Differential Privacy, Linguistic Fairness, and Training Data Influence: Impossibility and Possibility Theorems for Multilingual Language Models

Phillip Rust [1] **Anders Søgaard** [1]

Abstract

Language models such as mBERT, XLM-R, and BLOOM aim to achieve multilingual generalization or compression to facilitate transfer to a large number of (potentially unseen) languages. However, these models should ideally also be private, linguistically fair, and transparent, by relating their predictions to training data. Can these requirements be simultaneously satisfied? We show that multilingual compression and linguistic fairness are compatible with differential privacy, but that differential privacy is at odds with training data influence sparsity, an objective for transparency. We further present a series of experiments on two common NLP tasks and evaluate multilingual compression and training data influence sparsity under different privacy guarantees, exploring these trade-offs in more detail. Our results suggest that we need to develop ways to jointly optimize for these objectives in order to find practical trade-offs.

1. Introduction

One of the open challenges in AI is bridging the widening digital language divide by providing technologies that work well for all languages. Multilingual language models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020a), and BLOOM (Scao et al., 2022), facilitate transfer between closely related languages, enabling roll-out of technologies for low-resource languages, and are used for a wide range of real-world applications in many languages—e.g., from named entity recognition (Khalifa et al., 2021) to legal document classification (Wang & Banko, 2021). Generalization across languages is challenged by typological divides, language families, or scripts (Singh et al., 2019; Dufter & Schütze, 2020) and finding architectures that best facilitate such transfer, achieving optimal **multilingual compression** (Ravishankar & Søgaard, 2021) through parameter sharing (rather than compartmentalization), remains an open research problem.

With the widespread adaptation of multilingual language models also comes responsibility and requirements that models are trustworthy (Pruksachatkun et al., 2021). What does trustworthiness amount to for multilingual language models? A crucial requirement is that multilingual NLP models perform equally well across languages, not favoring any languages over others. (Choudhury & Deshpande, 2021) refer to this property as **linguistic fairness**. Linguistic fairness is defined as zero variance across language-specific losses, typically estimated on held-out data.[1]

Another crucial requirement is *transparency*, i.e., the ability to say *why* models make particular predictions. Methods to achieve transparency come in two flavors; Some methods—commonly referred to as feature attribution methods—present rationales behind predictions in terms of input token attributions, but such rationales are limited in that they cannot explain predictions motivated by the absence of input tokens or the presence of particular token combinations. Feature attribution methods have also been shown to be unreliable (Kindermans et al., 2019; Arun et al., 2020). Other methods highlight training data influence, i.e., provide influential data points as rationales for decisions. Often referred to as instance-based interpretability methods, they are argued to be more useful across different NLP tasks (Han et al., 2020; Han & Tsvetkov, 2021; Zhou et al., 2021b). We refer to the objective of achieving sparse training data influence, i.e. strong instance-interpretability, as **training data influence sparsity**. Finally, for many NLP applications, we further need our models to be private, for which **differential privacy** (DP; Dwork, 2006) provides a theoretically rigorous framework.

[1]Department of Computer Science, University of Copenhagen. Correspondence to: Phillip Rust <p.rust@di.ku.dk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

[1]This definition of linguistic fairness is an instantiation of *equal risk fairness* or overall performance parity, i.e., equal model performance across groups (Berk et al., 2018; Verma & Rubin, 2018; Williamson & Menon, 2019), which balances precision-based and recall-based metrics and is considered more relevant than calibration-based metrics for standard NLP applications. Since the three are mutually exclusive (Miconi, 2017), we ignore calibration and balance precision and recall.

The trustworthiness objectives as defined above have primarily been considered in a monolingual context, and are often (falsely) assumed to be independent (Ruder et al., 2022).[2] Our paper investigates *the extent to which these objectives align or are at odds*. We do so in a multilingual setting and show how multilinguality presents options and challenges.[3] Our theoretical contributions show that while privacy and linguistic fairness are compatible through multilingual compression, privacy and training data influence sparsity are not, and our empirical results indicate that these objectives interact in non-linear ways.[4]

Contributions We begin (in §2) with a theoretical exploration of differential privacy, training data influence, and linguistic fairness in the context of multilingual language models. We show that differential privacy and training data influence sparsity are fundamentally at odds, a result which is not limited to the multilingual setting. While differential privacy and fairness are often said to be at odds, we also show that differential privacy and linguistic fairness over languages are compatible in the multilingual setting, as a result of compression.

Subsequently (in §3–§5), we present empirical results on the impact of differentially private fine-tuning on multilingual compression and training data influence: We analyze the effect of such fine-tuning on the multilingual compression of large LMs and find that it is possible to achieve (i) high compression with strong privacy at the cost of performance; (ii) high compression with high performance at the cost of privacy; or (iii) privacy and accuracy at the cost of compression. Since we show in §2 that performance, privacy and compression *are theoretically* compatible, this leaves us with an open problem: How do we practically optimize for both performance, privacy and compression?

Furthermore, we compare four (proxy) metrics for quantifying multilingual compression—sentence retrieval, centered kernel alignment (CKA; Kornblith et al., 2019), IsoScore (Rudman et al., 2022), representational similarity analysis (RSA; Kriegeskorte et al., 2008; Edelman, 1998)—and discuss their usefulness for balancing these trade-offs.

Finally, we show that LMs exhibiting high multilingual compression are less instance-interpretable in that they make highlighting training data influence more difficult.

In sum, our work shows that *linguistically fair and private high-performance multilingual models are possible, even if learning them is challenging. However, training data influence methods will fail for such models*.

2. Theoretical Exploration

We consider language model learning and fine-tuning in a multilingual setting, in which our training data $D = D_1 \cup \ldots \cup D_{|L|}$ is the union of disjoint training data from $|L|$ different languages. We consider the interaction of differential privacy, training data influence and linguistic fairness with performance and compression in this setting.

Preliminaries We briefly introduce our formal definitions here: A randomized algorithm, here model, $\mathcal{M} : \mathcal{D} \to \mathcal{Y}$ is *ε_p-differentially private* (Dwork, 2006) iff for all adjacent datasets $D, D' \in \mathcal{D}$ and all $Y \subset \mathcal{Y}$, $\mathbb{P}(\mathcal{M}(D) \in Y) \leq \exp(\varepsilon_p) \cdot \mathbb{P}(\mathcal{M}(D') \in Y)$.[5] Adjacent means that the datasets differ by exactly one example x_{diff}.

