosine similarities** histograms for 1000 English/French/German/Italian/Romanian/Spanish sentences v. In yellow are the similarities between S_{AVG} and $S_{T_{rand}}$, in blue the similarities between S_{CLS} and $S_{T_{rand}}$, and in purple the similarities between S_{CLS} and S_{AVG}. **The y axes have different scales across the plots, but the x-axis has the same range [0,1] across the plots.**

B.2 COSINE SIMILARITY VS. SENTENCE LENGTH IN TOKENS

The y-axis is the length in tokens of each sentence, and on the x-axis there are three corresponding points for each sentence, for the cosine similarity between each pairing of sentence representations. We include the Spearman correlation coefficients as (S_{CLS}-$S_{T_{rand}}$, S_{CLS}-S_{AVG}, S_{AVG}-$S_{T_{rand}}$) tuples. We do not include the p-values in the table. For the situations when the p-values are $\geq 10^{-4}$, the absolute value of the correlation score is below 0.1 .

The plots and the included Spearman correlations show weak correlation between the sentence length and the cosine similarity in a very few cases: BERT on English data, for the (S_{AVG},$S_{T_{rand}}$) representation pair, DeBERTa on French, (stronger on) German and Spanish data, for the (S_{CLS}, $S_{T_{rand}}$) representation pair.

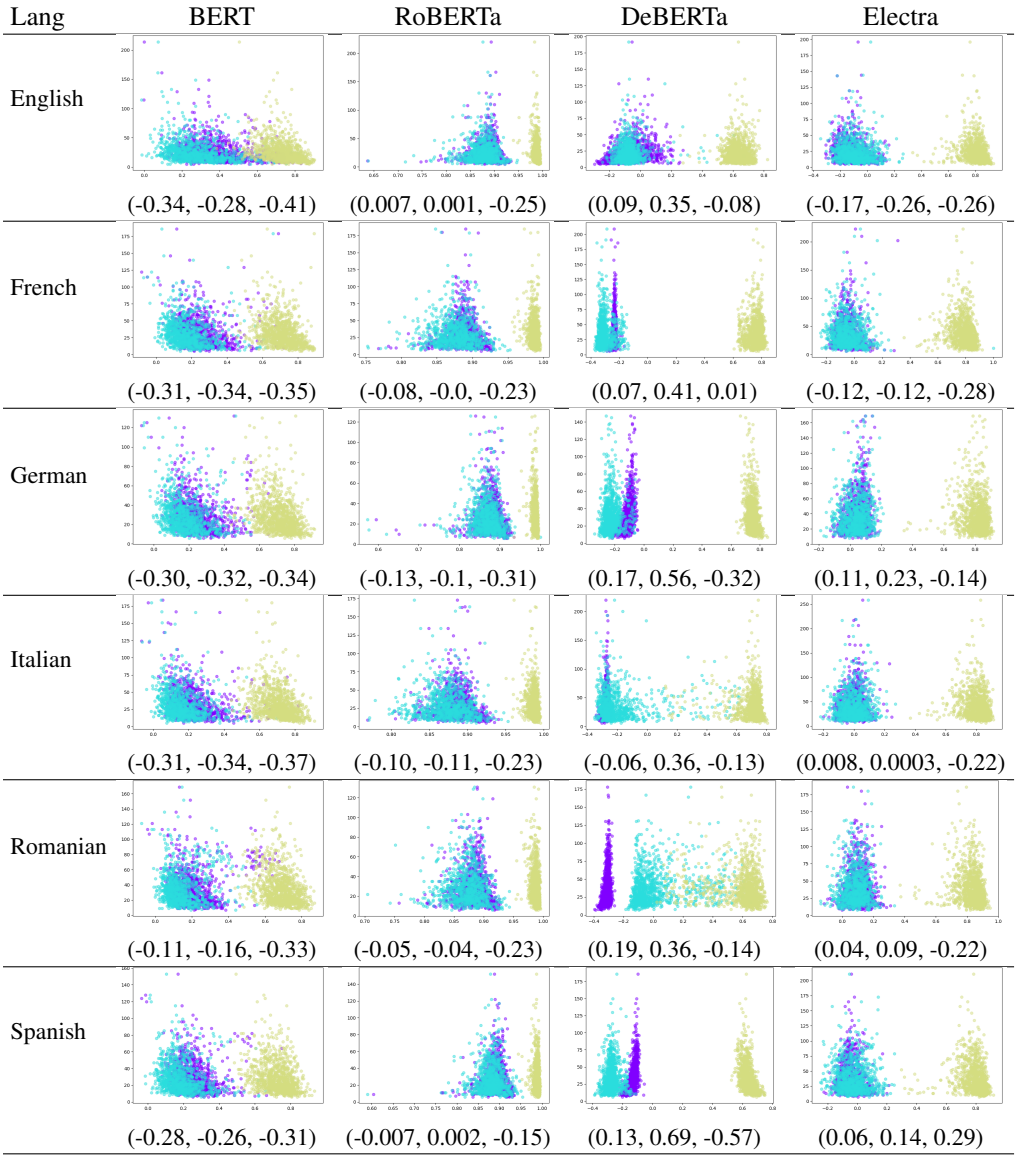


Figure 9: **Cosine similarities** (on the x-axis) plots for 1000 English/French/German/Italian/Romanian/Spanish sentences vs. the length in tokens of the sentences (on the y-axis). In yellow are the similarities between S_{AVG} and $S_{T_{rand}}$, in blue the similarities between S_{CLS} and $S_{T_{rand}}$, and in purple the similarities between S_{CLS} and S_{AVG}. The Spearman correlation coefficients are shown as (S_{CLS}-$S_{T_{rand}}$, S_{CLS}-S_{AVG}, S_{AVG}-$S_{T_{rand}}$) tuples. **The x and y axes have different scales across the plots.**

B.3 COSINE ANALYSIS WITHIN THE AVERAGE AND CLS SUBSPACES

The issue we explore is whether a comparison between vectors that considers each dimension separately is appropriate for objects as complex as embeddings produced by pretrained transformer models. As the averaged embedings and the [CLS] token embeddings seem to inhabit different subspaces, the question of whether the two subspaces have a "parallel" structure w.r.t. the cosine has come up. The plots below show the comparison of the cosine similarity within the two subspaces, and with human judgements. We used the STSB dataset[10]. As there is no training involved, we include the plots for all three subsets (train, validation, test). The plots are ordered by the similarities in the averaged embedding subspace. All plots show that the cosine does not reflect the wide range of similarity scores in the human annotations.

