Disentangling the Linguistic Competence of Privacy-Preserving BERT

Stefan Arnold, Nils Kemmerzell and **Annika Schreiner**
Friedrich-Alexander-Universität Erlangen-Nürnberg
Lange Gasse 20, 90403 Nürnberg, Germany
(stefan.st.arnold, nils.kemmerzell, annika.schreiner)@fau.de

Abstract

Differential Privacy (DP) has been tailored to address the unique challenges of text-to-text privatization. However, text-to-text privatization is known for degrading the performance of language models when trained on perturbed text. Employing a series of interpretation techniques on the internal representations extracted from BERT trained on perturbed pre-text, we intend to disentangle at the linguistic level the distortion induced by differential privacy. Experimental results from a representational similarity analysis indicate that the overall similarity of internal representations is substantially reduced. Using probing tasks to unpack this dissimilarity, we find evidence that text-to-text privatization affects the linguistic competence across several formalisms, encoding localized properties of words while falling short at encoding the contextual relationships between spans of words.

1 Introduction

Language Models (LM) (Devlin et al., 2018; Radford et al., 2018) are among the most successful applications of machine learning and applied in a diverse range of tasks such as classification, translation, summarization, and question answering.

However, concerns were raised that LMs (Carlini et al., 2019; Pan et al., 2020) in general and their embedding layers (Song and Raghunathan, 2020; Thomas et al., 2020) in particular memorize and disclose personally identifiable information.

To mitigate the risk of information leakage due to unintended memorization, *Differential Privacy* (DP) (Dwork et al., 2006) has been integrated into machine learning (Abadi et al., 2016) and LMs (McCann et al., 2017; Shi et al., 2022). DP formalizes privacy through a notion of indistinguishability which is accomplished by injecting additive noise.

While early adaptations of DP into LMs were applied to gradient updates (McMahan et al., 2017), there is a shift towards applying DP on raw text (Fernandes et al., 2019; Feyisetan et al., 2020; Qu et al., 2021) in the form of text-to-text privatization. This technique aims to provide plausible deniability (Bindschaedler et al., 2017) by perturbing words in a way that conceals authors and content.

Qu et al. (2021) applied text-to-text privatization to BERT (Devlin et al., 2018) and explored techniques for privacy-adaptive pre-training (*e.g.*, predicting a set of perturbed tokens for each masked position) and privacy-constrained fine-tuning. We complement this research direction by borrowing from range of techniques for model introspection to identify and localize the layer-wise alterations caused by perturbed text on internal representations and associate these with the retention and destruction of linguistic competence.

Drawing on a representational similarity analysis (Kriegeskorte et al., 2008), we measure a substantial dissimilarity between internal representations obtained from different privacy modalities. To connect this dissimilarity with linguistic formalisms, we conduct a series of probing tasks (Adi et al., 2016; Tenney et al., 2019b; Hewitt and Manning, 2019). By contrasting the probing accuracies for recovering a range of twelve linguistic formalisms, we uncover that linguistic formalisms relying on localized properties endure the perturbations introduced by text-to-text privatization while properties that require context information are less resilient.

Since internal representations of LMs are formed by an attention mechanism (Vaswani et al., 2017), we further investigate the distribution of attention patterns. By clustering the attention maps (Clark et al., 2019), we uncover that text-to-text privatization amplifies redundancy (Kovaleva et al., 2019).

2 Preliminaries

2.1 Language Models

Language Models (LMs) convert sentences composed of variable-length sequences of discrete tokens, such as *characters*, *subwords*, or *words*, into

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 65–75
December 7, 2023. ©2023 Association for Computational Linguistics

fixed-length continuous embeddings.

The introduction of the *Transformer* architecture (Vaswani et al., 2017) and variants based solely on a encoder (Devlin et al., 2018) or decoder (Radford et al., 2019) rapidly replaced recurrent architectures (Peters et al., 2018a). By relying entirely on a self-attention mechanism, transformers excel at modeling long-range interactions within text.

We focus on BERT (Devlin et al., 2018) with an uncased vocabulary, which exemplifies a family of transformers that produce bidirectional representations solely from the encoder block (Lan et al., 2019; Sanh et al., 2019; Liu et al., 2019b).

The conventional workflow for BERT consists of two stages: *pre-training* and *fine-tuning*. During pre-training, BERT is trained on a pre-text corpus using masked language modeling (prediction of randomly masked words) and next sentence prediction (binarized prediction whether text pairs are adjacent). Fine-tuning involves adding a fully-connected layer trained end-to-end on labeled data, allowing BERT to adapt to various task related to language understanding (Wang et al., 2018).

The internals of BERT comprise an embedding layer and multiple transformer layers. Once a text is tokenized into wordpieces (Wu et al., 2016), the embedding layer serves as a lookup table that contains a lexical representation for each token. Since BERT processes all token representations in parallel, the lexical representations need to be integrated with position and segment information. The transformer layers build on an attention mechanism that computes a scalar attention weight between each ordered pair of tokens and uses this weight to control the contextualization from every token regardless of its position or segment. Contextual representations together with attention maps provide the starting point for interpreting linguistic properties captured during pre-training (Tenney et al., 2019a) and retained after fine-tuning (Merchant et al., 2020).