A model $\mathcal{M}$ is said to be *ε_i-instance-interpretable*, i.e., having sparse training data influence, iff for any $D, D', D'' \in \mathcal{D}$ with $D' = D \setminus \{x_{diff}\}$, $D'' = D \setminus \{x'\}$, and $x_{diff} \neq x'$, where x_{diff} is the most influential training data point under leave-one-out influence,[6] it holds that $\mathbb{P}(\mathcal{M}(D) \in Y) - \mathbb{P}(\mathcal{M}(D') \in Y) > \exp(\varepsilon_i) \cdot (\mathbb{P}(\mathcal{M}(D) \in Y) - \mathbb{P}(\mathcal{M}(D'') \in Y))$. In other words, x_{diff} had more influence on $\mathcal{M}$ than any other data point x' by some margin $\exp(\varepsilon_i)$ (Koh & Liang, 2017).

A model $\mathcal{M}$ is said to be fair if for a group partitioning $g(D) \to D_{g_1}, \ldots, D_{g_n}$ into smaller samples and for some loss function ℓ, e.g., 0-1 loss, $\ell(\mathcal{M}(D_{g_i})) = \ell(\mathcal{M}(D_{g_j}))$ (Williamson & Menon, 2019). A model that is fair for a group partitioning by languages is said to be linguistically fair (Choudhury & Deshpande, 2021).

Finally, a model $\mathcal{M}$ exhibits perfect multilingual compression when it outputs identical representations for semantically equivalent inputs irrespective of the input language. Formally, for a pair of translation equivalent sentences, (i_j, i_q), the representations of i_j and i_q are identical at any layer l of the model, i.e $\mathcal{M}^l(i_j) = \mathcal{M}^l(i_q)$.

[2] One exception is a growing body of work showing fairness and differential privacy are at odds (Bagdasaryan et al., 2019; Cummings et al., 2019; Chang & Shokri, 2021; Hansen et al., 2022). While (Naidu et al., 2021) show that differential privacy and GradCAM (Selvaraju et al., 2019), a feature attribution method, are compatible, the interaction between differential privacy and training data influence remains unexplored.

[3] We are, to the best of our knowledge, first to consider differential privacy in a multilingual setting specifically, with the exception of work on differentially private neural machine translation (Kim et al., 2021).

[4] Our code is available at `https://github.com/xplip/multilingual-lm-objectives`.

[5] Note how standard empirical risk minimization is not private, since it is a linear combination of training samples near the decision boundary, and if D and D' differ in one of those, the classifier changes significantly.

[6] Leave-one-out here means $D' = D \setminus \{x_{diff}\}$ and is the gold standard for instance-based methods, which explains the close connection to DP where we also deal with adjacent datasets.

In the following paragraphs, we discuss under what conditions DP, training data influence, linguistic fairness, and multilingual compression are at odds or are compatible, and how these conditions align with common scenarios in multilingual NLP.[7]

Differential Privacy and Training Data Influence Sparsity We first show that differential privacy and training data influence sparsity are fundamentally at odds:

Theorem 1. *A model $\mathcal{M}$ becomes less ε_i-instance-interpretable as it becomes more ε_p-differentially private, and vice-versa.*

Proof. Let $\mathbb{P}(\mathcal{M}(D) \in Y)$ be abbreviated as p, $\mathbb{P}(\mathcal{M}(D') \in Y) = \mathbb{P}(\mathcal{M}(D \setminus \{x_{diff}\} \in Y)$ be abbreviated as p_d, and let $\mathbb{P}(\mathcal{M}(D'') \in Y) = \mathbb{P}(\mathcal{M}(D \setminus \{x'\} \in Y)$ be abbreviated as p_2. Assume that $\mathcal{M}$ is ε_i-instance-interpretable and ε_p-differentially private.

If $\mathcal{M}$ is ε_p-differentially private, it holds that

$$\begin{aligned} & p && \leq \exp(\varepsilon_p) \cdot p_d \\ \Rightarrow \quad & \exp(\varepsilon_p) && \geq \frac{p}{p_d} \end{aligned} \tag{1}$$

If $\mathcal{M}$ is also ε_i-instance-interpretable, it also holds that

$$\begin{aligned} (i) \quad & p - p_d > && \exp(\varepsilon_i)(p - p_2) \\ (ii) \Rightarrow \quad & p > && \exp(\varepsilon_i)(p - p_2) + p_d \\ (iii) \Rightarrow \quad & \frac{p}{p_d} > && \frac{\exp(\varepsilon_i)(p - p_2) + p_d}{p_d} \\ (iv) \Rightarrow \quad & \exp(\varepsilon_p) > && \frac{\exp(\varepsilon_i)(p - p_2)}{p_d} + 1 \end{aligned} \tag{2}$$

Step (iv) follows from Equation 1. We can now see from Equation 2 step (iv) that ε_p increases with increasing ε_i, i.e. the model becomes less differentially private as it becomes more instance-interpretable, and vice-versa. □

This result is not limited to the multilingual setting.

Differential Privacy and Linguistic Fairness Fairness and differential privacy are occasionally at odds, as shown in (Bagdasaryan et al., 2019; Cummings et al., 2019; Chang & Shokri, 2021; Hansen et al., 2022),[8] but in the multilingual setting, fairness and privacy can be compatible (for the common definitions above). We first note that there is a trivial solution to obtaining differential privacy and linguistic fairness (a joint optimum), namely randomness. This simply shows that the two objectives can be simultaneously satisfied. Next, imagine a perfectly compressed multilingual language model trained on a multi-parallel dataset.

Theorem 2. *If a model $\mathcal{M}_D$ trained on parallel data from $|L| \geq 2$ languages, $D = \{\ldots, i_1, \ldots, i_{|L|}, \ldots\}$, with i_j and i_q being translation equivalents, is perfectly multilingually compressed, then it is ε_p-differentially private.*

Proof. Since $\mathcal{M}_D$ is perfectly compressed, the representation of i_j is identical to i_q at any layer l, i.e., $\mathcal{M}^l_D(i_j) = \mathcal{M}^l_D(i_q)$. This gives us strong k-anonymity (Li et al., 2012) in the representation space of $\mathcal{M}_D$, with $k = |L|$ and all dimensions as quasi-identifiers. Since k-anonymity is not obtained through a deterministic (reversible) procedure, but a randomly initialized learning procedure with random sampling, and since our attributes are randomly initialized, k-anonymization entails differential privacy in our setting.[9] $\mathcal{M}_D$, given perfect compression and convergence, is 0-differentially private, i.e., the probability distribution of $\mathcal{M}_D$ is unaffected by the removal of any single row. □

It follows directly from perfect compression that $\mathcal{M}_D$ is also linguistically fair because identical representations imply identical performance across languages. It is therefore an immediate corollary of the above result that a linguistically fair model can be differentially private.

