Out-of-Tune rather than Fine-Tuned: How Pre-training, Fine-tuning and Tokenization Affect Semantic Similarity in a Historical, Non-Standardized Domain

Stella Verkijk
Vrije Universiteit Amsterdam
Huygens Institute
s.verkijk@vu.nl

Piek Vossen
Vrije Universiteit Amsterdam

Abstract

Domain-specific encoder language models have been shown to accurately represent semantic distributions as they appear in the pre-training corpus. However, the general consensus is that general language models can adapt to a domain through fine-tuning. Similarly, multilingual models have been shown to leverage transfer learning even for languages that were not present in their pre-training data. Contrastively, tokenization has also been shown to have a great impact on a models' abilities to capture relevant semantic information, while this remains unchanged between pre-training and fine-tuning. This raises the question whether word embeddings for subtokens in models are of sufficient semantic quality for a target domain if not learned for the same domain. In this paper, we compare how different models assign similarity scores to different semantic categories in a highly specialized, non-standardised domain: Early Modern Dutch as written in the archives of the Dutch East India Company. Since the language in this domain is from before spelling conventions were established, and noise accumulates due to the fact that the original handwritten text went through a Handwritten Text Recognition pipeline, this use-case offers a unique opportunity to study both domain-specific semantics as well as a highly complex tokenization task for lesser-resourced languages. Our results support findings in earlier work that fine-tuned models may pick up spurious correlations in the adaptation process and stop relying on relevant semantics learned during pre-training. All code and data are available on our repo.

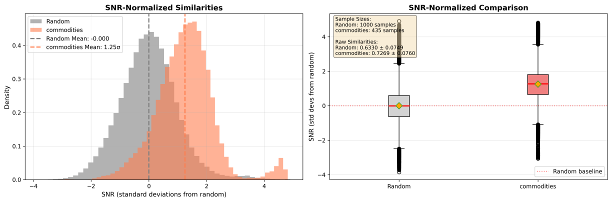

Figure 1: Signal-To-Noise ratio (SNR) for semantic similarity between strings referring to commodities (n=435) compared to similarity between random strings (n=1000) from the archives of the Dutch East India Company. A SNR of 1.25σ means the similarity between commodities is 1.25 standard deviations above the random similarity for this model. Embeddings for this figure were taken from the last layer of the base model of GloBERTise, a domain-specific encoder model.

1 Introduction

Accurately interpreting a language means interpreting the world it describes. Knowing whether a language model accurately represents a specific language or domain, taking into account the culture or world this specific language or domain describes, is not a trivial task. Models are often evaluated on downstream tasks such as named entity recognition or question answering, but the type of (linguistic) knowledge needed to solve these tasks can vary highly from task to task and from label to label, and therefore it is hard to pin down what a model is evaluated on exactly. For example, a question answering task can either rely heavily on syntax or on world knowledge depending on how you frame the task ('The doctor and the nurse have a coffee. What did she have?'). Similarly, it could be argued that recognizing a person in NER can rely more on syntax (since names often appear in similar contexts across languages and domains), while recognizing a ship might rely more on world knowledge (knowing that both a sailing ship and a plane can be a mode of transport depending on the world you live in). Additionally, models have been shown to rely on spurious correlations or shortcuts when performing downstream tasks (Du et al., 2021; Tang et al., 2023; Eshuijs et al., 2025). This results in task performance possibly not reflecting the linguistic or world knowledge we think models rely on when performing a specific task. Specifically, fine-tuned

Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026), pages 515–531
March 29, 2026 ©2026 Association for Computational Linguistics

models appear to be right for the wrong reasons because they pick up spurious correlations in the adaptation process (McCoy et al., 2019; Mosbach, 2023).

Larger language models are believed to have generalization abilities that enable them to use knowledge learned from one language or domain to interpret a language or domain the model has either not or only marginally been introduced to in its pre-training data. It remains unclear exactly what type of generalization abilities these models have. Generalizing syntactic knowledge of one Germanic language to another is one thing, but generalizing world knowledge from the modern society to that of a 17th-century colonial Dutch enterprise in the Indian Ocean world is another. We aim to provide more insight into whether models can do the latter.

Our paper provides an interpretable evaluation of how monolingual, multilingual and domain-specific models represent highly domain-specific semantics. We do so by evaluating different encoder Language Models on their abilities to cluster semantically similar concepts from the world of the Dutch East India Company (VOC): a colonial enterprise active in the Indian Ocean world between 1602 and 1799. We expect that general models not adapted to the domain lack the specific world knowledge needed to recognize domain-specific semantic clusters. For example, we expect a contemporary multilingual model to struggle with clustering ships described in early modern Dutch. We also study to what extent fine-tuning models on the domain-specific data enables general models to cluster domain-specific semantics.

In order to study whether a model accurately represents semantics relevant to a domain or culture, the first step is to recognize relevant tokens. After that, one can study how well these tokens are represented in the embedding space. We leverage an expert-annotated domain-specific dataset of named entities and semantic roles to select token sequences representing diverse concepts relevant to our domain. We study how well they are represented through subtoken sequence embeddings, both before and after finetuning. We compare between different models to test the following hypotheses:

- ▷ Multilingual models will struggle to accurately represent highly domain-specific semantics due to their tokenization and their lack of domain-specific world knowledge.
- ▷ Monolingual models will struggle for the same reasons, though in a lesser manner, since we expect tokenization to be better. We do not expect world knowledge of the domain to be better.
- ▷ The size of the model will not positively impact a models' abilities of representing domain-specific world knowledge since the size does not impact domain-specific world knowledge nor tokenization.
- ▷ A model pre-trained on the domain will perform best due to the optimized knowledge and tokenizer.
- ▷ Fine-tuned models will be more adapted to the domain and have better semantic representations, though not as good as a model with a dedicated tokenizer.

By testing these hypotheses, we investigate the impact of the amount and quality of the data that models are fine-tuned on in relation to the amount of (linguistic) variation in the target data or domain. We consider pre-training data of high quality when they are representative of the target domain. We evaluate this in a domain that features a high level of variance in the language. We expect that data quality is more important for capturing variance than the amount of pre-training data.

Our methodological contribution is a semantic Signal-to-Noise Ratio (SNR), an interpretable metric quantifying the extent to which within-concept similarity exceeds the model's baseline random similarity. See Figure 1 for an example.

2 Related Work

2.1 Related work in probing and ablation studies

Probing and analysing a model for linguistic or semantic knowledge is tricky because different types of knowledge might be stored in different layers of the model. Several studies find that encoder models follow a classical NLP pipeline within their layers, with knowledge of POS tagging being stored in earlier layers, then named entity recognition and then dependency parsing and coreference resolution in the last layers (Tenney et al., 2019; de Vries et al., 2020; Van Aken et al., 2019). However, findings in this area differ between studies.

While de Vries et al. finds that semantically rich POS tags become harder to identify in later layers, Jawahar et al. finds that semantic features are captured at the top of a model. Also, while de Vries et al. find that (some) information is lost in the last layers of a model, Van Aken et al. find that the last layer of a finetuned model is task-specific, suggesting that the final representation needed to solve a downstream task is stored in the last layer of a downstream model.