2.2 Differential Privacy

Differential Privacy (DP) (Dwork et al., 2006) transitioned from the field of statistical databases into machine learning (Song et al., 2013; Bassily et al., 2014; Abadi et al., 2016; Shi et al., 2022). DP operates on the principle of injecting additive noise so that model outputs are indistinguishable within the bounds of a privacy budget $\varepsilon > 0$, where $\varepsilon \to \infty$ represents no bound on the information leakage.

Table 1: Example chunk (truncated) from Wikipedia privatized with different privacy budgets. Highlighted words represent a mismatch between the original word and the surrogate word after privatization.

ε	**Example**
∞	’anarchism’, ’is’, ’a’, ’political’, ’philosophy’, ’and’, ’movement’, ’that’, ’is’, ’skeptical’, ’of’, ’authority’, ’and’, ’rejects’, ’all’, ’involuntary’, ’,’, ’coercive’, ’forms’, ’of’, ’hierarchy’, ’.’
10	’syndicalism’, ’situated’, ’a’, ’political’, ’pedagogy’, ’but’, ’movement’, ’that’, ’help’, ’signalled’, ’the’, ’recommendation’, ’18’, ’rejects’, ’four’, ’mobility’, ’,’, ’punitive’, ’forms’, ’on’, ’associations’, ’outset’

Equipped with a discrete vocabulary set $\mathcal{W}$, an embedding function $\phi : \mathcal{W} \to \mathbb{R}$, and a distance metric $d : \mathbb{R} \times \mathbb{R} \to [0, \infty)$, Feyisetan et al. (2020) formulated a randomized mechanism for text-to-text privatization grounded in metric differential privacy (Chatzikokolakis et al., 2013). Specifically, the randomized mechanism perturbs each word in a text by adding noise to the representation of the word derived from an embedding space (Mikolov et al., 2013) and projecting the noisy representation back to a discrete vocabulary using a nearest neighbor search. Since metric differential privacy scales the notion of indistinguishability by a distance $d(\cdot)$, this technique offers several benefits: (1) It ensures that the log-likelihood ratio of observing any substitution $\hat{w}$ given two words w and w' is bounded by $\varepsilon d\{\phi(w), \phi(w')\}$, providing plausible deniability (Bindschaedler et al., 2017) with respect to all $w \in \mathcal{W}$. (2) It produces similar substitutions $\hat{w}$ for any words w and w' that are close in the embedding space, alleviating the curse of dimensionality associated with randomized response (Warner, 1965).

Table 1 illustrates an example output obtained by querying the randomized mechanism for text-to-text privatization. Notice that the fidelity to the original text is proportional to the privacy budget. However, the example also shows that text-to-text privatization suffers from many constraints such as grammatical errors (Mattern et al., 2022), which spawned further developments aimed at improving both utility (Yue et al., 2021; Arnold et al., 2023; Chen et al., 2023) and privacy (Xu et al., 2020).

2.3 Model Introspection

Aimed at understanding the internals of language models, numerous interpretation techniques were developed to uncover which properties of a text are embedded in contextual representations. Prominent techniques include stimuli and diagnostic models.

Stimuli-based Probes. Linzen et al. (2016) assembled texts containing curated stimuli and evaluated the perplexity scores on masked stimuli as evidence for the presence or absence of linguistic knowledge. Using a fill-mask objective on stimuli was adopted to examine a range of linguistic properties, in particular *subject-verb agreement* (Gulordava et al., 2018; Marvin and Linzen, 2018; Lakretz et al., 2019; Goldberg, 2019; Ettinger, 2020).

Classifier-based Probes. Adi et al. (2016) eliminated the need for curating stimuli by setting up probing models. A probing model inputs internal representations as features annotated by linguistic properties of interest as labels and its accuracy score is directly interpreted as the extent to which linguistic properties are contained in the internal representation. Since probing models require few assumptions beyond the existence of model activations, they are widely used to assess the linguistic competence of language models (Belinkov et al., 2017; Conneau et al., 2018; Hupkes et al., 2018).

Considerable research is centered on the inspection of fixed-length sentence representations. Adi et al. (2016) introduced a probing suite to extract surface properties of sentences such as *length*, *content*, and *order*. Conneau et al. (2018) later recasted and extend these probing tasks by a broader set of linguistic properties, such as *tense* and *depth*.