While the assumptions of a perfectly compressed model and clean multi-parallel dataset rarely hold up in practice and there is no obvious way to satisfy them while maintaining utility, the practical significance of this result is a reminder that multilingual training converges toward k-anonymization, and that safe k-anonymization of the representation space, if obtained, would provide us differential privacy. In the absence of strong guarantees, increasing the number of training languages (larger k) would strengthen privacy (Li et al., 2012). Our empirical results below (§4) suggest that we can often obtain strong privacy and strong compression, but at the cost of performance.

3. Experimental Setup

In our experiments, we investigate the relation between the performance and multilingual compression of fine-tuned multilingual language models, and their privacy and training data influence. We rely on a commonly used multilin-

[7] Differential privacy meaningfully protects any individual training example. However, sensitive information may be repeated across many training examples, so ε-DP does not necessarily prevent leakage of such information at the granularity of individual people, real-world events, etc. For example, in our multilingual setting, an attacker may still gain access to a social security number learned by the model, but they will be unable to identify whether the number was leaked in a particular language.

[8] Several authors have considered practical trade-offs between privacy and fairness, including (Jagielski et al., 2019), (Lyu et al., 2020), (Pannekoek & Spigler, 2021), and (Liu et al., 2021b).

[9] The procedure also is not dependent on any individual input, because all individual data properties are either random (from initialization) or k-anonymous, by construction.

gual pretrained language model, which we fine-tune with different levels of (ε, δ)-differential privacy on two common NLP tasks and evaluate using metrics of compression and training data influence.[10] This section presents the pretrained language model, the tasks, the training protocol, the metrics of compression and training data influence, and the evaluation procedure.

Model We use a pretrained XLM-R Base (Conneau et al., 2020a), which is a 12-layer encoder-only transformer with ∼277M parameters and 250k vocabulary size trained on CC-100 (100 languages) via masked language modeling.

Tasks and Data We fine-tune in a zero-shot cross-lingual transfer setting for part-of-speech (POS) tagging and natural language inference (NLI). Why these tasks? First, while POS tagging is driven by lower-level syntactic features, NLI requires a higher-level understanding (Lauscher et al., 2020). Second, we can leverage *multi-parallel* corpora for multilingual fine-tuning and zero-shot cross-lingual transfer in both tasks, which helps eliminate confounders.[11]

For POS tagging, we use the Parallel Universal Dependencies (PUD) treebank from Universal Dependencies (UD) v2.8 (Nivre et al., 2020; Zeman et al., 2021), which contains 1000 sentences parallel across 15 languages. We train in 7 of these languages (FR, IT, JA, PT, TH, TR, ZH),[12] exclude English,[13] and use the remaining 7 languages (AR, DE, ES, HI, ID, KO, RU) for validation. This split ensures that (1) we both train and evaluate on typologically diverse language samples, (2) there exist additional UD v2.8 treebanks in our validation set languages that we can harness for testing, and (3) there exist parallel sentences in our training set languages that we can harness to evaluate multilingual compression. We use the test splits of the following treebanks for testing: Arabic-PADT, German-GSD, Spanish-GSD, Hindi-HDTB, Indonesian-GSD, Korean-Kaist, and Russian-SynTagRus. Appendix Table 4 lists the treebanks' sizes.[14]

For NLI, we rely on the XNLI dataset (Conneau et al., 2018), which contains (premise, hypothesis, label)-triplets multi-parallel across 15 languages. We, again, train in 7 of these languages (BG, ES, FR, HI, TR, VI, ZH), exclude the original English data, and validate in the remaining 7 languages (AR, DE, EL, RU, SW, TH, UR). We train and validate our models on the original XNLI validation data (7500 examples per language), and we test the models on the original test data (15000 examples per language) in the validation set languages.

The idea to train and validate on the same sentences (in different languages) while testing on sentences from different treebanks (as we do for POS) or a different dataset split (as for XNLI) is to induce a slight distributional shift between validation and test data for the same language sample. This shift lets us evaluate the regularization strength of the gradient noise added by the DP-optimizer.

Training We employ the standard fine-tuning procedures for token classification (POS) and sequence classification (XNLI) proposed by (Devlin et al., 2019). Similar to (Li et al., 2022), we use DP-AdamW (i.e., the DP-SGD algorithm (Abadi et al., 2016) applied to the AdamW optimizer with default hyperparameters (Loshchilov & Hutter, 2019; Kingma & Ba, 2015)) to train with (ε, δ)-DP. We evaluate 6 different privacy budgets with $\varepsilon \in \{1, 3, 8, 15, 30, \infty\}$.[15] We set $\delta = \frac{1e-4}{|D_{train}|}$ for POS, where $|D_{train}| = 7000$ is the length of the training dataset, and $\delta = 1e{-}6$ for XNLI.[16] The noise multiplier σ corresponding to a particular (ε, δ)-budget is determined numerically before training through binary search. Our implementation builds upon the optimized Opacus (Yousefpour et al., 2021) privacy engine by (Li et al., 2022).[17,18] We use the Rényi differential privacy (RDP; Mironov, 2017; Mironov et al., 2019) accountant with conversion to (ε, δ)-DP (Canonne et al., 2020). Hyper-parameter tuning on private data—which the POS and XNLI data in our study simulate—has been shown to incur additional privacy leakage (Liu & Talwar, 2019; Papernot & Steinke, 2022). Therefore, we try to keep hyper-parameter tuning to a minimum and rely on sensible priors to select a suitable range of hyper-parameters. For POS, we find that the range of good hyper-parameters for non-private settings transfers well to private settings if we just use slightly higher learning rates. For XNLI, we select hyper-parameters in a way that matches the sampling rate (Li et al., 2022) found to suit the NLI tasks in the GLUE benchmark (Wang et al., 2018) well.[19] Accordingly, we train with a maximum sequence length of 128 for 10 epochs

[10] For completeness, we explain the difference between ε-DP and (ε, δ)-DP in Appendix B.