Deciding which layers to probe for a specific task is not only hard because you might not know what type of knowledge is stored where, but also because knowing what type of knowledge is needed for a task is difficult. For example, different tag abstractions for POS are present in different layers (de Vries et al., 2020). Furthermore, Jawahar et al. find that knowledge on phrase syntax (i.e., knowing that tokens 'to' and 'demonstrate' belong together to form a verb phrase) is only present in lower layers and gets diluted in later layers. For a task like semantic role labelling, for which a model has to know something about phrase syntax (to recognise mentions) but also about high-level semantics (i.e. a ship can be an instrument in a Transportation event), it becomes hard to pin down which layers to probe. Even more so since SRL in our domain, which features very long sentences (Verkijk et al., 2024), should rely on knowledge of long-range dependencies, and according to Jawahar et al. encoders store long-distance dependency information in deeper layers. Additionally, it is known that one type of knowledge is usually also spread over different layers (de Vries et al., 2020; Vulic et al., 2020). Similarly, (Vulic et al., 2020) find that averaging across all layers of a model is beneficial across the board. On the other hand, they also state that for some tasks, scores go down when including higher layers into averaging. We therefore incorporate experiments probing different layers in our experiments (see Section 5.1).

Interestingly, although fine-tuning is known to greatly improve models' ability to solve tasks in unknown domains and languages, Van Aken et al. find that fine-tuning has little impact on models' semantic abilities. In more specific work comparing base models and fine-tuned versions of models, Mosbach finds that fine-tuning mostly affects the upper layers of models and that fine-tuning in some cases hurts probing performance in lower layers. It seems that when fine-tuning on a different domain than than a model is pre-trained on, the model relies most on change in the penultimate layer (Goerttler and Obermayer, 2022; Oh et al., 2020). However, Goerttler and Obermayer also find that earlier layers have to change more when solving cross-domain tasks than when solving domain tasks in order to adapt to the new input data. There are also findings that show fine-tuning sometimes hampers performance by introducing divergence between the training and test set, hurting generalization (Zhou and Srikumar, 2022). This study also finds that fine-tuning changes the geometry of a representation by pushing points belonging to the same level closer to each other, while largely preserving the original spatial structure of the data points. Zhou and Srikumar acknowledge that even though fine-tuning tends to provide strong performance, how fine-tuning manages to do so remains an open question.

Studying construction grammar (CC), Weissweiler et al. find that though models are able to recognize the structure (syntax) of CC, they fail to use its meaning. They see this as evidence that LMs still suffer from substantial shortcomings in central domains of linguistic knowledge.

As for multilingual models, results of all experiments by Vulic et al. point out that monolingual models contain much more lexical information for a target language than large multilingual models. Wu and Dredze (2020) caution against using mBERT alone for low resource languages, since the lowest resource languages seem to help better represented languages in the model, but not the other way around. They also point out the following dichotomy: although the number of shared subwords across languages correlates with cross-lingual performance (Wu and Dredze, 2019), mBERT has the capacity to learn cross-lingual representation without any vocabulary overlap (Wu and Dredze, 2019; Karthikeyan et al.). This indicates that the influence of tokenization on representation learning remains opaque. A recent study identifies token similarity and country similarity as pivotal factors in the success of a multilingual model, where the latter highlights the importance of shared cultural contexts (Nezhad et al., 2025).

2.2 Related work in Early Modern Dutch

Earlier work on NLP for the archives of the Dutch East India company has evaluated different encoder models on their downstream performance in this domain. Arnoult et al. (2025) show that fine-tuning a domain-specific model only pre-trained on data from the archives improves NER performance compared to much larger off-the-shelf multilingual pre-trained models. Verkijk et al. (2025) demonstrate the same results for the task of event mention detection. Additionally, Verkijk et al. find that language-specific models for contemporary Dutch underperform compared to multilingual models. A similar finding can be found in Arnoult et al. (2021). Although these studies perform elaborate analysis on downstream performance of the different models, they do not provide any probing experiment or ablation studies to explain these results. We aim to provide more context by doing so.

NER label	Examples
Location	*Sumatra*; *Batavia*
Commodity	*peper* (pepper); *buskruijt* (gunpowder)
Ship	*Abigail*; *de Bredamme*
Date	*14e, ultimo Februari*

Table 1: Named Entity categories

SRL label	Description
Agent	Animate or non-animate actor of any type of event
Patient	Animate or non-animate undergoer of any type of event
Time	Any indication of time ('yesterday', 'may 12th', 'this year')
Leaving Agents	Boats, people, animals, goods
Moving Instruments	Anything used for transportation (ships, elephants, vessels)
Given Patients	Anything gifted (land, money, elephants)
Enslaved Patients	Anyone being forced to be a slave
Enslaving Agents	Anyone forcing someone to be a slave

Table 2: SRL categories

3 Data

The data we work with are the annotated data provided by the GLOBALISE project[1]. They provide text from 27 different documents, comprising a total of 280 handwritten pages (27,8 MB), annotated with entities, event classes and semantic roles, according to their guidelines[2]. We use a subset of the types of annotated NEs for analysis. See Table 1 for a description of the NEs used in this study.

For the semantic roles, we made a selection of categories varying in granularity. It ranges from the general semantic role of *agent* to the more specific semantic role of *patient in an enslaving event* (i.e., the person who is enslaved). See Table 2 for an overview of our semantic role classes.

4 Defining Semantic Similarity

Semantics change over time, and are specific to a certain world or community. Defining 'a world' as a sociocultural and temporal container, we could say that semantics change as the world changes. Some semantics can be learned through linguistic patterns, which is the distributional semantics language models rely on. When it comes to assigning semantic similarity between word pairs, we would argue that depending on the word-meaning pairs, some represent a form of linguistic knowledge, some world knowledge, and most a combination of both. This is scalar and fuzzy, rather than completely distinguishable in two categories.

Linguistic variation, where multiple words refer to the same thing (synonyms), as well as ambiguity, where one word refers to multiple things, are forms of linguistic problems. Assigning high similarity to the words *closed* and *shut* is thus a clear proof of linguistic capacity, and something that can be directly inferred from the position of words in context. However, some semantic grouping relies more on knowledge of certain worldly customs, rather than knowledge of static attributes of items. For example, assigning high similarity to 'Kauris' (Cowrie) and 'contanten' (money) can be seen as proof of domain-specific world knowledge: In the Indian Ocean region, cowrie was used as shell money in the time the Dutch East India Company was active. In our experiments, we probe for semantic similarity between different groups of tokens. Depending on the semantic group, we are testing knowledge that is either leaning more towards the linguistic side of the spectrum, or towards the side representing world knowledge.

Note that both linguistic knowledge and world knowledge can also be placed on a spectrum of domain-specificity. For example, assigning high similarity to two tokens that should be the same word but have been spelled differently can be seen as linguistic knowledge specific to our domain (transcribed archival documents from a period before spelling conventions). Similarly, identifying cowrie as similar to money is more domain-specific than identifying cats and dogs as similar. We investigate to what extent (non-)domain specific models can model domain-specific linguistic as well as world knowledge, both before and after finetuning. Our selected entities and some of our selected semantic roles are highly domain-specific. This enables us to investigate how language models

[1] https://globalise.huygens.knaw.nl

[2] link to guidelines for events and semantic roles; link to guidelines for named entities

represent relatively unknown semantics.

Since we work with language models that learn distributional semantics, we should also take into account the patterns they may learn from syntax. We might expect a model to still recognize locations it has not seen during pre-training as locations because it has seen other locations in similar contexts and can thus rely on a (syntactic) pattern it has learned. However, we should also recognize that syntax of Early Modern Dutch is different from contemporary Dutch, and that HTR and a lack of spelling conventions introduce noise interfering with these patterns. Recognizing certain semantic roles, like a general *agent*, could be seen both as a syntactic as well as a semantic task. Recognizing that elephants can be transported on sailing ships to be gifted to local kings (i.e., act as a *commodity* or a *patient* in a Giving event) can be seen as knowledge of the world of the Dutch East India Company. We therefore want to reiterate the fact that we are evaluating models on various types of knowledge in combination.