Contrary to probing fixed-length sentence representations, probing suits exist that are tailored towards linguistic properties in word-level representations (Blevins et al., 2018; Peters et al., 2018b; Tenney et al., 2019b; Liu et al., 2019a). Tenney et al. (2019b) present *edge probing* in which a diagnostic model is given access only to span representations. From these span representations, the probing model aims to extract high-level linguistic properties which are expected to require complete sentence context. The analysis of intermediate layers of language models indicates that linguistic properties are captured in a hierarchical order (Peters et al., 2018b; Tenney et al., 2019a; Jawahar et al., 2019). This hierarchy is composed of signals ranging from surface abstractions in the lower layers, syntactic abstractions in the middle layers and semantic abstractions in the higher layers.

While prior probes on detecting syntactic structure lacked an explanation of whether structure is embedded as an entire parse tree (Conneau et al., 2018) or how such parse trees are embedded (Peters et al., 2018b), Hewitt and Manning (2019) proposed a *structural probe* to recover the topology of an entire parse tree and derive its parse depth. Using a linear transformation of the representation space, the structural probe shows evidence of a geometric representation that implicitly embeds sentence structure. The structural hypothesis formed by the linear transformation has recently been refined by a scaled isomorphic rotation (Limisiewicz and Mareček, 2020), kernelization using a radial-basis function (White et al., 2021), and projection onto hyperbolic space (Chen et al., 2021).

To examine how contextual representations are formed through the attention mechanism (Vaswani et al., 2017), recent research extended their analysis to role of attention in handling properties of text (Lin et al., 2019; Jo and Myaeng, 2020). The visualization of attention heatmaps and the calculation of the distribution of attention revealed interpretable positional patterns (Vig and Belinkov, 2019; Clark et al., 2019; Kovaleva et al., 2019) and strong correlations to linguistic properties (Clark et al., 2019; Htut et al., 2019; Ravishankar et al., 2021).

Limitations. Despite its popularity for model introspection, recent studies observed that linguistic properties are incidentally captured even without task relevance (Ravichander et al., 2020), casting doubt on the interpretations derived from attention maps (Jain and Wallace, 2019; Serrano and Smith, 2019; Brunner et al., 2019) and probing models (Tamkin et al., 2020). This prompted the design of control tasks (Hewitt and Liang, 2019; Ravichander et al., 2020), amnesic probing (Elazar et al., 2021; Jacovi et al., 2021), conditional probing (Hewitt et al., 2021), and orthogonal techniques for correlating contextual representations (Saphra and Lopez, 2018; Voita et al., 2019; Abdou et al., 2019).

3 Methodology

We follow the convention of denoting words and sentences using italic (w_i, s), and refer to their representations using bold $(\mathbf{w_i}, \mathbf{s})$, where the index i distinguishes words in a sentence. Let d be the dimension of a l-layer LM. Given a sentence s as a tokenized list of words $w \in \mathcal{W}$, the LM inputs a lexical vector representation for each word and computes a contextual vector representation $\mathbf{w}_i^l \in \mathbb{R}^d$ for the i-th word at the l-th layer.

We pre-train BERT models from-scratch following Devlin et al. (2018) on a dump of Wikipedia preprocessed with a privacy budget of $\epsilon \in \{10, \infty\}$, where 10 yields a privacy-preserving BERT and ∞ serves as our baseline for comparison. Apart from the difference in the privacy modality, training is identical to erase any confounding factors.

Equipped with BERT pre-trained on a corpus of Wikipedia with different privacy modalities, we intend to uncover how and where contextual representations produced by the model trained with differential privacy depart from those produced by the model trained without differential privacy. Following the experimental setup of Merchant et al. (2020), we address this question mainly through the lens of (unsupervised) representational similarity analysis and (supervised) probing models.

3.1 Similarity Analysis

We aim to compare the internals of language models that originate from pre-training under public and private training environments. Due to the lack of correspondence between activation patterns of models trained with different modalities, we need to abstract away from direct comparison of model activations. We instead leverage *Representational Similarity Analysis* (RSA) (Kriegeskorte et al., 2008) to correlate the dissimilarity structure between contextual representations. Building on dissimilarity structures rather than activation patterns, RSA is indifferent to the representation space.

We base our similarity analysis on higher-order comparisons introduced by Abdou et al. (2019). Given a set of language models trained under different (privacy) modalities M and a common set of sentences N, we extract representations as layer-wise activations from each M. Using any kernel that satisfies the axioms of a (dis)similarity metric, we can convert the extracted representations into pairwise dissimilarity matrices $\mathbb{R}^{n \times n}$. Each $N \times N$ dissimilarity matrix corresponds to the dissimilarity between the activation patterns associated with sentences pairs $n_i, n_j \in N$. Since the dissimilarity is intuitively zero when $n_i = n_j$, the dissimilarity matrix is symmetric along a diagonal. Using another kernel, we can now correlate the similarity between the flattened upper triangulars of the constructed dissimilarity matrices.

We adopt the *Cosine distance* as metric for the *intra*-space dissimilarity and *Spearman correlation* as metric for the *cross*-space similarity. The RSA is performed on a random subset of $5,000$ sentences drawn from WikiText (Merity et al., 2016).