[11] One limitation of this selection is that we only consider classification but no generative tasks, which could be worth exploring in the future.

[12] See Table 2 for language details.

[13] We exclude English to keep the number of languages balanced and because the combined corpus is already biased towards Indo-European with Latin scripts (see Table 2).

[14] Regardless of test split size, each language contributes equally to the mean accuracy reported in Figure 1.

[15] $\varepsilon = \infty$ refers to the standard, non-private setting.

[16] We deliberately use a larger δ for XNLI because it turned out to be much harder to achieve convergence than for POS. Even with the looser DP bounds from $\delta = 1e - 6$, we were unable to find a hyper-parameter setting for $\varepsilon = 1$ where the fine-tuned model was substantially better than random guessing.

[17] https://github.com/lxuechen/private-transformers

[18] We do not use ghost clipping, their proposed technique to fit larger batches on the GPU at the cost of training time, as we can still fit sufficiently large batches on our GPUs without.

[19] The sampling rate $q = \frac{B_{\text{train}}}{|D_{\text{train}}|}$, B denoting the batch size.

with a total batch size of 96 for POS and 30 epochs with batch size 512 for XNLI.[20] At each privacy budget, we train models (3 random initializations each) with 6 learning rates for POS ($1e{-}4$, $3e{-}4$, $5e{-}4$, $7e{-}4$, $1e{-}5$, $5e{-}5$, $7e{-}5$, $1e{-}6$) and 3 learning rates for XNLI ($3e{-}4$, $4e{-}4$, $5e{-}4$ for private models and $9e{-}5$, $1e{-}4$, $2e{-}4$ for non-private models). Based on the validation accuracy we then select the 5 best settings for each privacy level and task, listed in Appendix C. The learning rate is linearly decayed after 50 warm-up steps for POS and without warm-up for XNLI. We perform gradient clipping (per-sample in private settings) with a threshold of 0.1. Weight decay is set to 0.01.

Quantifying Multilingual Compression We present four metrics of multilingual compression: A common proxy task to measure the quality of cross-lingual representations is sentence retrieval (Artetxe & Schwenk, 2019; Dufter & Schütze, 2020; Libovický et al., 2020; Ravishankar & Søgaard, 2021; Liu et al., 2021c; Maronikolakis et al., 2021). (Dufter & Schütze, 2020) quantify the degree of multilingual compression using bidirectional sentence retrieval precision as follows:[21]

$$\mathrm{P} = \frac{1}{2m}\sum_{i=1}^{m} \mathbb{1}_{\arg\max_k R_{ik}=i} + \mathbb{1}_{\arg\max_k R_{ki}=i}. \quad (3)$$

Here, $R \in \mathbb{R}^{m\times m}$ denotes the matrix of cosine similarities $R_{ij} = \cos(e_i^q, e_j^r)$ between the m sub-word representations e_i^q and e_j^r from a LM at indices i and j for a set of parallel sentences in the languages q and r.[22]

(Kornblith et al., 2019) propose to use linear centered kernel alignment (CKA) as a similarity index for neural network representations. It is defined as

$$\mathrm{CKA}(X, Y) = \frac{\|Y^{\mathrm{T}}X\|_F^2}{\|X^{\mathrm{T}}X\|_F\|Y^{\mathrm{T}}Y\|_F}. \quad (4)$$

For LMs, the matrices X and Y are obtained by mean-pooling n sub-word representations at model layer l (Conneau et al., 2020b; Glavaš & Vulić, 2021). Typically, X and Y correspond to the representations from two different models for identical examples (Kornblith et al., 2019; Phang et al., 2021). We instead use the representations from a single model for a parallel sentence pair (s_q, s_r) in languages q and r as X and Y, respectively, to study the similarity of representations across languages, similar to (Muller et al., 2021) and (Conneau et al., 2020b). (Yang et al., 2022) also use CKA as a metric of compression.

IsoScore (Rudman et al., 2022) is an isotropy metric, computed as outlined in Appendix D, that quantifies the degree to which a point cloud uniformly utilizes the vector space. In our context, this point cloud corresponds to the n sub-word representations of all examples in a corpus at layer l. Prior work has shown that anisotropic representation spaces, such as the embedding spaces of large LMs (Ethayarajh, 2019), suffer from so-called *representation degeneration* (Gao et al., 2019), and that the isotropy of a model's representation space correlates with its task performance (Zhou et al., 2019; Wang et al., 2020; Zhou et al., 2021a; Rajaee & Pilehvar, 2021, *inter alia*). High isotropy also means languages are not compartmentalized and should therefore correlate with high compression.

Representational similarity analysis (RSA; Kriegeskorte et al., 2008; Edelman, 1998) was originally introduced in the field of cognitive neuroscience to analyze the similarity of fMRI activity patterns, but it is also applicable to neural network representations (Bouchacourt & Baroni, 2018; Chrupała, 2019; Chrupała & Alishahi, 2019; Lepori & McCoy, 2020; He et al., 2021, *inter alia*), e.g., to analyze their similarity across languages. RSA measures the similarity between the representational geometries (i.e., the arrangement in the vector space) of two sets of representations. The representational geometry is determined through pairwise (dis)similarity metrics, and similarity is typically measured using a rank-based correlation metric such as Spearman's ρ (Diedrichsen & Kriegeskorte, 2017).

Quantifying Training Data Influence Training data influence metrics can help us gain an understanding of the inner workings of a model (Koh & Liang, 2017; Yeh et al., 2018; Charpiat et al., 2019; Koh et al., 2019; Pruthi et al., 2020; Basu et al., 2020; K & Søgaard, 2021; Zhang et al., 2021; Kong & Chaudhuri, 2021, *inter alia*). Such metrics are approximations of leave-one-out-influence. (Pruthi et al., 2020) proposed a both effective and practical method, called $\mathrm{TracInCP}$,[23] to compute the influence of a training example z on the model's prediction for another example z', which could be a test example or z itself (called the self-influence). The influence is computed as follows:

$$\mathrm{TracInCP}(z, z') = \sum_{i=1}^{k} \eta_i \nabla\ell(\theta_i, z)\cdot\nabla\ell(\theta_i, z'), \quad (5)$$

[20] Note that using fixed-size batches technically breaks the privacy guarantees of RDP based on the Sampled Gaussian Mechanism (Mironov et al., 2019). We follow the convention of using fixed-size batches, circumventing potential out-of-memory GPU issues, as a proxy for the true privacy spending and performance (see (Li et al., 2022) and Appendix D.4 in (Tramèr & Boneh, 2021)).