5 Method

Our experiments consist of probing different layers of different models and calculating a Signal-To-Noise ratio (Figure 1) indicating the models' ability to semantically cluster domain-specific entities and semantic roles. We explained our choice of entities and semantic roles in Sections 3 and 4. In this section, we first explain our choice of models and layers to probe in Sections 5.1 and 5.2. We then continue to explain our probe and the Signal-To-Noise ratio we calculate in Section 5.3.

5.1 Experimental set-up

Our experiments feature three layer settings: probing the last layer, layers 2-4 (inclusive), and all layers. We choose the last layer because we want to compare base models and their fine-tuned versions and previous work has found later layers are most impacted by fine-tuning (Van Aken et al., 2019; Mosbach, 2023). We choose layers 2-4 because lower layers are known to isolate basic linguistic knowledge, like phrase syntax (Jawahar et al., 2019). Also, in a study where embeddings of different layers of XLM-Roberta are tested as input for a Named Entity Disambiguation task, layer 3 performed best (Tufa et al., 2023). Since we are dealing with models of different amounts of layers, we therefore select the range of layers 2-4. Finally, we probe all layers together since previous studies have found knowledge is spread throughout layers of encoder models (de Vries et al., 2020; Vulic et al., 2020), and we expect both our tasks and especially our SRL task to rely on a variety of different types of knowledge.

For entities, we probe six base models (two contemporary Dutch, two contemporary multilingual, one Dutch historical and one domain-specific (Early Modern Dutch from VOC archives)). We do so to compare multilingual, monolingual and domain-specific models. For the domain-specific model, there is a fine-tuned version for NER available on GLOBALISE's Huggingface. We use this to test whether similarity between representations of named entities increases after fine-tuning. For semantic roles we probe three models. These are the two multilingual models and the domain-specific model. We opt for these three models because they perform best in downstream tasks (Verkijk et al., 2025; Arnoult et al., 2025) and because there are fine-tuned versions for SRL of these models available. We compare results between base models and their fine-tuned versions to see if similarity between representations of semantic roles improves after fine-tuning.

Lastly, we also compare the semantic similarity between entities with three base models and their versions fine-tuned for SRL to see whether fine-tuning on SRL leaves the semantic representations for entities intact or not.

5.2 Models

We investigate two contemporary multilingual models (mBERT (Devlin et al., 2019) and XLM-R (Liu et al., 2019)), two monolingual models of contemporary Dutch (BERTje (De Vries et al., 2019) and RobBERT (Delobelle et al., 2020)) and two historical Dutch models (GysBERT (Manjavacas and Fonteyn, 2022) and GloBERTise (Verkijk et al., 2025)). Each of these pairs consists of one BERT and one RoBERTa-based architecture. Only one of the six models is completely domain-specific. GloBERTise has been pre-trained on digitized texts from the archives of the Dutch East India Company (VOC) only. For an overview of sizes of the models and their pre-training data see Table 7 in Appendix A. No data cleaning or filtering was performed, as the authors of the model preferred it to learn from the noisy data it would later also have to predict on. The other historical Dutch model, GysBERT, was trained on language from a span of

Layers 2-4				
Base Model	loc.	ships	comm.	dates
GloBERTise	0.21	0.27	-0.18	1.55
XLM-R	-0.39	-0.04	-0.34	0.30
mBERT	-0.80	-0.48	-0.55	0.46
RobBERT	-0.34	-0.53	-0.94	1.00
BERTje	-1.22	-1.04	-1.53	0.19
GysBERT	-0.42	-0.52	-0.50	1.25
All layers				
Base Model	loc.	ships	comm.	dates
GloBERTise	0.48	0.50	0.26	1.39
XLM-R	-0.38	-0.14	-0.42	0.44
mBERT	-0.53	-0.27	-0.16	0.32
RobBERT	-0.42	-0.48	-1.13	0.55
BERTje	-0.65	-0.58	-1.11	0.42
GysBERT	0.08	-0.08	-0.03	1.04
Last layer				
Base Model	loc.	ships	comm.	dates
GloBERTise	0.96	0.86	1.25	0.67
XLM-R	-0.36	-0.19	-0.48	0.32
mBERT	0.09	0.08	0.43	0.19
RobBERT	-0.25	-0.31	-0.84	0.44
BERTje	0.71	0.48	0.26	0.64
GysBERT	0.46	0.29	0.44	0.80

Table 3: SNR values for **entity types** by base model across different layer configurations.
All results are statistically significant with $p < .001$ (t-test).

almost 500 years, and is therefore a broader model. It did not include language from the VOC archives. Furthermore, they used a perplexity threshold to filter out very noisy data before pre-training. For the domain-specific model as well as the two multilingual models, GLOBALISE provides fine-tuned versions, which we also evaluate.

5.3 SNR calculation

We do not train classifiers with the embeddings we extract from the models. Instead, we opt for a direct probe. We chose to do so because, although semantic similarity does not provide a complete image of a model's knowledge and/or linguistic capacities, probing with external classifiers introduces an extra layer of noise (de Vries et al., 2020), which can be avoided by directly comparing the models' representations.

When working with cosine similarity of LM embeddings, there are many things that can negatively influence the results. The main issue here is that internal representations in encoder models occupy a narrow cone in the vector space (Luo et al., 2021). This means that one model may have a very high general similarity between all tokens, whereas another model has a much lower general similarity. This can be shown by calculating the similarity between embeddings of random strings. For example, calculating this for mBERT on 1000 random strings from our corpus results in a similarity score of 0.44, where for XLM-R this is 0.99. Therefore, it is problematic comparing similarity scores of different models to each other without any normalization or data transformation. This problem of a narrow cone is also sometimes referred to as representation degeneration or anisotropy (Godey et al., 2024). Some solutions that are proposed for this are *clipping*, where positional outliers are identified and deleted from embeddings (Luo et al., 2021), or *whitening*, a method often used in Machine Learning where data are transformed to have zero mean and unit variance in all directions, with no correlation between features. For LMs, this essentially means you fit and transform on your embeddings to lay them out over a wider space. However, clipping only deals with noise from position information and whitening has been shown to have issues (Forooghi et al., 2024) and is not very interpretable.

We propose a more interpretable calculation of a Signal-To-Noise Ratio for semantic string similarity.

$$\text{SNR} = \frac{sim_{\text{cat}} - sim_{\text{rand}}}{\sigma_{\text{rand}}}$$

where:

- sim_{cat} is the mean pairwise cosine similarity between strings referring to a certain category
- sim_{rand} is the mean pairwise cosine similarity between random strings of up to five tokens
- σ_{rand} is the standard deviation of pairwise similarities for random embeddings

By using this calculation, we essentially compute a $z - score$ showing how many standard deviations the category's mean similarity is from the random baseline. See Figure 1 for an example of a plot where similarity between an entity category is very different from the random similarity. We gather embeddings for multi-token strings by taking the mean of the stacked embeddings per token.