3.2 Linguistic Probing

We aim to connect the dissimilarity between contextual representations with linguistic properties. To discern and locate the extent to which linguistic properties of texts are captured, we employ probing tasks at word-level and sentence-level representations for a range of surface, syntactic, and semantic formalisms. Note that BERT uses tokenization into subwords. Since word-level probes require access to word representations, we map subword representations to word representations by element-wise mean pooling over all subword components.

Surface Probe. We evaluate surface properties using the setup for sentence-level probing assembled by Adi et al. (2016). To form sentence representations $\mathbf{s} \in \mathbb{R}^d$, we use element-wise mean pooling. Without access to a sentence s and any of its words w, the surface proprieties to extract are *length*, *content*, and *order*. The length task measures to what extent a sentence representation $\mathbf{s}$ encodes the length $|s|$ of a sentence s. The length task is formulated as a multi-class classification for a balanced set of binned lengths in intervals $[0, 35)$, $[35, 41)$, $[41, 46)$, $[46, 52)$, $[52, \infty)$. The content task measures the extent to which a sentence representation $\mathbf{s}$ encodes the identities of words w in a sentence. The content task is formulated as a binary classification in the form $(\mathbf{s}, \mathbf{w}) \in \{0, 1\}$, where 0 denotes $w \notin s$ and 1 denotes $w \in s$, respectively. The order task measures the extent to which a sentence representation $\mathbf{s}$ encodes the order of words w_i, w_j. Given a sentence representation $\mathbf{s}$ and two word representation $\mathbf{w_i}$, $\mathbf{w_j}$ of words appearing in a sentence, the content task is formulated as a binary classification in the form $(\mathbf{s}, \mathbf{w_i}, \mathbf{w_j}) \in \{0, 1\}$, where 0 denotes $\mathbf{w_i} \prec \mathbf{w_j}$ and 1 denotes $\mathbf{w_i} \succ \mathbf{w_j}$, respectively. All surface probes are performed on sentences from the training set reflecting their presumably most accurate representations.

Linguistic Probe. To evaluate linguistic properties , we employ *edge probes* (Tenney et al., 2019b) and *structural probes* (Hewitt and Manning, 2019) as two complementary probes at word-level.

The purpose of edge probing is to measure the extent to which contextual representations capture syntactic dependencies and semantic abstractions. Instead of supplying a probing model with a pooled sentence representation $\mathbf{s}$, edge probing decomposes the probing task into a common format so that the probing model only receives labeled spans $[\mathbf{w}_i^l, \mathbf{w}_j^l)$ and (optionally) $[\mathbf{w}_u^l, \mathbf{w}_v^l)$. With access only to contextual representations within the end-exclusive spans, the probing model must label the relation between these spans and their role in the sentence. Derived from evaluation on tagged benchmark datasets, we report the micro-averaged harmonic mean of the precision and recall for labeling *part-of-speech tags*, *constituency phrases*, *dependency relations* as syntactic tasks, and *entity types*, *entity relations*, *semantic roles*, and *coreference mentions* as semantic tasks.

The structural probe is designed to measure the representation of syntactic structure. The probe identifies whether the geometric space under linear transformation $B \in \mathbb{R}^{k \times d}$, where k is the rank of the transformation and d is the dimensionality of the representation, captures the depth of words or distances between words in a parse tree. We adjust the rank to the dimensionality $k = d$. The *depth* probe measures the distance from root $\forall i$ in a parse tree. It is defined by $\|\mathbf{w}_i^l\|_B = (B\mathbf{w}_i^l)^T(B\mathbf{w}_i^l)$. The depth probe is evaluated based on the accuracy of the root word and the correlation between the predicted order of words and ordering specified by the depth in the parse tree. The *distance* probe measures the pairwise distances $\forall i, j$ within a parse tree. It is defined by $\|\mathbf{w}_i^l - \mathbf{w}_j^l\|_B = (B(\mathbf{w}_i^l - \mathbf{w}_j^l)^T(B(\mathbf{w}_i^l - \mathbf{w}_j^l))$. The distance probe is evaluated by correlating the predicted distances between pairs of words with distances metrics specified by the parse tree and by converting the predicted distances between pairs of words into a minimum spanning tree and scoring it against the parse tree using the Undirected Unlabeled Attachment Score (UUAS).

4 Experiments

We initiate our model introspection by examining the performance in terms of perplexity scores. Figure 1 reveals that BERT trained on a corpus of text subjected to text-to-text privatization converges to a notably (but reasonably) worse perplexity score at 61.45 (compared to 6.82). Since perplexity is a measure for assessing the proficiency of language models in predicting the next word in a sentence, the elevated value in this context connotes a diminished ability for language modeling. To elucidate the linguistic alterations that lead to the degradation of the perplexity score, we pursue a layer-wise ablation of linguistic properties captured in the internal representations of privacy-preserving BERT.