[21] Note that (Dufter & Schütze, 2020) also consider word alignment in their multilinguality score. We omit this task as it is not trivial to obtain ground truth alignments in our setup.

[22] The sub-word representations are taken from the LM's layer l and mean-pooled over the sequence length (excluding special tokens).

[23] "CP" stands for checkpoint; the method approximates $\mathrm{TracInIdeal}$, which is impractical to compute, through model checkpoints taken during training (Pruthi et al., 2020).

where η_i is the learning rate and $\nabla\ell(\theta_i, z)$ is the gradient of the loss w.r.t. the model parameters θ_i and inputs z for the i-th model checkpoint. We will use $\mathrm{TracInCP}$ as an approximation of training data influence in our experiments.

Evaluation We evaluate our models both during and after fine-tuning. For POS, we evaluate every 100 steps, and for XNLI, every 200 steps. We measure zero-shot cross-lingual transfer performance on the validation and test data by accuracy (token-level for POS and sequence-level for XNLI). To account for randomness, we take the mean of the best 5 seeds for each privacy budget.

The measures of multilingual compression (sentence retrieval precision, CKA, IsoScore, RSA) are computed using distinct evaluation corpora comprising parallel sentences for all languages pairs in the respective training set language sample. For models trained on XNLI, we use 3000 sentence pairs per language pair from the TED 2020 corpus (Reimers & Gurevych, 2020) and 3500 pairs from the WikiMatrix dataset (Schwenk et al., 2021). For models trained for POS, we use 3500 pairs from TED 2020, 3500 pairs from WikiMatrix, and 900 pairs from Tatoeba,[24,25,26] numbers chosen based on availability and memory usage.

Following (Dufter & Schütze, 2020), we evaluate the models at layers 0 and 8, which complement each other well with regard to the properties they capture, e.g., multilinguality and task-specificity (Choenni & Shutova, 2020; de Vries et al., 2020; Muller et al., 2021). We compute the sentence retrieval precision between language pairs and take the mean.[27] The IsoScore is computed for the contextualized representations of all examples in the respective corpus at once. In contrast, CKA and RSA scores are also computed per language pair, and then averaged across those.[28] For RSA, we use $\mathrm{D} = 1 - \text{Spearman's } \rho$ and $\mathrm{S} = \text{Spearman's } \rho$ as the dissimilarity and similarity metrics, respectively.[29] Finally, we average results for all four metrics across TED 2020, WikiMatrix, and Tatoeba, the two layers, and the 5 best seeds for each privacy budget.

[24]https://tatoeba.org

[25]We extract sentence pairs from Tatoeba using the tatoebatools library (https://github.com/LBeaudoux/tatoebatools).

[26]We exclude TH from the WikiMatrix and Tatoeba evaluation sets for POS as there are insufficiently many sentence pairs available between TH and the remaining languages.

[27]Sentence retrieval is bidirectional (see Eq. 3). Given $|L|$ languages, we therefore average over the full $\mathbb{R}^{|L|\times|L|}$ language pair matrix, only excluding the main diagonal.

[28]CKA and RSA are symmetrical. Given $|L|$ languages, we thus only use the upper triangle of the $\mathbb{R}^{|L|\times|L|}$ language pair matrix, still excluding the main diagonal.

[29]This is consistent with (Zhelezniak et al., 2019) and (Lepori & McCoy, 2020) who show that Spearman's ρ is more suitable for RSA with embeddings than conventional similarity metrics such as cosine similarity.

For comparison, we also compute all metrics for the original pretrained and a randomly initialized XLM-R model.

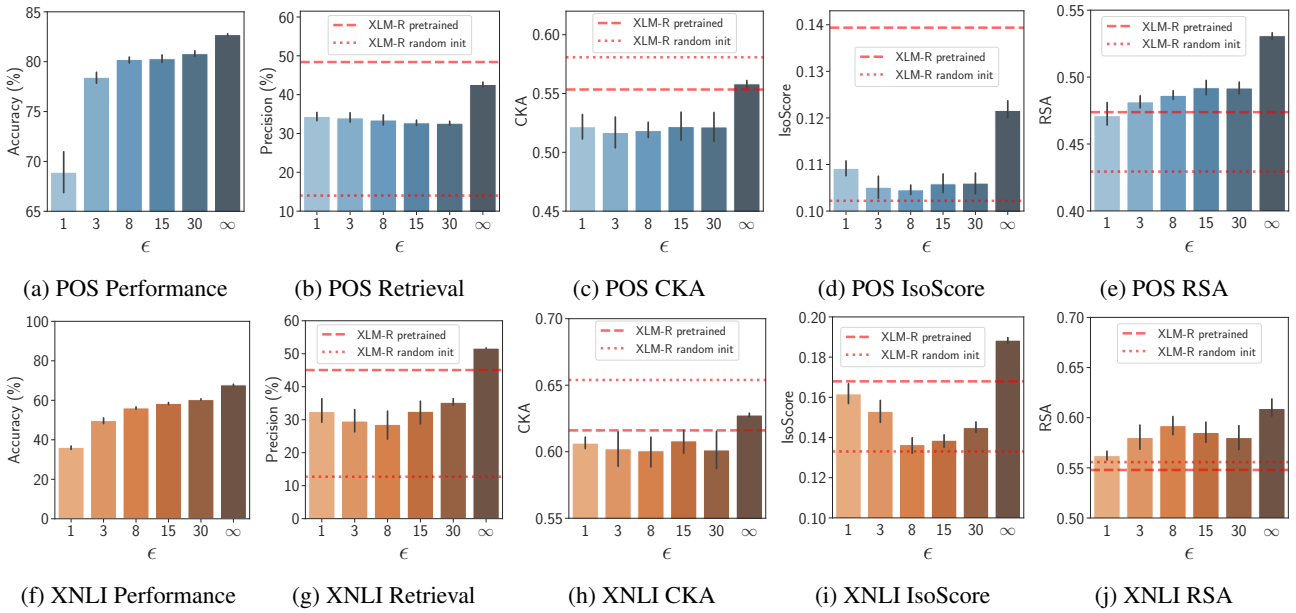
4. Results

Privacy, Compression, Performance We now empirically investigate the relationship between differential privacy, multilingual compression, and cross-lingual transfer performance. We present aggregated results in Figure 1 and non-aggregated results in Appendix G. We observe that the zero-shot accuracy *decreases* as we fine-tune with stronger privacy guarantees (Figures 1a and 1f), which is expected due to the *privacy–utility tradeoff* (Geng et al., 2020). In particular, the relatively small sizes of our training datasets make private LM fine-tuning more challenging (Kerrigan et al., 2020; Habernal, 2021; Senge et al., 2022; Yu et al., 2022) because, for a fixed number of update steps, the gradient noise added per update step grows as the size of the training dataset decreases (Tramèr & Boneh, 2021; McMahan et al., 2018). Note that although the private models tend to underperform the non-private models by a large margin on the validation set (>30% for XNLI, as shown in Appendix Table 6), the performance gap on the test set is noticeably smaller, showing that training with differential privacy, like other noise injection schemes (Bishop, 1995), is also a form of regularization.