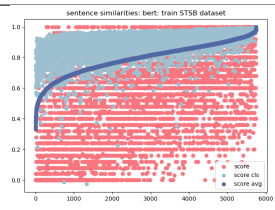
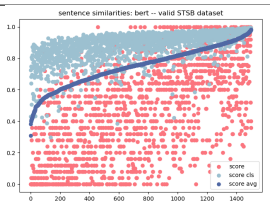
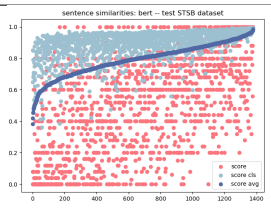
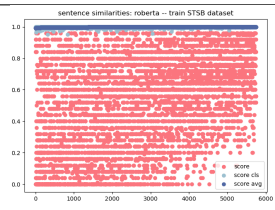
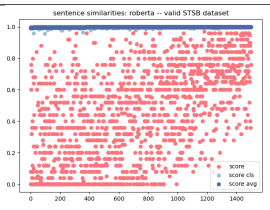
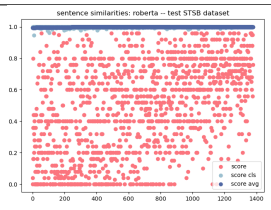
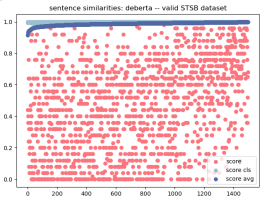
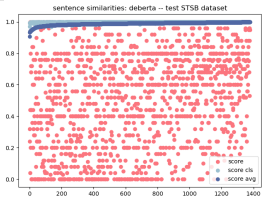
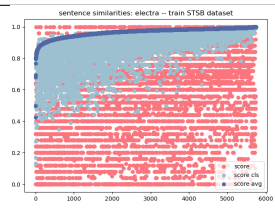
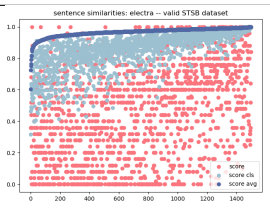
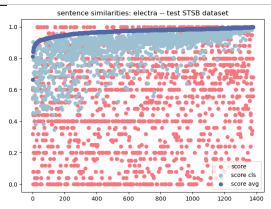
model	train	valid	test
BERT			
spearman	0.17 / 0.52 / 0.60	0.34 / 0.62 / 0.70	0.19 / 0.53 / 0.58
pearson	0.14 / 0.55 / 0.52	0.31 / 0.62 / 0.63	0.15 / 0.54 / 0.52
kendall	0.12 / 0.37 / 0.43	0.23 / 0.45 / 0.51	0.13 / 0.38 / 0.41
RoBERTa			
spearman	0.25 / 0.42 / 0.33	0.34 / 0.56 / 0.45	0.25 / 0.45 / 0.42
pearson	0.09 / 0.39 / 0.17	0.15 / 0.50 / 0.25	0.06 / 0.40 / 0.23
kendall	0.17 / 0.29 / 0.22	0.23 / 0.40 / 0.31	0.18 / 0.32 / 0.29
DeBERTa			
spearman	0.27 / 0.25 / 0.81	0.32 / 0.41 / 0.72	0.23 / 0.29 / 0.70
pearson	0.21 / 0.17 / 0.70	0.22 / 0.30 / 0.53	0.15 / 0.22 / 0.50
kendall	0.18 / 0.17 / 0.62	0.22 / 0.28 / 0.52	0.16 / 0.20 / 0.51
Electra			
spearman	0.24 / 0.27 / 0.82	0.37 / 0.40 / 0.83	0.20 / 0.23 / 0.80
pearson	0.20 / 0.25 / 0.70	0.30 / 0.29 / 0.70	0.13 / 0.18 / 0.67
kendall	0.17 / 0.19 / 0.63	0.26 / 0.28 / 0.64	0.14 / 0.15 / 0.61

Table 1: Cosine similarity comparisons between human annotations and similarities within the [CLS] and averaged token subspaces on the STSB data. Correlation scores are:
corr(human, S_{CLS}
/ corr(human, S_{AVG}) / corr(S_{CLS}, S_{AVG})

[10] https://huggingface.co/datasets/sentence-transformers/stsb

B.4 EUCLID SIMILARITY HISTOGRAMS

Euclid similarities are computed as (1/(euclid$_{dist(x,y)}$ + 1).