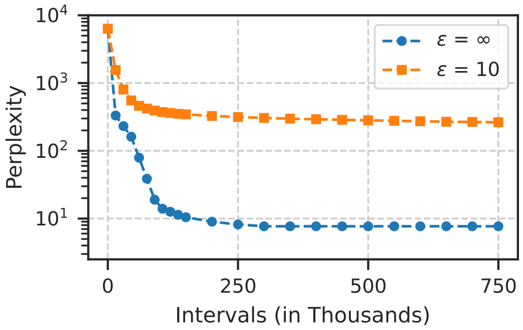


Figure 1: Interval-wise learning progress of BERT from $26,903,298$ chunks generated from Wikipedia.

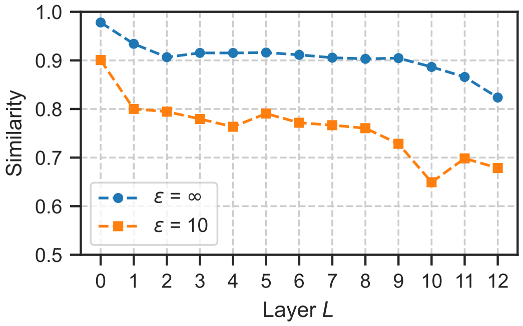
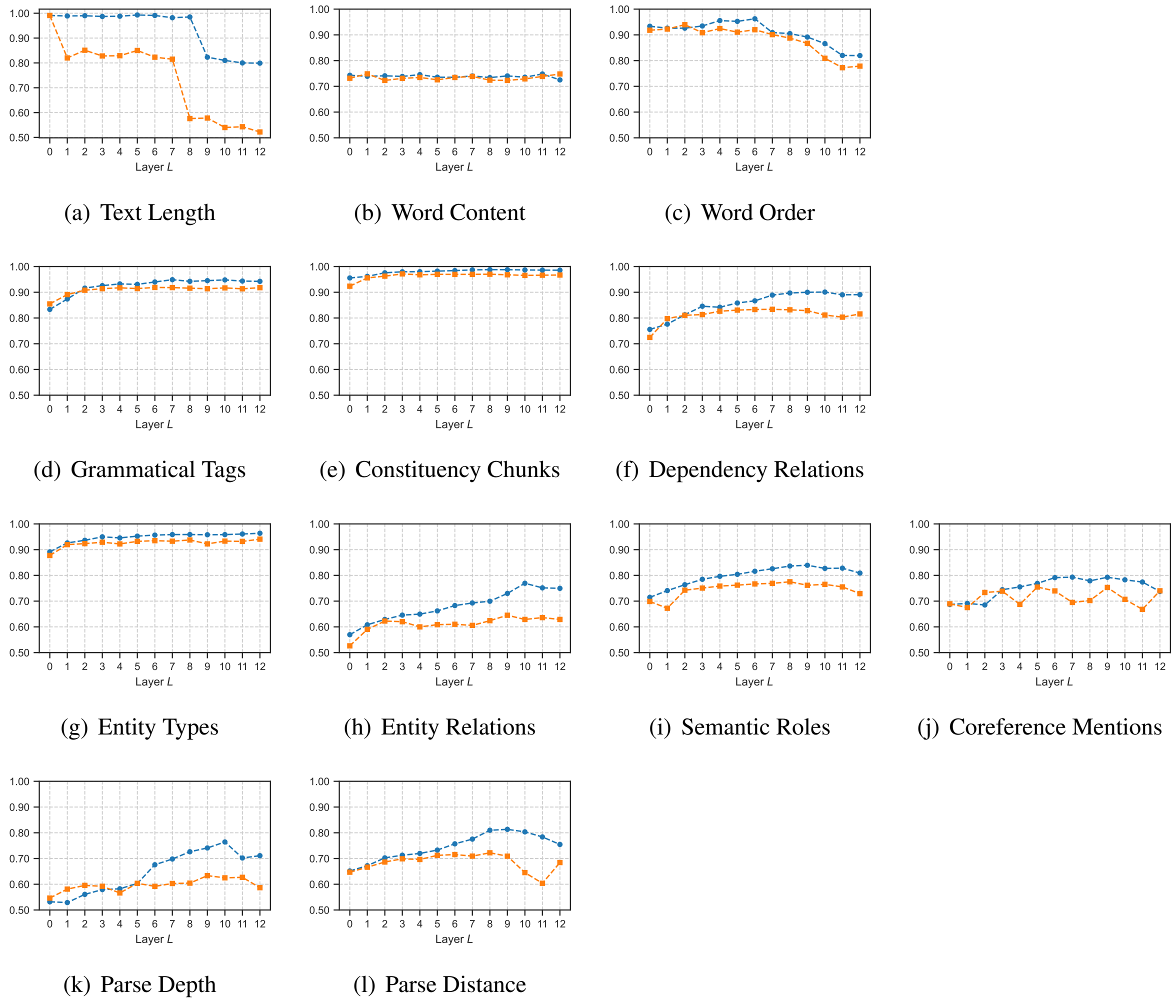


Figure 2: Layer-wise representational similarity of BERT for $5,000$ samples randomly drawn from WikiText.