Figures 1b and 1g display sentence retrieval precision when fine-tuning with different privacy budgets. The highest compression is achieved by the non-private models. The second-highest compression is achieved for $\varepsilon = 1$, our most private models. Both suggest non-linear privacy–compression interactions, with POS showing lowest compression for $\varepsilon = 30$ (or higher) and XNLI showing lowest compression for $\varepsilon = 8$. The results are very similar for IsoScore (Figures 1d, 1i) and also similar, albeit less pronounced for CKA (Figures 1c, 1h).[30] RSA, in contrast, exhibits very low scores for highly private models; see Appendix E.

These results show that we can achieve *strong compression and strong performance at the cost of privacy* ($\varepsilon = \infty$), *strong compression and strong privacy at the cost of performance* ($\varepsilon = 1$), or *trade-off performance and privacy at the cost of compression* (e.g., $\varepsilon = 8$). It may seem counterintuitive that multilingual compression and cross-lingual transfer performance are not strictly correlated. However, in the fine-tuning setting, we can sacrifice task-specific knowledge in favor of multilingual compression, which leads to poor performance. Vice-versa, a model may ex-

[30]The randomly initialized XLM-R model shows high CKA scores. This is explained by the high dimensionality ($d = 768$) of the contextualized representations, considering that CKA saturates with increasing network width (Kornblith et al., 2019), and the high centroid similarity of random activations.



(a) POS Performance (b) POS Retrieval (c) POS CKA (d) POS IsoScore (e) POS RSA

(f) XNLI Performance (g) XNLI Retrieval (h) XNLI CKA (i) XNLI IsoScore (j) XNLI RSA

Figure 1: Task performance, sentence retrieval, CKA, IsoScore, and RSA results when fine-tuning with different privacy guarantees (∞=non-private). We add the original pretrained XLM-R and XLM-R with randomly initialized weights for comparison. The results show how non-private fine-tuning balances multilingual compression and task performance. Strongly private fine-tuning ($\varepsilon = 1$) is compatible with high compression (retrieval, CKA, IsoScore), but not with task performance. For medium levels of privacy (e.g., $\varepsilon = 8$), we see the result of balancing privacy and task performance at the expense of multilingual compression.

ploit spurious correlations in the data to make correct predictions without actually relying on cross-lingual signal. An example for the former case is the pretrained (but not fine-tuned) XLM-R, which scores highly in multilingual compression (as displayed in Figure 1) but has poor cross-lingual transfer performance in the downstream tasks.

We also find that in some fine-tuning settings, e.g., $\varepsilon = \infty$, the multilingual compression surpasses that of the pretrained XLM-R. While (Liu et al., 2021c) have previously shown that sentence retrieval performance typically drops (i.e., compression worsens) over the course of fine-tuning (which we confirm in Appendix Fig. 5), this finding clearly shows that there are exceptions. Future work may investigate this further.

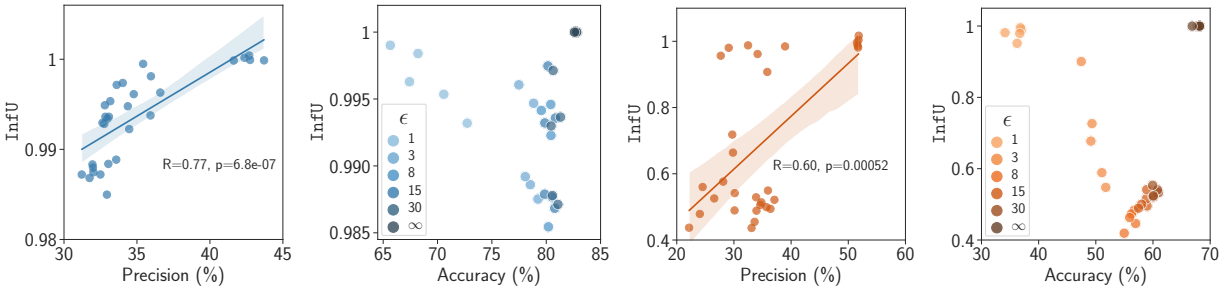
Lastly, retrieval and CKA scores are always highest between typologically similar languages and languages overrepresented in pretraining (see Table 2 for a comparison across languages) *across all levels of privacy*, as shown by the non-aggregated results in the Appendix Figures 6–13. This finding thus extends conclusions from prior work (Pires et al., 2019; Wu & Dredze, 2019; K et al., 2020; Lauscher et al., 2020) to private models.

5. More multilingual, less interpretable?

Metric To answer this question, we introduce InfU (**Inf**luence **U**niformity), a measure of uniformity based on TracInCP influence scores for each training example in the multiparallel dataset $D = \{\ldots, i_1, \ldots, i_{|L|}, \ldots\}$, with i_j and i_k translation equivalents. We compute InfU for $\mathcal{M}$ and the translation equivalents $i = \{i_1, \ldots, i_{|L|}\}$ as follows:

$$\text{InfU}(i) = \frac{1}{|L|} \sum_{k=1}^{|L|} \text{H}(\sigma(\text{TracInCP}(i_k, i))) \qquad (6)$$

where H is the entropy with $log_{|L|}$ and σ is a softmax used to obtain a probability distribution over influence scores. InfU is maximized (InfU $= 1$) for uniform influence scores, fulfilling $\text{TracInCP}(i_j, i_k) = \text{TracInCP}(i_q, i_r)$, $\forall j, k, q, r \in L$. This means a perfectly multilingual model that yields equivalent representations for translation equivalent examples obtains InfU $= 1$. In this scenario of maximum uniformity our model is also the least instance-interpretable because training data influence is minimally sparse, so we cannot easily identify influential examples for a prediction. We use InfU to study to what extent influence sparsity aligns with the metrics privacy and cross-lingual performance.