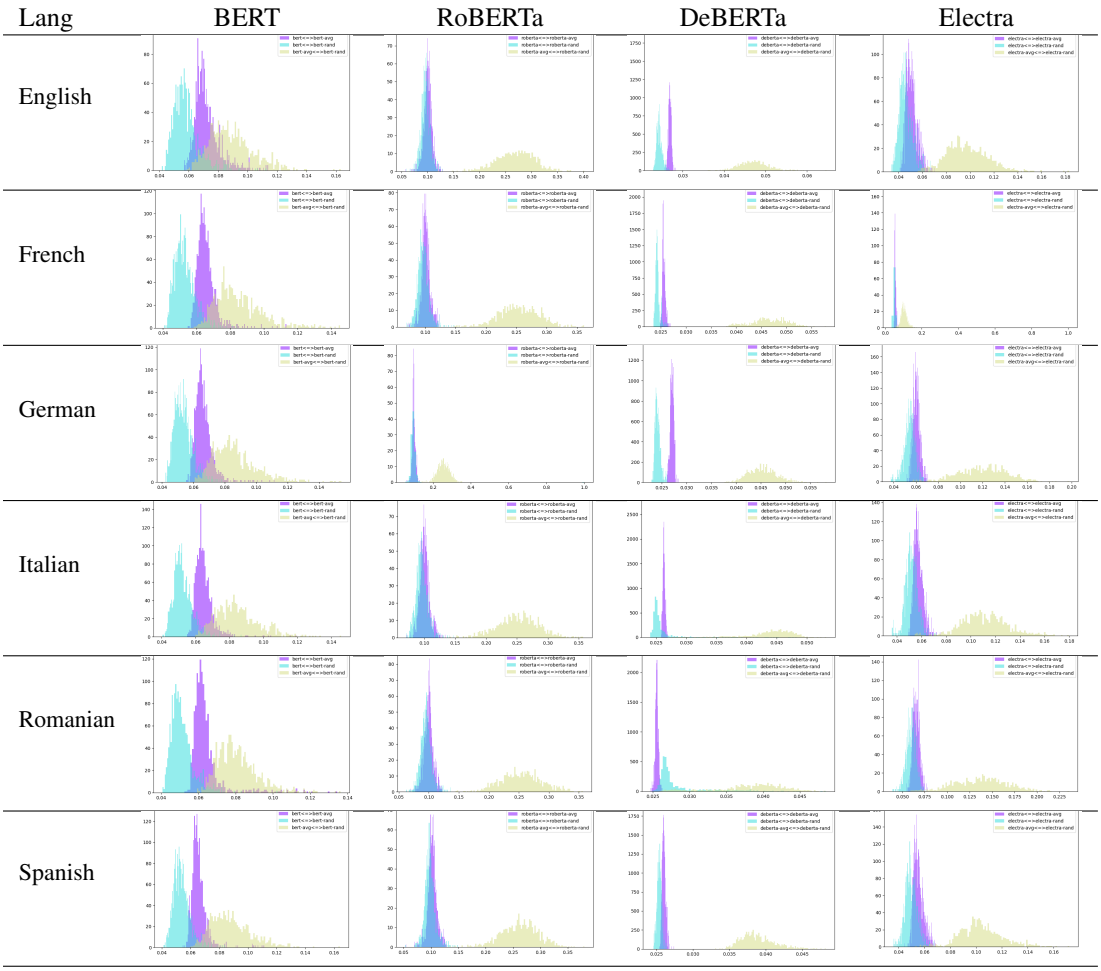


Figure 10: **Euclid similarities** (1/(euclid$_{dist(x,y)}$ + 1) histograms for 1000 English/ French/ German/ Italian/ Romanian/ Spanish sentences. In yellow are the distances between S$_{AVG}$ and S$_{T_{rand}}$, in blue the distances between S$_{CLS}$ and S$_{T_{rand}}$, and in purple the distances between S$_{CLS}$ and S$_{AVG}$. **Both x and y axes have different scales across the plots.**

C TASK RESULTS

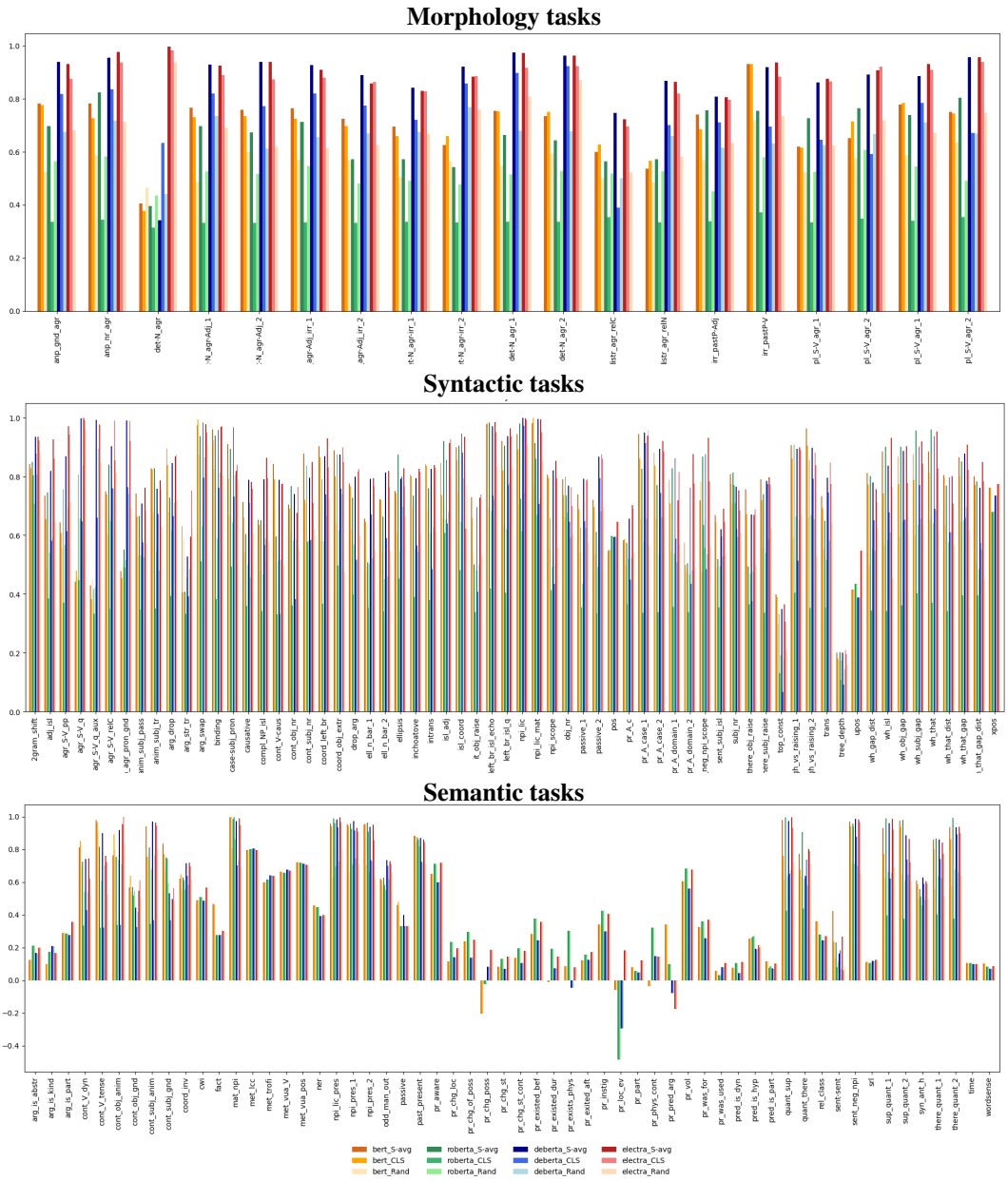


Figure 11: Detailed results on the FlashHolmes benchmark, on morphology, syntax and semantic tasks