4.1 Similarity Results

In line with correlation coefficients, RSA scores have value range of $[-1, +1]$, where $+1$ indicates that the models produce a similar internal representation and -1 indicates that the models diametrically opposed in latent space. Since these theoretical bounds are unlikely in practice, we establish an empirical bound on RSA by correlating the dissimilarity structures of BERT models with identical architecture but different initialization. We observe that the average similarity bounds at 0.9051. By correlating the dissimilarity structures between BERT and BERT trained on perturbed text, we find a remarkable drop to 0.7601, signifying a substantial departure between their internal representations.

To locate the variations in the internal representations on different layers of the BERT architecture, we present the layer-wise RSA results in Figure 2. Note that BERT models typically maintain consistently high RSA values across all layers, whereas



(a) Text Length

(b) Word Content

(c) Word Order

(d) Grammatical Tags

(e) Constituency Chunks

(f) Dependency Relations

(g) Entity Types

(h) Entity Relations

(i) Semantic Roles

(j) Coreference Mentions

(k) Parse Depth

(l) Parse Distance

Figure 3: Layer-wise probing results for BERT under public (blue circles) and private (orange squares) training modalities. Surface properties according to Adi et al. (2016) are depicted in Figures 3(a), 3(b), and 3(c). Syntactic properties according to Tenney et al. (2019b) are depicted in Figures 3(d), 3(e), and 3(f). Semantic properties according to Tenney et al. (2019b) are depicted in Figures 3(g), 3(h), 3(i), and 3(j). Structural properties according to Hewitt and Manning (2019) are depicted in Figures 3(k) and 3(l).

our BERT model trained on perturbed text starts with relatively high RSA values at the lexical representation layer at 0.9007 and declines with contextual representations layers to 0.6784, indicating a sharper deviation in the representation space. This pattern carries significant implications for our understanding of the impact of text-to-text privatization. Since the lexical representation corresponds to occurrence characteristics, this indicates that private BERT fails to capture context information.

4.2 Probing Results

Assuming that the substantial divergence arises from the fact that privacy-preserving BERT forms its contextual representation based on different linguistic properties than BERT, we are interested in discovering which linguistic properties are captured despite being trained on perturbed text.

Figure 3 depicts the probing results. The layer-wise probing results are shaped similarly but the consistently lower scores across all properties indicate that the linguistic competence is compromised when text-to-text privatization is are applied.

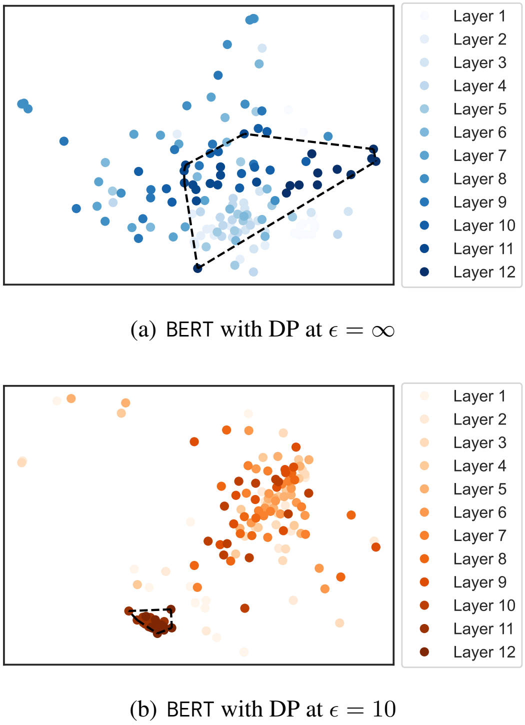
Surface. Starting from the sentence-level probes, we notice distinct patterns in the details captured about surface properties. With a deficit of -0.2770, there is a marked difference related to the encoded text length. Contrasting this deficiency, details concerning content and order show a higher degree of consistency, reflecting deviations of $+0.0230$ and -0.0410, respectively. To grasp the implications of surface properties, we recall the argumentation of Adi et al. (2016) that representations containing information about length and order are more suited for syntactic tasks while representations that excel at content are more suited for semantic tasks.

Linguistic. We continue with linguistic properties at word-level. From syntactic probes, we observe that a significant portion of information about grammatical tags and constituency chunks are retained at -0.0246 and -0.0187, while less emphasis is placed on capturing dependency relations, resulting in a reduction of -0.0751. From semantic probes, we notice that information about entity types is missing by only -0.0229, while entity relations and semantic roles experience a more substantial drop of -0.1209 and -0.0798. From structural probes, which test whether a representation encodes topology, we consolidate the findings from the syntactic probe on dependency relations. Scored against a discrete solution in the form of the root word or minimum spanning tree, the representations contain information about the root word with a score of 0.5866 and the parse tree with a score of 0.6843, representing decrements of -0.1244 and -0.0703, respectively.