(a) POS InfU – Retrieval (b) POS InfU – Perf. (c) XNLI InfU – Retrieval (d) XNLI InfU – Perf.

Figure 2: Linear fit and Pearson correlation between the influence uniformity InfU and sentence retrieval precision (2a, 2c) and InfU versus downstream performance for different levels of privacy (2b, 2d). We see significant positive correlations between retrieval precision and InfU, suggesting a negative correlation between multilingual compression and training data influence sparsity. For task performance, we see the trade-off between training data influence sparsity (InfU) and privacy, which aligns with our theoretical expectations (§2).

Setup We use 1000 training examples and compute TracInCP scores from the last 3 model checkpoints, taken every 100 steps, with their corresponding learning rates.[31]

Results and Analysis We plot the mean InfU against the mean sentence retrieval precision for our fine-tuned models and compute Pearson's R in Figures 2a and 2c. For both tasks, there is a significant ($p < 0.05$) strong positive correlation between the InfU score and multilingual compression as determined through sentence retrieval. This supports the idea that *multilingual compression is at odds with training data influence*. See also how highly private and low-performing models score highly in InfU (Figures 2b, 2d); and non-private and high-performing models do the same. For medium levels of privacy we, however, see a trade-off characterized by lower InfU, i.e., better instance-interpretability, and medium performance. Strong *privacy* guarantees, sparse training data influence estimates, and performance are incompatible, because the high-performing models are strictly low in privacy and training data influence sparsity, and the models high in privacy are strictly low in performance and training data influence sparsity.

6. Related Work

While privacy, fairness, and interpretability *individually* have enjoyed ample attention from the research community in recent years Liu et al. (2021a); Mehrabi et al. (2021); Søgaard (2021), the interactions between these objectives have not been explored much (Ruder et al., 2022). Some prior work has focused on the interactions between group fairness and differential privacy, suggesting that the two objectives are at odds, although this relationship also depends on the selected notion of fairness (Bagdasaryan et al., 2019; Cummings et al., 2019; Chang & Shokri, 2021; Hansen et al., 2022). Somewhat in contrast to this work, we show that linguistic fairness (group fairness over linguistic communities) and differential privacy may align for multilingual language models. Furthermore, (Naidu et al., 2021) and (Shokri et al., 2021) have studied the interaction between privacy and feature attribution methods for model explainability. While the former show that privacy and feature attribution methods can align, the latter find that model explanations are at risk of membership inference attacks. Closest to our work is contemporaneous work by (Strobel & Shokri, 2022) who discuss the interactions of data privacy with fairness, explainability, and robustness. Our work differs from theirs in that we are particularly concerned with multilingual language models and we consider instance-based interpretability methods while they consider feature attribution methods. Strobel & Shokri (2022) also call for more research at the intersection of different objectives rather than working on one at a time.

7. Conclusion

We presented a preliminary investigation of how multilingual compression, differential privacy, training data influence, and linguistic fairness interact in multilingual models. We found that privacy and influence are incompatible, while privacy and linguistic fairness, often said to be at odds, are theoretically compatible through multilingual compression. We also explored these interactions empirically. Our results support the idea that high multilingual compression can be achieved either while optimizing for performance or while optimizing for privacy, but that by trading off privacy and performance, we compromise compression. Finding practical trade-offs between *all* these di-

[31] Since the learning rate changes every training step, we use the learning rate from the end of each checkpointing interval.

mensions remains an open challenge. Finally, we introduced a new diagnostic metric, influence uniformity, which we used to validate that privacy and training data influence sparsity are incompatible, and that the interactions between privacy, training data influence sparsity, and multilingual compression are, therefore, also non-linear.

Ethical Aspects and Broader Impact

It is crucial that NLP goes beyond performance and studies the interaction of objectives such as privacy, interpretability, and fairness, also in multilingual NLP (Ruder et al., 2022). Our work aims to provide a starting point for further research in this area. Our empirical investigation, including the models we train, fully relies on publicly available models and data. Moreover, we do not create any new datasets. Therefore, we foresee no misuse of the results of our work.

Acknowledgements

We thank the anonymous reviewers and members of the CoAStaL group for their helpful feedback and suggestions. Phillip Rust is funded by the Novo Nordisk Foundation (grant NNF 20SA0066568).

References

Abadi, M., Chu, A., Goodfellow, I. J., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318, Vienna, Austria, 2016. ACM. URL https://doi.org/10.1145/2976749.2978318.

Artetxe, M. and Schwenk, H. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, March 2019. URL https://aclanthology.org/Q19-1038.

Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M. D., and Kalpathy-Cramer, J. Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging. *medRxiv*, 2020. URL https://www.medrxiv.org/content/early/2020/07/30/2020.07.28.20163899.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 15453–15462, Vancouver, BC, Canada, 2019. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/hash/fc0de4e0396fff257ea362983c2dda5a-Abstract.html.

Basu, S., You, X., and Feizi, S. On second-order group influence functions for black-box predictions. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 715–724, Online, 2020. PMLR. URL http://proceedings.mlr.press/v119/basu20b.html.

Berk, R. A., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50: 3 – 44, 2018. URL https://doi.org/10.1177/0049124118782533.

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. URL https://dl.acm.org/doi/10.1162/neco.1995.7.1.108.

Bouchacourt, D. and Baroni, M. How agents see things: On visual representations in an emergent language game. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 981–985, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1119.

Canonne, C. L., Kamath, G., and Steinke, T. The discrete gaussian for differential privacy. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 15676–15688, Online, 2020. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/hash/b53b3a3d6ab90ce0268229151c9bde11-Abstract.html.

Chang, H. and Shokri, R. On the privacy risks of algorithmic fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 292–303, Los Alamitos, CA, USA, 2021. IEEE Computer Society. URL https://doi.ieeecomputersociety.org/10.1109/EuroSP51992.2021.00028.

Charpiat, G., Girard, N., Felardos, L., and Tarabalka, Y. Input similarity from the neural network perspective. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R.

(eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5343–5352, Vancouver, BC, Canada, 2019. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/hash/c61f571dbd2fb949d3fe5ae1608dd48b-Abstract.html.

Choenni, R. and Shutova, E. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2009.12862.