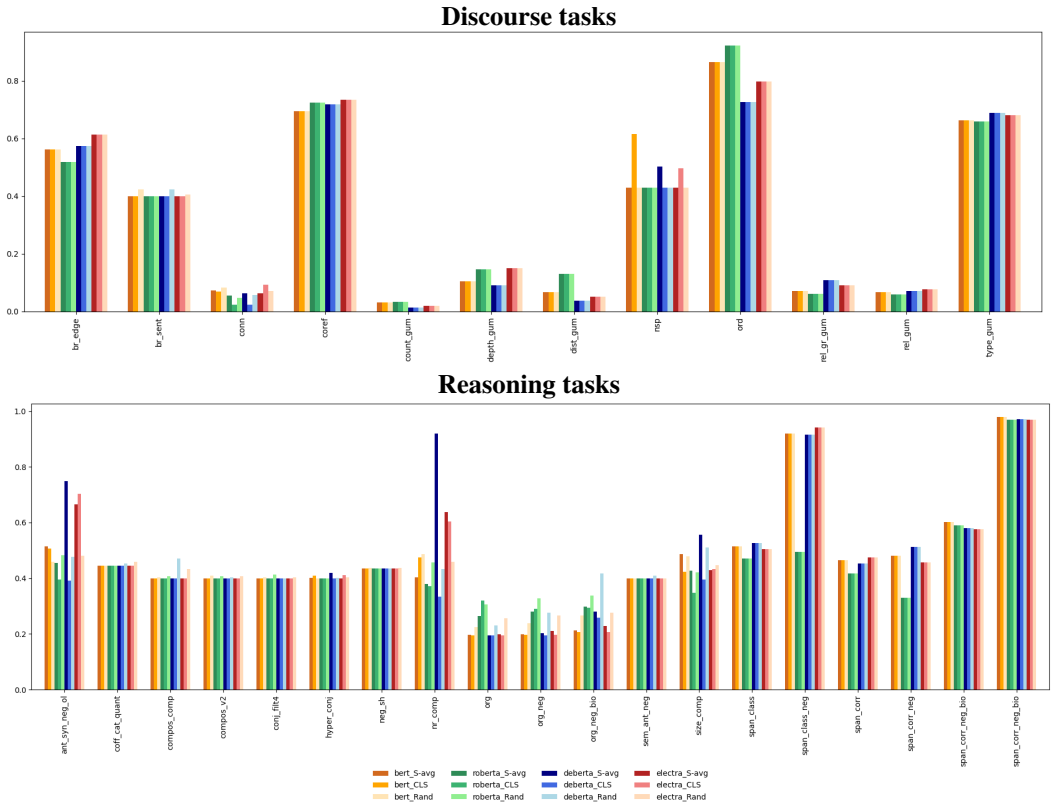


Figure 12: Detailed results on the FlashHolmes benchmark, on discourse and reasoning tasks

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-anaphor_gender_agreement	0.782	0.776	0.525	0.696	0.336	0.564	**0.939**	0.818	0.676	0.931	0.875	0.682
blimp-anaphor_number_agreement	0.782	0.727	0.586	0.825	0.345	0.582	0.954	0.835	0.717	**0.977**	0.937	0.712
blimp-determiner_noun_agreement_1	0.755	0.752	0.549	0.664	0.336	0.514	**0.974**	0.898	0.678	0.973	0.918	0.809
blimp-determiner_noun_agreement_2	0.736	0.75	0.594	0.644	0.336	0.529	**0.963**	0.922	0.677	**0.963**	0.924	0.873
blimp-determiner_noun_agreement_irregular_1	0.694	0.66	0.505	0.572	0.336	0.49	**0.842**	0.721	0.675	0.83	0.829	0.668
blimp-determiner_noun_agreement_irregular_2	0.626	0.66	0.565	0.542	0.334	0.476	**0.922**	0.857	0.768	0.884	0.885	0.758
blimp-determiner_noun_agreement_with_adj_2	0.759	0.735	0.601	0.673	0.333	0.517	0.938	0.773	0.611	**0.939**	0.873	0.620
blimp-determiner_noun_agreement_with_adj_irregular_1	0.765	0.725	0.57	0.713	0.334	0.546	**0.927**	0.82	0.656	0.909	0.879	0.613
blimp-determiner_noun_agreement_with_adj_irregular_2	0.725	0.698	0.568	0.573	0.333	0.481	**0.889**	0.775	0.670	0.857	0.864	0.627
blimp-determiner_noun_agreement_with_adjective_1	0.767	0.73	0.487	0.697	0.333	0.528	**0.929**	0.82	0.736	0.925	0.89	0.692
blimp-distractor_agreement_relational_noun	0.537	0.567	0.485	0.573	0.334	0.526	**0.868**	0.702	0.659	0.864	0.819	0.582
blimp-distractor_agreement_relative_clause	0.6	0.628	0.501	0.565	0.354	0.518	**0.746**	0.39	0.501	0.723	0.696	0.524
blimp-irregular_past_participle_adjectives	0.742	0.685	0.569	0.757	0.338	0.451	**0.808**	0.712	0.616	0.806	0.796	0.633
blimp-irregular_past_participle_verbs	0.932	0.931	0.718	0.754	0.372	0.58	0.92	0.695	0.633	**0.936**	0.884	0.736
blimp-irregular_plural_subject_verb_agreement_1	0.619	0.615	0.523	0.727	0.334	0.525	0.861	0.647	0.626	**0.875**	0.866	0.624
blimp-irregular_plural_subject_verb_agreement_2	0.651	0.716	0.577	0.764	0.349	0.608	0.892	0.591	0.667	0.908	**0.922**	0.720
blimp-regular_plural_subject_verb_agreement_1	0.779	0.785	0.588	0.739	0.341	0.544	0.885	0.785	0.711	**0.93**	0.909	0.672
blimp-regular_plural_subject_verb_agreement_2	0.75	0.746	0.637	0.804	0.354	0.49	**0.957**	0.671	0.670	**0.957**	0.938	0.747
zorro-agreement_determiner_noun-between_neighbors	0.407	0.378	0.466	0.395	0.315	0.435	0.343	0.633	0.442	**0.996**	0.983	0.937
average	0.706	0.698	0.559	0.667	0.339	0.521	0.871	0.740	0.652	**0.904**	0.878	0.696

Table 2: FlashHolmes morphology tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-adjunct_island	0.735	0.655	0.487	0.746	0.384	0.539	0.818	0.581	0.592	**0.925**	0.861	0.646
blimp-animate_subject_passive	0.743	0.663	0.531	0.665	0.346	0.531	0.709	0.577	0.527	**0.761**	0.685	0.573
blimp-animate_subject_trans	0.827	0.824	0.678	**0.828**	0.351	0.546	0.759	0.673	0.479	0.787	0.755	0.628
blimp-causative	0.714	0.661	0.544	0.605	0.36	0.501	**0.788**	0.71	0.456	0.782	0.758	0.605
blimp-complex_NP_island	0.652	0.634	0.478	0.65	0.341	0.467	0.791	0.566	0.589	**0.863**	0.794	0.581