Considering the nature of the linguistic properties and the degree to which they decline under privacy constraints, it is noticeable that formalisms closely related to basic characteristics of words display a considerable degree of preservation, whereas formalisms tied to complex relationships within spans of words undergo a substantial degree of deterioration. This intriguing pattern suggests that while localized properties endure the perturbations of text-to-text privatization, the ability of language models to maintain contextual constructs can be severely hindered by text-to-text privatization.

Since text-to-text privatization builds on word-level differential privacy (Mattern et al., 2022), a plausible explanation for this phenomenon could be rooted in the nature of its randomized mechanism, which has been observed to disproportionately affect linguistic properties (Arnold et al., 2023). This insight underscores the interplay between perturbation strategies and the necessity of accurately conveying different types of linguistic formalisms.

Attention. Since contextual representations are mainly formed by the mechanism of self-attention (Vaswani et al., 2017), we could attribute the alterations in the representations to the fact that the attention mechanism (somehow) fails to discriminate certain linguistic properties. We attempt to answer this hypothesis by analyzing the distributional patterns of attention maps.



(a) BERT with DP at $\epsilon = \infty$

(b) BERT with DP at $\epsilon = 10$

Figure 4: Divergence-based clustering of attention maps extracted from $1,000$ random samples of `WikiText`.

Once for each training modality, we obtain attention maps for $1,000$ randomly selected sentences and rearrange the attention maps from their subwords in line with Vig and Belinkov (2019). For attentions drawn to a split-up word, we sum up the attention weights over its subwords. For attentions stemming from a split-up word, we average all weights from its subwords. Following Clark et al. (2019), we calculate the distance between all pairs of attention maps using the Janson-Shannon divergence and visualize the distances grouped by layer using multidimensional scaling in Figure 4.

Assuming that attention heads that are clustered closely together perform similar linguistic roles in forming the internal representation, we conclude from the distributional patterns that text-to-text privatization amplifies the redundancy that is already present in attention heads as revealed by Kovaleva et al. (2019). This is most evident by comparing the overlap of the attention maps in rear layers.

Considering that Li et al. (2018) showed that encouraging the attention mechanism to have diverse behaviors can improve performance, we find another possible explanation for the lack of linguistic competence in privacy-preserving language models and their deteriorated level of perplexity.

5 Conclusion

Assuming that the performance loss of language models caused by text-to-text privatization can be attributed to the destruction of linguistic competence (Merendi et al., 2022), we set to disentangle the layer-wise alterations of perturbations to the internal representations of a language model.

By employing a series of techniques for model introspection (Adi et al., 2016; Hewitt and Manning, 2019; Tenney et al., 2019b), we tested the internal representations formed by language models for linguistics properties across several formalisms.

From the perspective of linguistic competence, experimental results from our layer-wise model introspection indicate that privacy preservation can considered conservative as language models subjected to text-to-text privatization retain a hierarchical order of linguistic formalisms (Peters et al., 2018b; Tenney et al., 2019a; Jawahar et al., 2019). However, text-to-text privatization shows to have a cumulative impact on the linguistic competence of language models, affecting aspects ranging from surface-level properties to linguistic constructs across syntactic, semantic, and structural formalisms. We further notice that basic properties of words are less disrupted than complex relations between words that require context information.

Limitations. Most assumptions and findings of this study are grounded in probing. Although probing enjoys much support as a technique for interpreting the internals of language models (Abadi et al., 2016; Conneau et al., 2018; Tenney et al., 2019b; Hewitt and Manning, 2019), recent studies dispute with conclusion derived from probing due to the fact that probing may not entail task relevance (Ravichander et al., 2020). We side with those viewing probing as a tool for model introspection, but nonetheless caution that our probing results may not be the appropriate technique for discerning the differences of private training modalities. Given the wide range of probing tasks and the fact that our probing results show a consistent pattern of competencies, we are convinced that this study contributes novel privacy implications.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318.

Mostafa Abdou, Artur Kulmizev, Felix Hill, Daniel M Low, and Anders Søgaard. 2019. Higher-order comparisons of sentence encoder representations. *arXiv preprint arXiv:1909.00303*.

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

Stefan Arnold, Dilara Yesilbas, and Sven Weinzierl. 2023. Guiding text-to-text privatization by syntax. In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 151–162, Toronto, Canada. Association for Computational Linguistics.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th annual symposium on foundations of computer science*, pages 464–473. IEEE.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Vincent Bindschaedler, Reza Shokri, and Carl A Gunter. 2017. Plausible deniability for privacy-preserving data synthesis. *arXiv preprint arXiv:1708.07975*.

Terra Blevins, Omer Levy, and Luke Zettlemoyer. 2018. Deep rnns encode soft hierarchical syntax. *arXiv preprint arXiv:1805.04218*.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2019. On identifiability in transformers. *arXiv preprint arXiv:1908.04211*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. 2019. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX Security Symposium (USENIX Security 19)*, pages 267–284.