6 Results

Table 3 shows the SNR scores of six base models for entities. As we can see, the domain-specific GloBERTise demonstrates the strongest signals. A dramatic result is demonstrated by RobBERT

Layers 2-4								
Base	enslaved_pt	leaving_ag	moving_ins	given_pt	enslaving_ag	agents	patients	time
GloBERTise	1.30	-0.14	0.31	-0.26	0.71	-0.12	-0.21	0.17
XLM-R	0.37	0.06	0.35	-0.15	0.45	0.15	0.13	-0.09
mBERT	1.12	-0.03*	0.11	-0.04*	0.52	0.09	0.17	-0.05
Finetuned								
GloBERTise	1.11	-0.34	-0.13	-0.45	0.42	-0.41	-0.49	0.22
XLM-R	0.63	0.05	0.43	-0.14	0.56	0.19	0.18	-0.14
mBERT	1.31	-0.02**	0.25	-0.02**	0.67	0.12	0.11	0.02
All layers								
Base	enslaved_pt	leaving_ag	moving_ins	given_pt	enslaving_ag	agents	patients	time
GloBERTise	1.44	0.01	0.42	-0.03*	0.92	0.07	-0.01	0.12
XLM-R	0.58	0.09	0.30	-0.20	0.49	0.14	0.10	-0.09
mBERT	1.05	0.02	0.18	0.07	0.54	0.12	0.18	-0.11
Finetuned								
GloBERTise	0.93	-0.27	-0.22	-0.34	0.56	-0.30	-0.35	-0.01*
XLM-R	0.78	0.05	0.24	-0.14	0.46	0.14	0.03	-0.04
mBERT	1.16	-0.01**	0.14	0.01**	0.68	0.12	0.03	0.03
Last layer								
Base	enslaved_pt	leaving_ag	moving_ins	given_pt	enslaving_ag	agents	patients	time
GloBERTise	1.21	0.26	0.29	0.30	0.59	0.17	-0.03	-0.10
XLM-R	0.57	0.13	0.23	-0.10	0.34	0.13	0.13	0.02
mBERT	0.99	0.15	0.25	0.26	0.33	0.09	0.15	0.01**
Finetuned								
GloBERTise	0.28	-0.21	-0.52	-0.49	0.56	-0.18	-0.23	-0.60
XLM-R	0.62	-0.15	-0.36	-0.21	0.00**	-0.11	-0.09	-0.17
mBERT	0.83	-0.08	-0.47	0.20	0.62	0.05	0.17	-0.58

Table 4: SNR values for **semantic roles** by model (base vs. fine-tuned) and across layer configurations. All results are statistically significant with $p < .001$ (t-test) except for results indicated with *, which are significant with $p < .05$ and those indicated with **, which are not significant.

and XLM-R, giving a negative signal for three out of four entities even in the best performing layer setting. This demonstrates how contemporary language models can struggle clustering entities specific to a certain domain and time. On the other hand, BERTje performs more reasonable. We investigate whether this is because of tokenization in Section 7. The results show that the last layer captures most information on all entities except dates, indicating that in our case recognizing most entities might be a task that relies more on semantics than other types of named entity recognition.

A striking difference is that the category of dates in the entities is much easier to cluster than the category of time in the semantic roles (Table 4). This is an example of how linguistic variety complicates semantic modelling. References to dates mostly follow a similar pattern (‘19 Julij 1601’, ‘’den 14=e julij’, ‘28=en passado’), while time indications for semantic role labelling are way more varied, including strings such as ‘voorledene dingsdag den 23 der gepasseerde maand april des morgens ten Elf uuren’ (*past tuesday the 23rd of the passed month of april in the morning at eleven hours*), ‘thans’ (*now*) and ‘snachts’ (*at night*). Looking at base models, GloBERTise is the only model showing a positive SNR of above 0.10 for time in two of the three layer settings. This indicates that the domain-specific model is more capable of dealing with variation, supporting our hypothesis.

The fine-tuned version of the GloBERTise model for NER shows drastically different results than the base model (see Figure 2), with an SNR for locations of -1.13, for ships of -1.84 and of commodities of -0.57 (versus the positive values of 0.96, 0.86 and 1.25 from the base model). This is completely counter-intuitive, since we expected fine-tuned models to adapt to the domain and cluster references to entities with higher inter-similarity than random tokens.

Looking at Table 4, another stark difference can be seen between the positive signals produced by

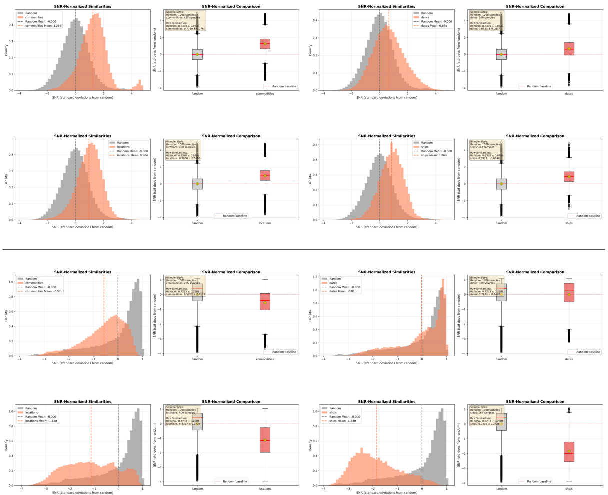

Figure 2: Plots of SNR values for **entities** probed from the last layer of **GloBERTise base model** (above the midline) and the last layer of **GloBERTise fine-tuned for NER** (underneath the midline).

the last layer of base models as opposed to the negative signals produced by the last layer of finetuned models. Even the most specific semantic roles produce a weaker signal in the fine-tuned versions of the models. Fine-tuned models show an even bigger drop in SNR scores for semantic roles that feature more variation (including *Time*). Fine-tuning does not seem to help overcoming a high level of variation in the data. For a visualization of how representations of semantic roles in the last layer change after fine-tuning see Figures 6-8 in Appendix B.

When analysing how fine-tuning on SRL affects the models' representations of entities, let us first turn to Figure 3. Here, we depict GloBERTise's SNR values and standard deviations for all entities, both in its base-model form (above the line) and in its fine-tuned form (underneath the line). For plots like these for mBERT and XLM-R, see Figures 4 and 5 in the Appendix B. We consistently see across models how the similarity between entities gets disrupted by fine-tuning on SRL, contrary to what is sometimes found in earlier work that finds that higher layers remain close to the original representations after fine-tuning (Zhou and Srikumar, 2022). We can also see how the similarity score of random strings decreases after fine-tuning. What is striking here is that the difference between random tokens and categories seems to disappear consistently after fine-tuning. This means that the vector space only becomes denser: all tokens shift toward the domain data, and representations of random tokens shift most of all.

In general (both for entities and semantic roles), the largest model with respect to amount of parameters as well as the amount of pre-training data, namely XLM-R, does not perform better than mBERT, the smaller multilingual model of the two. GloBERTise, the model with the least pre-training data, outperforms all other models.

7 Error Analysis: Tokenization

We study a highlight of our results in light of tokenization, namely the case of commodities. The domain-specific model has a high signal of similarity for this group, while other models do not and one monolingual Dutch model even has a negative signal. It is interesting to see that the monolingual model with a RoBERTa architecture and a ByteLevel-BPE tokenizer (RobBERT) performs notably worse than the monolingual model with a BERT architecture and WordPiece tokenizer (BERTje).

Tokenizer	# subtokens	% subtok. overlap historical lexicon
GysBERT	944	21%
GloBERTise	1004	16%
RobBERT	1131	13%
XLM-R	1161	6%
mBERT	1216	5%
BERTje	1221	11%
Original token count = 505; overlap lexicon = 48%		

Table 5: Amount of subtokens per model for the entity category of **commodities**

We start our analysis with some general statistics. In Table 5 we show in how many subtokens each model divides the total set of 505 tokens referring to commodities. We also calculate the percentage of overlap that exists between these subtokens and a hisotrical lexicon for OCR and OCR-postcorrection made for Dutch from the period of approx. 1550 - approx. 1970, developed by the Institute for Dutch Language[3]. Since this lexicon also contains very short, non-semantic subtokens (as it was made for OCR), we take a subset of the lexicon containing all (sub)tokens of 4 or more letters. This reduces the lexicon to a total of 499187 (sub)tokens.