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-coord_str_constr_complex_left_branch	0.902	0.865	0.58	0.79	0.367	0.58	0.87	0.739	0.665	**0.929**	0.831	0.724
blimp-coord_str_constr_object_extraction	0.888	0.801	0.637	0.874	0.498	0.617	0.875	0.759	0.739	**0.899**	0.85	0.476
blimp-drop_argument	0.775	0.766	0.621	0.729	0.4	0.569	0.8	0.516	0.516	0.817	**0.827**	0.599
blimp-ellipsis_n_bar_1	0.656	0.645	0.492	0.508	0.352	0.503	0.792	0.67	0.522	**0.794**	0.767	0.548
blimp-ellipsis_n_bar_2	0.723	0.72	0.543	0.664	0.34	0.45	0.813	0.589	0.46	**0.818**	0.763	0.547
blimp-existential_there_object_raising	**0.757**	0.671	0.5	0.494	0.364	0.469	0.67	0.375	0.478	0.669	0.688	0.497
blimp-existential_there_subject_raising	0.791	0.718	0.558	0.74	0.336	0.54	0.785	0.774	0.587	**0.797**	0.772	0.624
blimp-expletive_it_object_raising	0.729	0.659	0.544	0.5	0.337	0.48	0.695	0.409	0.499	0.729	**0.739**	0.532
blimp-inchoative	0.806	0.801	0.698	0.736	0.39	0.552	0.793	0.565	0.542	**0.825**	0.815	0.604
blimp-intransitive	**0.841**	0.837	0.685	0.76	0.379	0.606	0.826	0.485	0.629	0.838	0.829	0.615
blimp-left_branch_island_echo_question	0.978	0.984	0.692	0.983	0.418	0.683	0.97	0.735	0.722	**0.985**	0.95	0.835
blimp-left_branch_island_simple_question	0.921	0.821	0.63	0.904	0.405	0.62	0.937	0.771	0.777	**0.964**	0.937	0.828
blimp-only_npi_scope	0.807	0.796	0.55	0.658	0.413	0.494	0.822	0.435	0.498	**0.853**	0.795	0.559
blimp-passive_1	0.738	0.689	0.541	0.625	0.355	0.434	**0.791**	0.648	0.624	0.787	**0.791**	0.576
blimp-passive_2	0.72	0.696	0.534	0.611	0.335	0.492	0.867	0.793	0.681	**0.876**	0.859	0.691
blimp-principle_A_c_command	0.583	0.586	0.5	0.574	0.365	0.52	0.657	0.448	0.525	**0.702**	0.69	0.541
blimp-principle_A_case_1	0.944	0.86	0.654	0.826	0.336	0.577	0.949	0.913	0.657	0.939	**0.956**	0.563
blimp-principle_A_case_2	0.882	0.836	0.657	0.771	0.338	0.532	0.894	0.744	0.662	**0.921**	0.886	0.657
blimp-principle_A_domain_1	0.788	0.708	0.544	0.828	0.356	0.536	**0.862**	0.588	0.507	0.719	0.816	0.567
blimp-principle_A_domain_2	0.575	0.499	0.478	0.505	0.339	0.468	0.762	0.434	0.479	**0.877**	0.775	0.49
blimp-principle_A_domain_3	0.64	0.649	0.495	0.524	0.335	0.493	0.839	0.533	0.515	**0.877**	0.82	0.548
blimp-principle_A_reconstruction	0.948	0.88	0.58	0.752	0.333	0.557	0.808	0.467	0.584	**0.949**	0.599	0.585
blimp-sentential_negation_npi_scope	0.719	0.781	0.477	0.867	0.634	0.556	0.876	0.484	0.534	**0.931**	0.783	0.561
blimp-sentential_subject_island	0.669	0.645	0.49	0.519	0.354	0.493	0.62	0.595	0.527	**0.69**	0.645	0.534
blimp-tough_vs_raising_1	0.907	0.86	0.592	**0.908**	0.404	0.663	0.893	0.513	0.651	0.899	0.891	0.671
blimp-tough_vs_raising_2	**0.964**	0.904	0.661	0.855	0.353	0.669	0.898	0.792	0.655	0.882	0.839	0.526
blimp-transitive	0.732	0.693	0.553	0.65	0.354	0.516	0.796	0.746	0.581	**0.847**	0.774	0.641
blimp-wh_island	0.886	0.742	0.563	0.901	0.342	0.583	0.836	0.679	0.609	**0.931**	0.766	0.452
blimp-wh_questions_object_gap	0.867	0.774	0.593	0.901	0.361	0.645	0.888	0.653	0.653	**0.902**	0.779	0.607
blimp-wh_questions_subj_gap	0.876	0.789	0.6	**0.957**	0.402	0.635	0.9	0.771	0.657	0.919	0.844	0.63
blimp-wh_questions_subj_gap_long_dist	**0.812**	0.772	0.498	0.802	0.344	0.555	0.78	0.651	0.548	0.757	0.714	0.625
blimp-wh_vs_that_no_gap	0.885	0.813	0.649	**0.959**	0.371	0.641	0.936	0.689	0.685	0.952	0.827	0.615
blimp-wh_vs_that_no_gap_long_dist	**0.805**	0.796	0.634	0.769	0.341	0.576	0.797	0.609	0.612	0.801	0.708	0.565

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
blimp-wh_vs_that_with_gap	0.865	0.802	0.645	0.851	0.395	0.651	0.878	0.66	0.696	**0.908**	0.822	0.613
blimp-wh_vs_that_with_gap_long_dist	0.8	0.771	0.511	0.784	0.394	0.486	0.762	0.574	0.55	**0.849**	0.783	0.449
const	0.163	0.163	0.164	0.128	0.128	0.128	**0.186**	**0.186**	**0.186**	0.18	0.18	0.18
const_max_depth		0.815	0.668		0.78	0.658		0.812	0.539		**0.829**	0.635
const_node_length		0.879	0.721		0.89	0.73		**0.903**	0.588		0.871	0.676
context-object_number	0.705	0.691	0.609	**0.767**	0.361	0.581	0.74	0.383	0.586	0.676	0.763	0.548
context-subj_number	**0.878**	0.72	0.707	0.837	0.578	0.582	0.795	0.582	0.588	0.849	0.71	0.619
context-verb_causative	**0.842**	0.791	0.687	0.596	0.332	0.511	0.789	0.333	0.517	0.775	0.777	0.498
flesch		0.171	0.019		0.605	0.082		**0.846**	0.002		0.48	-0.074
pos	0.547	0.551	0.548	0.598	0.597	0.598	0.594	0.595	0.594	**0.646**	0.645	0.645
senteval-bigram_shift	0.841	0.828	0.786	0.85	0.804	0.706	**0.936**	0.878	0.806	0.935	0.922	0.88
senteval-obj_number	0.791	0.737	0.66	**0.799**	0.735	0.678	0.769	0.645	0.592	0.763	0.698	0.599
senteval-sentence_length	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**	**0.048**
senteval-subj_number	0.808	0.774	0.707	**0.813**	0.771	0.71	0.766	0.62	0.593	0.752	0.685	0.608
senteval-top_constituents	**0.398**	0.391	0.298	0.334	0.131	0.191	0.349	0.067	0.177	0.365	0.307	0.209
senteval-tree_depth	0.2	0.181	0.174	0.203	0.108	0.173	0.2	0.091	0.145	**0.209**	0.196	0.156
upos	0.415	0.415	0.415	0.434	0.434	0.434	0.388	0.388	0.388	**0.548**	**0.548**	**0.548**
xpos	0.761	0.761	0.761	0.68	0.68	0.68	0.736	0.735	0.736	**0.775**	**0.775**	**0.775**
zorro-agr_subj_v-across_prepositional_phrase	0.643	0.608	0.556	0.757	0.371	0.567	0.868	0.614	0.694	**0.971**	0.943	0.714
zorro-agr_subj_v-across_relative_clause	0.75	0.738	0.6	0.84	0.348	0.647	0.904	0.76	0.758	**0.99**	0.855	0.813
zorro-agr_subj_v-in_question_with_aux	0.429	0.383	0.447	0.417	0.333	0.423	**0.992**	0.659	0.894	0.978	0.975	0.847
zorro-agr_subj_v-in_simple_question	0.441	0.48	0.457	0.806	0.447	0.656	0.998	0.648	0.836	**1.0**	0.991	0.96
zorro-anaphor_agr-pronoun_gender	0.477	0.454	0.494	0.551	0.49	0.516	**0.991**	0.764	0.74	0.989	0.921	0.691
zorro-arg_str-dropped_argument	**0.896**	0.838	0.68	0.728	0.393	0.615	0.845	0.665	0.71	0.869	0.875	0.719
zorro-arg_str-swapped_arguments	0.974	**0.994**	0.874	0.936	0.512	0.63	0.983	0.795	0.865	0.978	0.95	0.906
zorro-arg_str-transitive	0.63	0.405	0.59	0.409	0.333	0.455	0.527	0.392	0.484	0.596	**0.754**	0.63
zorro-binding-principle_a	0.96	0.923	0.827	0.939	0.383	0.588	0.958	0.762	0.813	0.968	**0.97**	0.804
zorro-case-subjective_pronoun	0.91	0.793	0.669	0.895	0.493	0.643	**0.967**	0.731	0.805	0.819	0.841	0.787
zorro-ellipsis-n_bar	0.751	0.744	0.608	**0.874**	0.453	0.545	0.792	0.799	0.696	0.829	0.767	0.663
zorro-filler-gap-wh_question_object	0.959	0.848	0.769	**0.997**	0.785	0.774	0.909	0.781	0.743	0.967	0.915	0.756
zorro-filler-gap-wh_question_subject	0.99	0.904	0.827	**1.0**	0.902	0.673	0.942	0.797	0.812	0.912	0.933	0.786
zorro-island-effects-adjunct_island	0.846	0.738	0.659	0.921	0.607	0.654	0.857	0.641	0.679	0.914	**0.926**	0.799
zorro-island-effects-coord_str_constr	0.899	0.855	0.746	0.904	0.482	0.646	**0.946**	0.882	0.795	0.936	0.623	0.771