Konstantinos Chatzikokolakis, Miguel E Andrés, Nicolás Emilio Bordenabe, and Catuscia Palamidessi. 2013. Broadening the scope of differential privacy using metrics. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 82–102. Springer.

Boli Chen, Yao Fu, Guangwei Xu, Pengjun Xie, Chuanqi Tan, Mosha Chen, and Liping Jing. 2021. Probing bert in hyperbolic spaces. *arXiv preprint arXiv:2104.03869*.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. 2023. A customized text sanitization mechanism with differential privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5747–5758, Toronto, Canada. Association for Computational Linguistics.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

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

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Natasha Fernandes, Mark Dras, and Annabelle McIver. 2019. Generalised differential privacy for text document processing. In *International Conference on Principles of Security and Trust*, pages 123–148. Springer, Cham.

Oluwaseyi Feyisetan, Borja Balle, Thomas Drake, and Tom Diethe. 2020. Privacy-and utility-preserving textual analysis via calibrated multivariate perturbations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 178–186.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

John Hewitt, Kawin Ethayarajh, Percy Liang, and Christopher D Manning. 2021. Conditional probing: measuring usable information beyond a baseline. *arXiv preprint arXiv:2109.09234*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *arXiv preprint arXiv:1909.03368*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R Bowman. 2019. Do attention heads in bert track syntactic dependencies? *arXiv preprint arXiv:1911.12246*.

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

Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. Contrastive explanations for model interpretability. *arXiv preprint arXiv:2103.01378*.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

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

Jae-young Jo and Sung-Hyon Myaeng. 2020. Roles and utilization of attention heads in transformer-based neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3404–3417.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of bert. *arXiv preprint arXiv:1908.08593*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:4.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. *arXiv preprint arXiv:1810.10183*.

Tomasz Limisiewicz and David Mareček. 2020. Introducing orthogonal constraint in structural probes. *arXiv preprint arXiv:2012.15228*.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: getting inside bert's linguistic knowledge. *arXiv preprint arXiv:1906.01698*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Nelson F Liu, Matt Gardner, Yonatan Belinkov, Matthew E Peters, and Noah A Smith. 2019a. Linguistic knowledge and transferability of contextual representations. *arXiv preprint arXiv:1903.08855*.

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

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

Justus Mattern, Benjamin Weggenmann, and Florian Kerschbaum. 2022. The limits of word level differential privacy. *arXiv preprint arXiv:2205.02130*.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. *Advances in neural information processing systems*, 30.

H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2017. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963*.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to bert embeddings during fine-tuning? *arXiv preprint arXiv:2004.14448*.

Federica Merendi, Felice Dell'Orletta, and Giulia Venturi. 2022. On the nature of bert: Correlating fine-tuning and linguistic competence. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3109–3119.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1314–1331. IEEE.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. Deep contextualized word representations.

Matthew E Peters, Mark Neumann, Luke Zettlemoyer, and Wen-tau Yih. 2018b. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*.

Chen Qu, Weize Kong, Liu Yang, Mingyang Zhang, Michael Bendersky, and Marc Najork. 2021. Natural language understanding with privacy-preserving bert. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1488–1497.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2020. Probing the probing paradigm: Does probing accuracy entail task relevance? *arXiv preprint arXiv:2005.00719*.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention can reflect syntactic structure (if you let it). *arXiv preprint arXiv:2101.10927*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Naomi Saphra and Adam Lopez. 2018. Understanding learning dynamics of language models with svcca. *arXiv preprint arXiv:1811.00225*.

Sofia Serrano and Noah A Smith. 2019. Is attention interpretable? *arXiv preprint arXiv:1906.03731*.

Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. 2022. Selective differential privacy for language modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2848–2859, Seattle, United States. Association for Computational Linguistics.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*, pages 377–390.

Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*, pages 245–248. IEEE.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. *arXiv preprint arXiv:2004.14975*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019a. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*.

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. 2019b. What do you learn from context? probing for sentence structure in contextualized word representations. *arXiv preprint arXiv:1905.06316*.

Aleena Thomas, David Ifeoluwa Adelani, Ali Davody, Aditya Mogadala, and Dietrich Klakow. 2020. Investigating the impact of pre-trained word embeddings on memorization in neural networks. In *International Conference on Text, Speech, and Dialogue*, pages 273–281. Springer.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*.

Elena Voita, Rico Sennrich, and Ivan Titov. 2019. The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives. *arXiv preprint arXiv:1909.01380*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Stanley L Warner. 1965. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.

Jennifer C White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. *arXiv preprint arXiv:2105.10185*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zekun Xu, Abhinav Aggarwal, Oluwaseyi Feyisetan, and Nathanael Teissier. 2020. A differentially private text perturbation method using a regularized mahalanobis metric. *arXiv preprint arXiv:2010.11947*.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. 2021. Differential privacy for text analytics via natural text sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3853–3866, Online. Association for Computational Linguistics.