The fact that GysBERT shows highest overlap with the lexicon makes sense since it is a general model of historical Dutch and less specialized than GloBERTise. We see that GloBERTise

[3] see https://taalmaterialen.ivdnt.org/download/tstc-int-historische-woordenlijst/

	olie	olij	oliphanten	eliphanten	contanten	comptanten	pistoolen	oologsmunitie
olie	1.00	0.80	0.67	0.69	0.60	0.63	0.78	0.76
ol-ij	0.80	1.00	0.82	0.80	0.70	0.68	0.82	0.83
ol-ip-h-anten	0.67	0.82	1.00	0.96	0.74	0.74	0.81	0.82
e-ip-h-anten	0.69	0.80	0.96	1.00	0.73	0.74	0.83	0.84
cont-anten	0.60	0.70	0.74	0.73	1.00	0.77	0.74	0.73
com-pt-anten	0.63	0.68	0.74	0.74	0.77	1.00	0.71	0.77
pist-ool-en	0.78	0.82	0.81	0.83	0.74	0.71	1.00	0.85
ool-og-sm-un-itie	0.76	0.83	0.82	0.84	0.73	0.77	0.85	1.00

Table 6: Cosine similarity matrix for **RobBERT**. Higher values (greener) indicate higher similarity.

and RobBERT have a similar amount of subtokens and percentage of overlap. Hence, we cannot explain the difference in their performance directly from the statistics of their subtoken representation. This seems to indicate that domain-specific distributional semantics could be more important than domain-specific tokenization.

We continue our analysis by looking in more detail at how certain words are tokenized. We choose to compare RobBERT and BERTje since they had similar pre-training data but differ in tokenizer type. We selected interesting cases from the set of tokens that refer to commodities, where certain pairs should have a similarity of 1 since they mean the same: *olie* and *olij* (oil); *oliphanten* and *eliphanten* (elephants); *contanten* and *comptanten* (money). Also, *pistoolen* (pistols) and *oologsmunitie* (war munition) should be more similar. This selection of tokens is an interesting test bed because for both models, there is different overlap in subtokens across pairs. For example, in the case of BERTje, *oologsmunitie*, *olij* and *oliphanten* share the subtoken 'o', whereas in the case of RobBERT, *oliphanten* and *olij* share the same subtoken 'ol'. In the case of RobBERT, the spelling variants *contanten* and *comptanten* share the subtoken 'anten', whereas for BERTje, they share the very common subtoken 'en'.

Results can be found in Table 6 (above) and 8 (Appendix C). As an upper bound, we also show results of GloBERTise in Table 9 (Appendix C). We can see how BERTje completely misses the similarity between the two strings for *oil*. It only shows a weak similarity for *comptanten* and *contanten*, undistinguishable from the similarity it gives for *comptanten* and either of the words for elephants. In that sense, RobBERT seems to do a bit better here. On the other hand, RobBERT sees *pistoolen* and *oologsmunitie* as similar to all other tokens. Also, RobBERT sees *oliphanten* and *olij* as similar, possibly because of their shared subtoken 'ol'.

8 Discussion

We return to the hypotheses presented in the introduction of this paper. Our results confirm the hypothesis that multilingual models and monolingual models struggle to accurately represent highly domain-specific semantics and that a model pretrained on the domain performs best. Our results seem to indicate that this is mostly because of a lack of specific world knowledge and in a lesser extent due to tokenization. There is no clear obvious distinction in performance between multilingual models and contemporary models not tuned to the domain. This finding can be extended to other low-resource scenarios, for example when language models do not accurately represent a culture (world) described in a low-resource language. Our results also confirm the hypothesis that neither the size of a model (paramater-wise) nor the size of its pre-training data positively impact knowledge on domain-specific semantics, even when fine-tuned on the domain. This is a finding that likely extends in some way to decoder models, since pretraining strategies of these models are similar to the encoder models studied here. Finally, our results show that fine-tuning models does not make the representations of relevant tokens more semantically accurate, on the contrary, it generally makes them less semantically accurate. These results might indicate a form of catastrophic forgetting. This raises the question: what knowledge do fine-tuned models actually use? If the semantics of relevant concepts deteriorate, the model must rely on task-specific shortcuts that are neither explainable nor likely to generalize well. Our finding is highly counterintuitive, and further research is needed to confirm our results and investigate possible causes.

9 Limitations and ethical considerations

Our work presents several limitations. Firstly, it would be good to gain more insight into the specific reasons why all models successfully cluster *enslaved patients* and *enslaving agents*. Apart from the fact it makes sense because they both only concern humans, it might also have to do with a high level of lexical overlap (low variety), especially since there were specific ways to refer to enslaved people. Moreover, a more fine-grained ablation analysis could be beneficial, analysing what each layer in the model adds to or deletes from a semantic representation. Furthermore, we only study encoder models in this paper and have not tested the representations of generative (decoder) models. Future research could investigate i) to what extent these models provide accurate representations for the semantic categories proposed in this paper and ii) whether instruction tuning improves or disrupts semantic representations. Lastly, it remains an open discussion what type of semantic clustering is informative for certain tasks. The group of commodities contains concepts that in some ways, for example in traditional (non-distributional) semantic theory, should not be seen as similar at all, such as oil and elephants.

It is important to be aware of the fact that a domain-specific model for the archives we work with, since it accurately represents the world the archives describe, is biased in ways that do not align with contemporary ethical values. For example, this model is probable to see people as a type of commodity, since they did so in the Dutch East India Company.

Acknowledgements

This research was funded by the Dutch Research Council (NWO) under project number 175.2019.003 and is part of the GLOBALISE project. We thank dr. Lodewijk Petram for proofreading.

References

Sophie Arnoult, Brecht Nijman, and Leon van Wissen. 2025. Fine-grained named-entity recognition for the east-india company domain. *Anthology of Computers and the Humanities*, 3:953–967.

Sophie I Arnoult, Lodewijk Petram, and Piek Vossen. 2021. Batavia asked for advice. pretrained language models for named entity recognition in historical texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 21–30.

Wietse De Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. Bertje: A dutch bert model. *arXiv preprint arXiv:1912.09582*.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about bert's layers? a closer look at the nlp pipeline in monolingual and multilingual models. *Findings of the Association for Computational Linguistics: EMNLP*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *Findings of the Association for Computational Linguistics: EMNLP*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu models. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Leon Eshuijs, Shihan Wang, and Antske Fokkens. 2025. Short-circuiting shortcuts: Mechanistic investigation of shortcuts in text classification. *arXiv preprint arXiv:2505.06032*.

Ali Forooghi, Shaghayegh Sadeghi, and Jianguo Lu. 2024. Whitening not recommended for classification tasks in llms. In *The 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, page 285.

Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48.

Thomas Goerttler and Klaus Obermayer. 2022. Similarity of pre-trained and fine-tuned representations. *arXiv preprint arXiv:2207.09225*.

Ganesh Jawahar, Benoıt Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

K Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

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

Ziyang Luo, Artur Kulmizev, and Xiaoxi Mao. 2021. Positional artefacts propagate through masked language model embeddings. In *Proceedings of the 59th annual meeting of the Association for Computational Linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5312–5327.