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
zorro-local_attractor-in_question_with_aux	<u>0.669</u>	0.585	0.525	<u>0.807</u>	0.399	0.642	<u>0.878</u>	0.645	0.729	**0.969**	0.908	0.807
zorro-npi_licensing-matrix_question	0.981	**0.999**	0.767	<u>0.914</u>	0.859	0.668	<u>0.996</u>	0.706	0.723	<u>0.994</u>	0.951	0.851
zorro-npi_licensing-only_npi_licensor	<u>0.944</u>	0.892	0.704	<u>0.98</u>	0.725	0.614	**1.0**	0.972	0.744	<u>0.998</u>	0.992	0.909
average	0.752	0.705	0.577	0.723	0.435	0.547	0.795	0.622	0.596	**0.819**	0.774	0.615

Table 3: FlashHolmes syntactic tasks results. In **bold** are the overall best results, and <u>underlined</u> are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
arg-is-abstract	0.125	0.125	0.125	**0.212**	**0.212**	**0.212**	0.167	0.167	0.165	0.2	0.2	0.2
arg-is-kind	0.098	0.098	0.098	0.174	0.175	0.176	**0.209**	0.208	**0.209**	0.168	0.168	0.168
arg-is-particular	0.29	0.29	0.29	0.286	0.285	0.285	0.277	0.277	0.278	**0.359**	**0.359**	**0.359**
blimp-existential_there_quantifiers_1	0.861	0.802	0.556	**0.868**	0.402	0.634	0.859	0.741	0.627	0.841	0.775	0.512
blimp-existential_there_quantifiers_2	0.936	0.863	0.578	**0.993**	0.378	0.677	0.934	0.894	0.659	0.939	0.898	0.639
blimp-matrix_quest_npi_licensor_pres	0.995	**1.0**	0.807	0.986	**1.0**	0.862	0.973	0.704	0.733	0.989	0.948	0.691
blimp-npi_present_1	0.956	0.946	0.66	0.957	0.927	0.706	**0.973**	0.915	0.746	0.93	0.91	0.711
blimp-npi_present_2	0.953	0.954	0.702	**0.965**	0.91	0.664	0.939	0.737	0.727	0.95	0.858	0.622
blimp-only_npi_licensor_present	0.959	0.941	0.625	0.99	0.962	0.698	0.982	0.934	0.726	**0.995**	0.966	0.8
blimp-sent_neg_npi_licensor_present	0.972	0.942	0.712	0.956	0.943	0.711	**0.987**	0.877	0.699	0.985	0.966	0.652
blimp-superlative_quantifiers_1	0.932	0.773	0.626	**0.99**	0.398	0.661	0.96	0.621	0.657	0.986	0.922	0.687
blimp-superlative_quantifiers_2	0.976	0.939	0.645	**0.98**	0.378	0.612	0.888	0.738	0.643	0.863	0.723	0.606
context-object_animacy	0.765	0.894	0.8	0.753	0.339	0.736	0.918	0.339	0.71	0.955	**0.999**	0.83
context-object_gender	0.566	**0.638**	0.543	0.57	0.518	0.548	0.444	0.326	0.421	0.547	0.613	0.504
context-subject_animacy	0.942	0.756	0.752	0.814	0.346	0.679	**0.969**	0.367	0.765	0.966	0.945	0.806
context-subject_gender	**0.835**	0.77	0.671	0.75	0.744	0.594	0.533	0.367	0.501	0.496	0.566	0.484
context-verb_dynamic	0.813	**0.852**	0.691	0.724	0.334	0.542	0.742	0.43	0.54	0.743	0.623	0.519
context-verb_tense	**0.981**	0.969	0.913	0.816	0.324	0.61	0.898	0.324	0.702	0.762	0.723	0.525
cwi	0.49	0.49	0.49	0.51	0.51	0.51	0.488	0.488	0.488	**0.568**	**0.568**	**0.568**
event_structure-distributive	0.678	0.678	0.678	0.675	0.675	0.675	0.691	0.691	0.691	**0.694**	**0.694**	**0.694**
event_structure-event	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**	**0.241**
event_structure-has-natural-parts	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**	**0.452**
event_structure-has-similar-parts	0.302	0.302	0.302	0.284	0.284	0.284	0.337	0.337	0.337	**0.423**	**0.423**	**0.423**
event_structure-is-dynamic	0.364	0.364	0.364	0.363	0.363	0.363	**0.395**	**0.395**	**0.395**	0.376	0.376	0.376
event_structure-is-telic	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**	**0.466**
factuality	**0.466**	**0.466**	**0.466**	0.278	0.278	0.278	0.278	0.278	0.278	0.303	0.303	0.303
metaphor-lcc	0.796	0.796	0.796	0.803	0.803	0.803	**0.806**	**0.806**	**0.806**	0.798	0.798	0.798
metaphor-trofi	0.601	0.601	0.601	0.617	0.617	0.617	**0.642**	**0.642**	**0.642**	0.638	0.638	0.638
metaphor-vua_pos	**0.721**	**0.721**	**0.721**	0.719	0.719	0.719	0.712	0.712	0.712	0.708	0.708	0.708
metaphor-vua_verb	0.665	0.665	0.665	0.657	0.657	0.657	**0.678**	**0.678**	**0.678**	0.672	0.672	0.672
ner	**0.457**	0.456	**0.457**	0.449	0.449	0.449	0.393	0.393	0.393	0.401	0.401	0.401
passive	0.46	**0.48**	0.418	0.333	0.333	0.333	0.399	0.333	0.333	0.333	0.333	0.334
pred-is-dynamic	0.078	0.078	0.075	0.105	0.105	0.105	0.046	0.046	0.046	**0.111**	**0.111**	**0.111**