Enrique Manjavacas and Lauren Fonteyn. 2022. Non-parametric word sense disambiguation for historical languages. In *Proceedings of the 2nd international workshop on natural language processing for digital humanities*, pages 123–134.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Marius Mosbach. 2023. Analyzing pre-trained and fine-tuned language models. In *The Big Picture Workshop*, page 123.

Sina Bagheri Nezhad, Ameeta Agrawal, and Rhitabrat Pokharel. 2025. Beyond data quantity: Key factors driving performance in multilingual language models. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 225–239.

Jaehoon Oh, Hyungjun Yoo, ChangHwan Kim, and Se-Young Yun. 2020. Boil: Towards representation change for few-shot learning. *arXiv preprint arXiv:2008.08882*.

Ruixiang Tang, Dehan Kong, Longtao Huang, and 1 others. 2023. Large language models can be lazy learners: Analyze shortcuts in in-context learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4645–4657.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Wondimagegnhue Tufa, Lisa Beinborn, and Piek Vossen. 2023. A wordnet view on crosslingual contextualized language models. *University of the Basque Country in Donostia-San Sebastian Basque Country*, page 14.

Betty Van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1823–1832.

Stella Verkijk, Pia Sommerauer, and Piek Vossen. 2024. Studying language variation considering the re-usability of modern theories, tools and resources for annotating explicit and implicit events in centuries old text. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 174–187.

Stella Verkijk, Piek Vossen, and Pia Sommerauer. 2025. Language models lack temporal generalization and bigger is not better. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20629–20637.

Ivan Vulic, Edoardo M Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pre-trained language models for lexical semantics. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Leonie Weissweiler, Valentin Hofmann, Abdullatif Köksal, and Hinrich Schütze. 2022. The better your syntax, the better your semantics? probing pretrained language models for the english comparative correlative. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *Proceedings of the 5th Workshop on Representation Learning for NLP*, page 120.

Yichu Zhou and Vivek Srikumar. 2022. A closer look at how fine-tuning changes bert. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1046–1061.

Appendix A: Additional information accompanying Section 5

	Param	tok/byt in data
GysBERT	110M	7.1B /
BERTje	109M	2.4B / 12GB
RobBERT	117M	6.6B / 39GB
mBERT	179M	/
XLM-R	279M	/ 2.5TB
GloBERTise	117M	/ 6GB

Table 7: Models with parameter and pre-training data size. All info missing in this table could not be found in the relevant papers.

Appendix B: Additional SNR Plots Accompanying Section 6

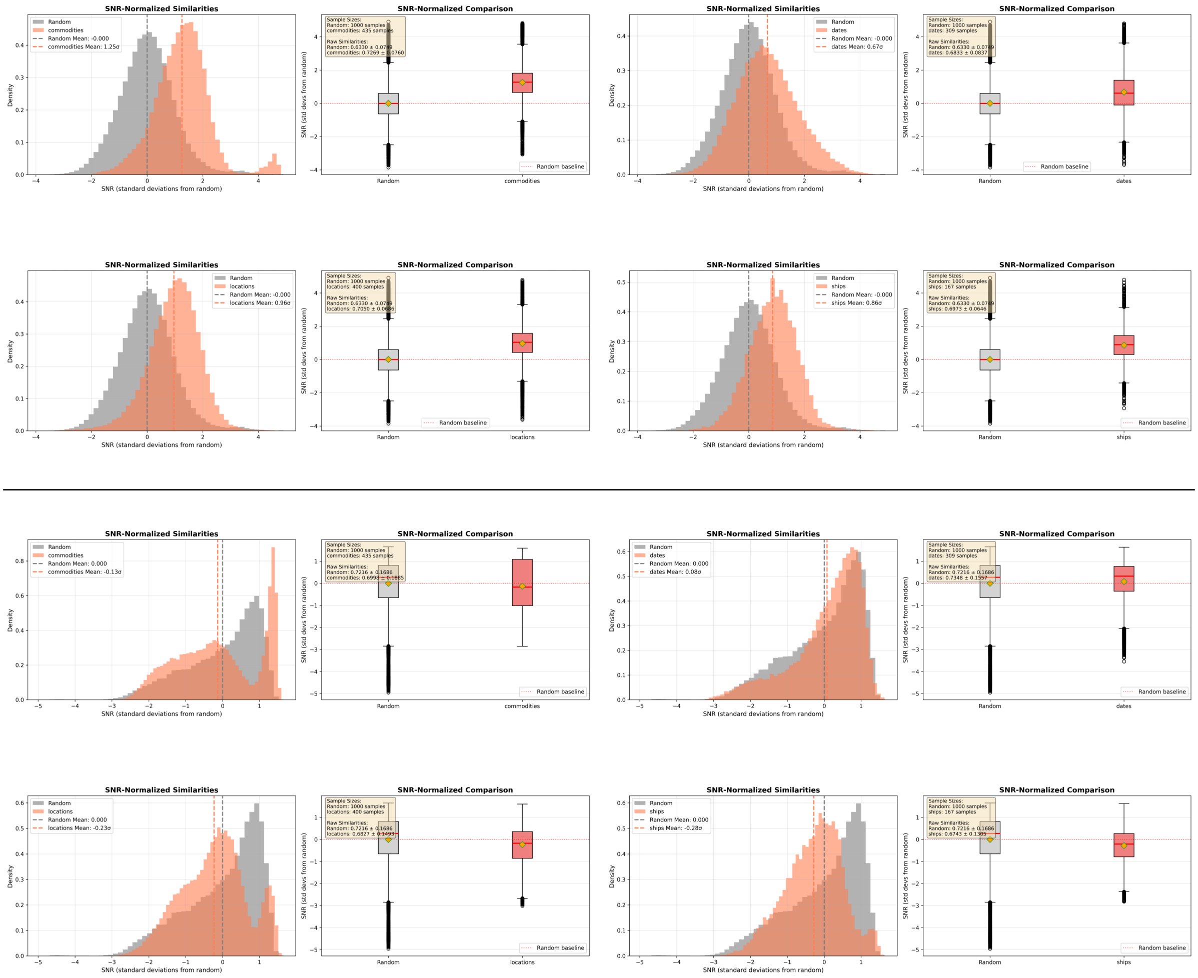

Figure 3: Plots of SNR values for **entities** probed from the last layer of **GloBERTise** base model (above the midline) and the last layer of GloBERTise fine-tuned for SRL (underneath the midline)

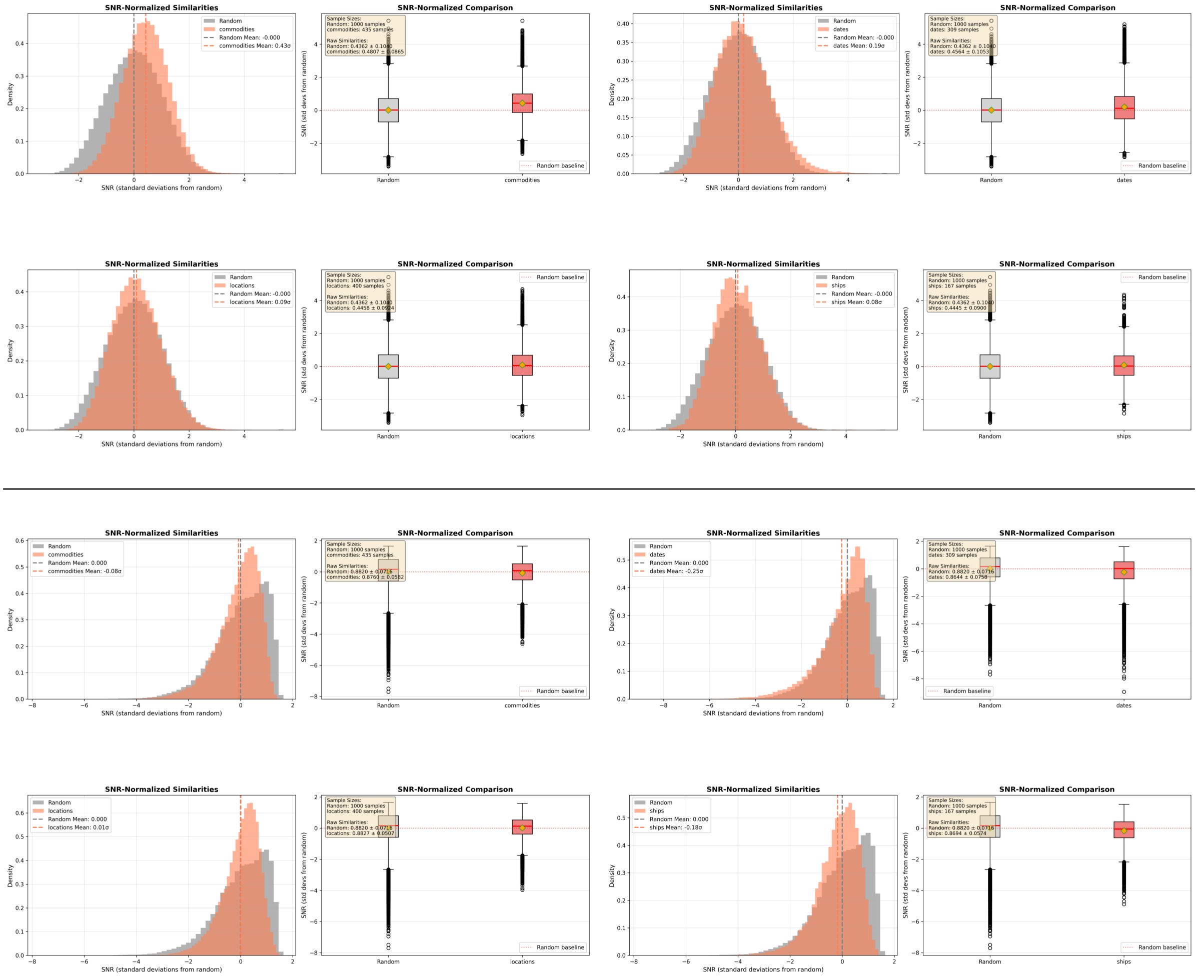

Figure 4: Plots of SNR values for **entities** probed from the last layer of **mBERT** base model (above the midline) and the last layer of mBERT fine-tuned for SRL (underneath the midline)

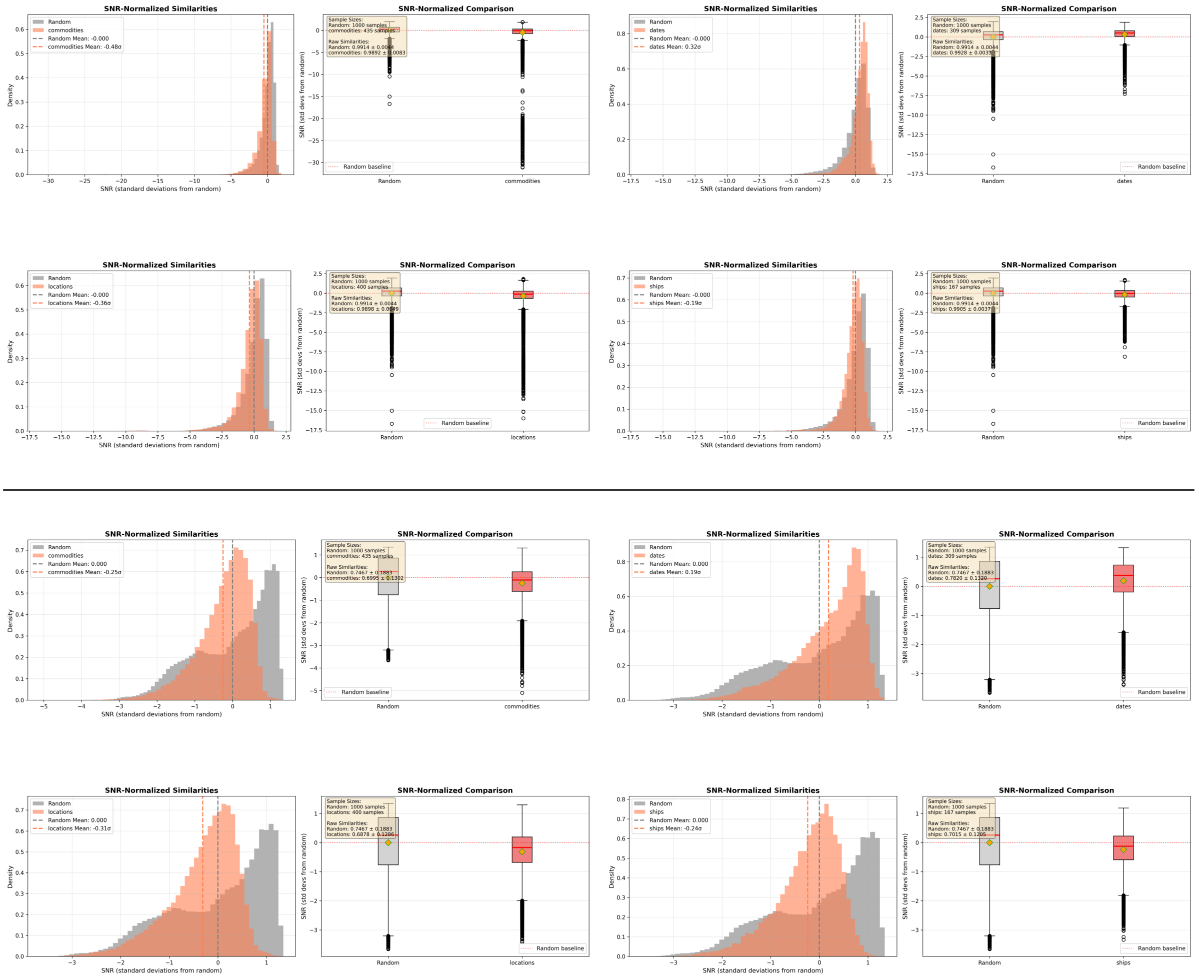

Figure 5: Plots of SNR values for **entities** probed from the last layer of **XLM-R** base model (above the midline) and the last layer of XLM-R fine-tuned for SRL (underneath the midline)