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
pred-is-hypothetical	0.256	0.257	0.256	0.264	**0.271**	**0.271**	0.193	0.191	0.195	0.216	0.199	0.216
pred-is-particular	0.117	0.116	**0.118**	0.076	0.086	0.086	0.074	0.071	0.071	0.102	0.104	0.103
protoroles-awareness	0.651	0.651	0.651	0.714	0.714	0.714	0.599	0.599	0.599	**0.72**	**0.72**	**0.72**
protoroles-change_of_location	0.115	0.115	0.115	**0.235**	**0.235**	**0.235**	0.141	0.141	0.141	0.197	0.197	0.197
protoroles-change_of_possession	0.238	0.238	0.238	**0.298**	**0.298**	**0.298**	0.139	0.139	0.139	0.25	0.25	0.25
protoroles-change_of_state	0.082	0.082	0.082	0.132	0.132	0.132	0.071	0.071	0.071	**0.144**	**0.144**	**0.144**
protoroles-change_of_state_continuous	0.138	0.138	0.138	**0.197**	**0.197**	**0.197**	0.107	0.107	0.107	0.179	0.179	0.179
protoroles-changes_possession	-0.203	-0.203	-0.203	-0.024	-0.024	-0.024	0.085	0.085	0.085	**0.185**	**0.185**	**0.185**
protoroles-existed_after	0.122	0.122	0.122	0.157	0.157	0.157	0.126	0.126	0.126	**0.173**	**0.173**	**0.173**
protoroles-existed_before	0.283	0.283	0.283	**0.377**	**0.377**	**0.377**	0.244	0.244	0.244	0.359	0.359	0.359
protoroles-existed_during	-0.011	-0.011	-0.011	**0.194**	**0.194**	**0.194**	0.074	0.074	0.074	0.146	0.146	0.146
protoroles-exists_as_physical	0.086	0.086	0.086	**0.304**	**0.304**	**0.304**	-0.045	-0.045	-0.045	0.079	0.079	0.079
protoroles-instigation	0.343	0.343	0.343	**0.425**	**0.425**	**0.425**	0.301	0.301	0.301	0.406	0.406	0.406
protoroles-location_of_event	-0.059	-0.059	-0.059	-0.485	-0.485	-0.485	-0.292	-0.292	-0.292	**0.184**	**0.184**	**0.184**
protoroles-makes_physical_contact	-0.035	-0.035	-0.035	**0.323**	**0.323**	**0.323**	0.149	0.149	0.149	0.146	0.146	0.146
protoroles-partitive	0.079	0.079	0.079	0.059	0.059	0.059	0.048	0.048	0.048	**0.121**	**0.121**	**0.121**
protoroles-predicate_changed_argument	**0.342**	**0.342**	**0.342**	0.099	0.099	0.099	-0.078	-0.078	-0.078	-0.175	-0.175	-0.175
protoroles-sentient	0.678	0.678	0.678	**0.748**	**0.748**	**0.748**	0.614	0.614	0.614	0.724	0.724	0.724
protoroles-stationary	-0.1	-0.1	-0.1	**0.04**	**0.04**	**0.04**	-0.154	-0.154	-0.154	-0.0	-0.0	-0.0
protoroles-volition	0.606	0.606	0.606	**0.685**	**0.685**	**0.685**	0.56	0.56	0.56	0.677	0.677	0.677
protoroles-was_for_benefit	0.327	0.327	0.327	0.36	0.36	0.36	0.259	0.259	0.259	**0.369**	**0.369**	**0.369**
protoroles-was_used	0.058	0.058	0.058	0.031	0.031	0.031	0.079	0.079	0.079	**0.105**	**0.105**	**0.105**
relation-classification	**0.362**	**0.362**	**0.362**	0.281	0.281	0.281	0.244	0.244	0.244	0.27	0.27	0.271
senteval-coordination_inversion	0.623	0.649	0.555	0.628	0.614	0.554	0.715	0.639	0.583	**0.721**	0.698	0.614
senteval-odd_man_out	0.622	0.612	0.575	0.63	0.584	0.552	**0.736**	0.7	0.619	0.73	0.714	0.661
senteval-past_present	**0.885**	0.883	0.844	0.873	0.86	0.818	0.87	0.725	0.716	0.862	0.843	0.716
senteval-word_content	**0.084**	0.04	0.02	0.02	0.0	0.014	0.006	0.0	0.002	0.006	0.004	0.005
sentiment-sentence	**0.425**	0.237	0.133	0.234	0.082	0.113	0.163	0.184	0.07	0.269	0.061	0.084
srl	0.114	0.114	0.114	0.105	0.105	0.105	0.118	0.118	0.118	0.125	**0.126**	**0.126**
synonym-antonym-hard	0.608	0.588	0.5	0.559	0.514	0.458	**0.63**	0.591	0.494	0.606	0.59	0.499
time	0.105	0.105	0.105	**0.106**	**0.106**	**0.106**	0.099	0.099	0.099	0.098	0.098	0.098
wordsense	**0.104**	**0.104**	**0.104**	0.084	0.084	0.084	0.071	0.071	0.071	0.086	0.086	0.086
zorro-quantifiers-existential_there	0.773	0.678	0.636	**0.905**	0.438	0.621	0.638	0.74	0.581	0.803	0.789	0.716

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
zorro-quantifiers-superlative	0.981	0.76	0.616	**0.997**	0.425	0.634	0.973	0.652	0.673	0.995	0.93	0.722
average	0.463	0.449	0.398	0.468	0.386	0.405	0.436	0.373	0.374	**0.474**	0.46	0.409