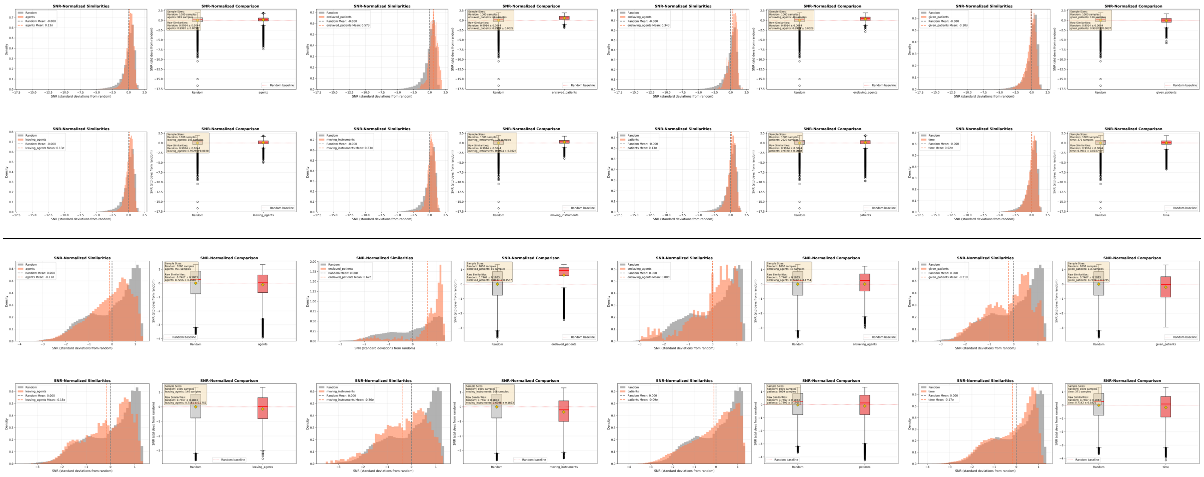

Figure 6: Plots of SNR values for **semantic roles** probed from the last layer of **XLM-R** base model (above the midline) and the last layer of XLM-R fine-tuned for SRL (underneath the midline)

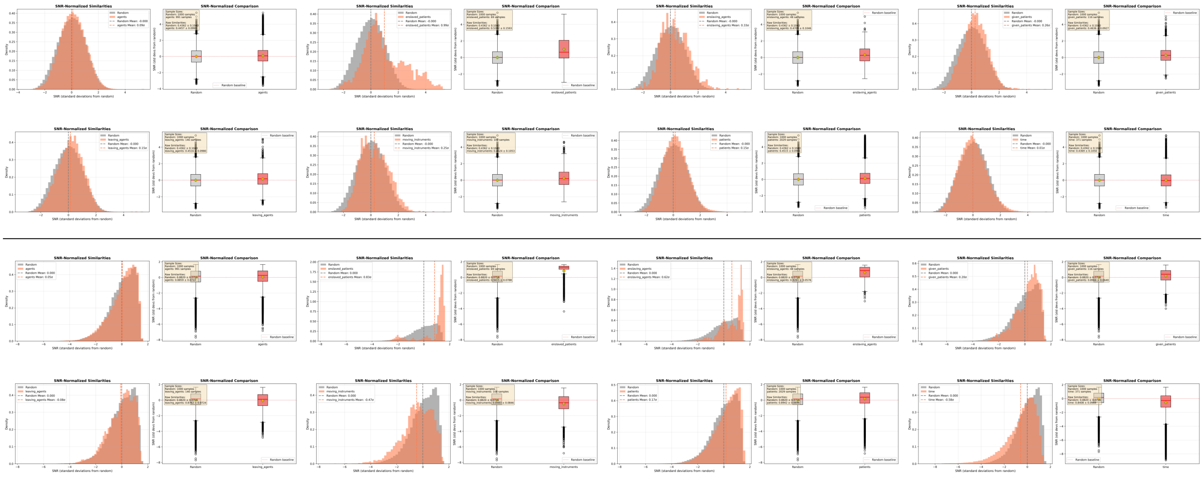

Figure 7: Plots of SNR values for **semantic roles** probed from the last layer of **mBERT** base model (above the midline) and the last layer of mBERT fine-tuned for SRL (underneath the midline)

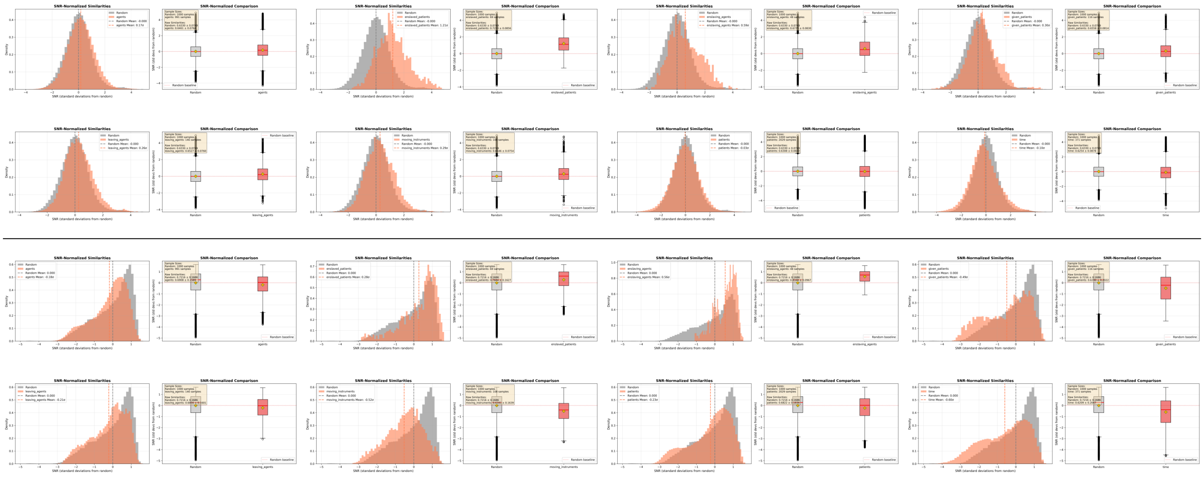

Figure 8: Plots of SNR values for **semantic roles** probed from the last layer of **GloBERTise** base model (above the midline) and the last layer of GloBERTise fine-tuned for SRL (underneath the midline)

Appendix C: Additional similarity matrices accompanying Section 7

	olie	olij	oliphanten	eliphanten	contanten	comptanten	pistoolen	oologsmunitie
olie	1.00	0.65	0.56	0.56	0.66	0.56	0.65	0.55
o-lij	0.65	1.00	0.72	0.72	0.66	0.69	0.65	0.73
o-lip-han-ten	0.56	0.72	1.00	1.00	0.65	0.74	0.66	0.80
e-lip-han-ten	0.56	0.72	1.00	1.00	0.65	0.74	0.66	0.80
contant-en	0.66	0.66	0.65	0.65	1.00	0.74	0.80	0.66
comp-tant-en	0.56	0.69	0.74	0.74	0.74	1.00	0.74	0.73
pistool-en	0.65	0.65	0.66	0.66	0.80	0.74	1.00	0.68
o-olo-g-s-mun-itie	0.55	0.73	0.80	0.80	0.66	0.73	0.68	1.00

Table 8: Cosine similarity matrix for **BERTje**. Higher values (greener) indicate higher similarity.

	olie	olij	oliphanten	eliphanten	contanten	comptanten	pistoolen	oologsmunitie
olie	1.00	0.99	0.78	0.78	0.70	0.76	0.63	0.60
olij	0.99	1.00	0.78	0.78	0.71	0.75	0.64	0.60
ol-iphanten	0.78	0.78	1.00	0.91	0.73	0.73	0.76	0.75
el-iphanten	0.78	0.78	0.91	1.00	0.70	0.74	0.72	0.69
cont-anten	0.70	0.71	0.73	0.70	1.00	0.83	0.73	0.69
comp-tanten	0.76	0.75	0.73	0.74	0.83	1.00	0.68	0.66
p-ist-oolen	0.63	0.64	0.76	0.72	0.73	0.68	1.00	0.78
ool-og-sm-unitie	0.60	0.60	0.75	0.69	0.69	0.66	0.78	1.00

Table 9: Cosine similarity matrix for **GloBERTise**. Higher values (greener) indicate higher similarity.