Table 4: FlashHolmes semantic tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
bridging-edge	0.563	0.563	0.563	0.519	0.519	0.519	0.574	0.574	0.574	**0.614**	**0.614**	**0.614**
bridging-sentence	0.4	0.4	**0.424**	0.4	0.4	0.4	0.4	0.4	**0.424**	0.4	0.4	0.405
coref	0.695	0.695	0.695	0.725	0.725	0.725	0.718	0.718	0.718	**0.735**	**0.735**	**0.735**
discourse-connective	0.073	0.068	0.082	0.055	0.023	0.047	0.064	0.023	0.056	0.064	**0.093**	0.071
gum-rst-edu-count	0.031	0.031	0.031	**0.034**	**0.034**	**0.034**	0.014	0.014	0.014	0.02	0.02	0.02
gum-rst-edu-depth	0.105	0.105	0.105	0.146	0.146	0.146	0.091	0.091	0.091	**0.151**	**0.151**	**0.151**
gum-rst-edu-distance	0.067	0.067	0.067	**0.13**	**0.13**	**0.13**	0.038	0.038	0.038	0.052	0.052	0.052
gum-rst-edu-relation	0.067	0.067	0.067	0.059	0.059	0.059	0.072	0.072	0.072	**0.076**	**0.076**	**0.076**
gum-rst-edu-relation-group	0.071	0.071	0.071	0.061	0.061	0.061	**0.109**	**0.109**	**0.109**	0.092	0.092	0.092
gum-rst-edu-successively	**0.486**	**0.486**	**0.486**	0.483	0.483	0.483	0.483	0.483	0.483	0.479	0.479	0.479
gum-rst-edu-type	0.664	0.664	0.664	0.658	0.658	0.658	**0.689**	**0.689**	**0.689**	0.68	0.68	0.68
next-sentence-prediction	0.429	**0.615**	0.429	0.429	0.429	0.429	0.502	0.429	0.43	0.429	0.497	0.429
ordering	0.865	0.865	0.865	**0.923**	**0.923**	**0.923**	0.727	0.727	0.727	0.798	0.798	0.798
averages	0.347	**0.361**	0.350	0.356	0.353	0.355	0.345	0.336	0.340	0.353	**0.361**	0.354

Table 5: FlashHolmes discourse tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

Data	BERT			RoBERTa			DeBERTa			Electra		
	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand	S-avg	CLS	Rand
SemAntoNeg	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	**0.408**	0.4	0.4	0.4
bioscope-negation-span-classify	**0.979**	**0.979**	**0.979**	0.969	0.969	0.969	0.971	0.971	0.971	0.97	0.97	0.97
bioscope-negation-span-classify	**0.601**	**0.601**	**0.601**	0.589	0.589	0.589	0.58	0.58	0.58	0.577	0.577	0.577
bioscope-org-negation	0.212	0.207	0.266	0.299	0.294	0.338	0.279	0.259	**0.417**	0.228	0.207	0.276
fuse-negation-span-classify	0.919	0.919	0.919	0.495	0.495	0.495	0.917	0.917	0.917	**0.942**	**0.942**	**0.942**
fuse-negation-span-correspondence	0.48	0.48	0.48	0.33	0.33	0.33	**0.512**	**0.512**	**0.512**	0.458	0.458	0.458
fuse-org-negation	0.2	0.196	0.239	0.28	0.29	**0.329**	0.203	0.196	0.275	0.211	0.196	0.267
olmpics-antonym_synonym_negation	0.515	0.507	0.46	0.455	0.396	0.483	**0.749**	0.392	0.477	0.665	0.703	0.48
olmpics-coffee_cats_quantifiers	0.444	0.444	0.444	0.444	0.444	0.444	0.444	0.444	0.453	0.444	0.444	**0.46**
olmpics-composition_v2	0.4	0.4	**0.409**	0.4	0.4	0.407	0.4	0.4	0.404	0.4	0.4	0.407
olmpics-compositional_comparison	0.4	0.4	0.404	0.4	0.4	0.407	0.4	0.4	**0.47**	0.4	0.4	0.434
olmpics-conjunction_filt4	0.4	0.4	0.403	0.4	0.4	**0.412**	0.4	0.4	0.4	0.4	0.4	0.403
olmpics-hypernym_conjunction	0.401	0.408	0.4	0.4	0.4	0.4	**0.42**	0.4	0.402	0.4	0.41	0.404
olmpics-number_comparison_age_compare_masked	0.404	0.476	0.487	0.38	0.372	0.457	**0.919**	0.333	0.433	0.638	0.603	0.459
olmpics-size_comparison	0.486	0.423	0.479	0.427	0.347	0.421	**0.556**	0.394	0.51	0.429	0.434	0.447
sherlock-negation	0.434	0.434	0.436	0.434	0.434	0.434	0.435	0.434	0.435	0.434	0.434	**0.437**
speculation-org	0.196	0.195	0.224	0.265	**0.319**	0.306	0.195	0.195	0.23	0.198	0.195	0.256
speculation-span-classify	0.514	0.514	0.514	0.47	0.47	0.47	**0.526**	**0.526**	**0.526**	0.505	0.505	0.505
speculation-span-correspondence	0.465	0.465	0.465	0.417	0.417	0.417	0.453	0.453	0.453	**0.475**	**0.475**	**0.475**
average	0.466	0.466	0.474	0.434	0.430	0.448	**0.514**	0.453	0.488	0.483	0.482	0.477

Table 6: FlashHolmes reasoning tasks results. In **bold** are the overall best results, and underlined are the results for the best performing variation for each transformer.

D PROBING FOR STRUCTURE

train on	test on	BERT	RoBERTa	DeBERTa	Electra
CLS	CLS	0.896 (0.088)	0.789 (0.027)	0.227 (0.058)	0.955 (0.006)
	AVG	0.910 (0.078)	0.793 (0.026)	0.130 (0.025)	0.971 (0.003)
	RAND	0.919 (0.070)	0.792 (0.023)	0.139 (0.019)	0.966 (0.002)
AVG	CLS	1.000 (0.000)	0.943 (0.013)	0.174 (0.020)	0.999 (0.001)
	AVG	0.999 (0.001)	0.936 (0.017)	0.325 (0.087)	0.997 (0.001)
	RAND	1.000 (0.000)	0.939 (0.018)	0.327 (0.096)	0.999 (0.001)
RAND	CLS	0.998 (0.001)	0.888 (0.009)	0.163 (0.023)	0.999 (0.001)
	AVG	0.998 (0.002)	0.895 (0.004)	0.233 (0.048)	0.998 (0.001)
	RAND	0.997 (0.003)	0.886 (0.005)	0.221 (0.048)	0.997 (0.003)

Table 7: Detailed results on detecting chunk structure in sentence embeddings. Averaged F1 scores (standard deviation) over three runs.