Choudhury, M. and Deshpande, A. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 12710–12718, Online, 2021. AAAI Press. URL https://ojs.aaai.org/index.php/AAAI/article/view/17505.

Chrupała, G. Symbolic inductive bias for visually grounded learning of spoken language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6452–6462, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1647.

Chrupała, G. and Alishahi, A. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2952–2962, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1283.

Conneau, A., Rinott, R., Lample, G., Williams, A., Bowman, S., Schwenk, H., and Stoyanov, V. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1269.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.747.

Conneau, A., Wu, S., Li, H., Zettlemoyer, L., and Stoyanov, V. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034, Online, July 2020b. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.536.

Cummings, R., Gupta, V., Kimpara, D., and Morgenstern, J. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, UMAP'19 Adjunct, pp. 309–315, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367110. URL https://doi.org/10.1145/3314183.3323847.

de Vries, W., van Cranenburgh, A., and Nissim, M. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4339–4350, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.389.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 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)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1423.

Diedrichsen, J. and Kriegeskorte, N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13 (4):1–33, 04 2017. URL https://doi.org/10.1371/journal.pcbi.1005508.

Dufter, P. and Schütze, H. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.358.

Dwork, C. Differential privacy. In Bugliesi, M., Preneel, B., Sassone, V., and Wegener, I. (eds.), *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), Part II*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12, Venice, Italy, 2006. Springer. URL https://doi.org/10.1007/11787006_1.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. URL https://doi.org/10.1561/0400000042.

Edelman, S. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4): 449–467, 1998. URL https://doi.org/10.1017/s0140525x98001253.

Ethayarajh, K. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1006.

Gao, J., He, D., Tan, X., Qin, T., Wang, L., and Liu, T. Representation degeneration problem in training natural language generation models. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019. OpenReview.net. URL https://openreview.net/forum?id=SkEYojRqtm.

Geng, Q., Ding, W., Guo, R., and Kumar, S. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pp. 89–99, Online, 26–28 Aug 2020. PMLR. URL https://proceedings.mlr.press/v108/geng20a.html.

Glavaš, G. and Vulić, I. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3090–3104, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.270.

Habernal, I. When differential privacy meets NLP: The devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1522–1528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.114.

Han, X. and Tsvetkov, Y. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4398–4409, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.374.

Han, X., Wallace, B. C., and Tsvetkov, Y. Explaining black box predictions and unveiling data artifacts through influence functions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5553–5563, Online, July 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.492.

Hansen, V. P. B., Neerkaje, A. T., Sawhney, R., Flek, L., and Søgaard, A. The impact of differential privacy on group disparity mitigation. In *Proceedings of the Fourth Workshop on Privacy in Natural Language Processing at NAACL 2022*, Seattle, United States, 2022. Association for Computational Linguistics. URL https://arxiv.org/abs/2203.02745.

He, R., Liu, L., Ye, H., Tan, Q., Ding, B., Cheng, L., Low, J., Bing, L., and Si, L. On the effectiveness of adapter-based tuning for pretrained language model adaptation. 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)*, pp. 2208–2222, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.172.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Malvajerdi, S. S., and Ullman, J. Differentially private fair learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3000–3008, Long Beach, CA, USA, 09–15 Jun 2019. PMLR. URL https://proceedings.mlr.press/v97/jagielski19a.html.

K, K. and Søgaard, A. Revisiting methods for finding influential examples. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2111.04683.

K, K., Wang, Z., Mayhew, S., and Roth, D. Cross-lingual ability of multilingual BERT: an empirical study. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Online, 2020. OpenReview.net. URL https://openreview.net/forum?id=HJeT3yrtDr.

Kerrigan, G., Slack, D., and Tuyls, J. Differentially private language models benefit from public pre-training.

In *Proceedings of the Second Workshop on Privacy in NLP*, pp. 39–45, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.privatenlp-1.5.

Khalifa, M., Abdul-Mageed, M., and Shaalan, K. Self-training pre-trained language models for zero- and few-shot multi-dialectal Arabic sequence labeling. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 769–782, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.65.

Kim, S., Bisazza, A., and Turkmen, F. Using confidential data for domain adaptation of neural machine translation. In *Proceedings of the Third Workshop on Privacy in Natural Language Processing*, pp. 46–52, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.privatenlp-1.6.

Kindermans, P.-J., Hooker, S., Adebayo, J., Alber, M., Schütt, K. T., Dähne, S., Erhan, D., and Kim, B. The (un)reliability of saliency methods. In *Explainable AI*, 2019. URL https://doi.org/10.1007/978-3-030-28954-6_14.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. URL http://arxiv.org/abs/1412.6980.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1885–1894, Sydney, NSW, Australia, 2017. PMLR. URL http://proceedings.mlr.press/v70/koh17a.html.

Koh, P. W., Ang, K., Teo, H. H. K., and Liang, P. On the accuracy of influence functions for measuring group effects. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5255–5265, Vancouver, BC, Canada, 2019. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/hash/a78482ce76496fcf49085f2190e675b4-Abstract.html.

Kong, Z. and Chaudhuri, K. Understanding instance-based interpretability of variational auto-encoders. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems (NeurIPS)*, Online, 2021. Curran Associates, Inc. URL https://arxiv.org/abs/2105.14203.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, Long Beach, CA, USA, 2019. PMLR. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Kriegeskorte, N., Mur, M., and Bandettini, P. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. ISSN 1662-5137. URL https://www.frontiersin.org/article/10.3389/neuro.06.004.2008.

Lauscher, A., Ravishankar, V., Vulić, I., and Glavaš, G. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4483–4499, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.363.

Lepori, M. and McCoy, R. T. Picking BERT's brain: Probing for linguistic dependencies in contextualized embeddings using representational similarity analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3637–3651, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. URL https://aclanthology.org/2020.coling-main.325.

Lhoest, Q., Villanova del Moral, A., Jernite, Y., Thakur, A., von Platen, P., Patil, S., Chaumond, J., Drame, M., Plu, J., Tunstall, L., Davison, J., Šaško, M., Chhablani, G., Malik, B., Brandeis, S., Le Scao, T., Sanh, V., Xu, C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., Cistac, P., Goehringer, T., Mustar, V., Lagunas, F., Rush, A., and Wolf, T. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational

Linguistics. URL https://aclanthology.org/2021.emnlp-demo.21.

Li, N., Qardaji, W., and Su, D. On sampling, anonymization, and differential privacy or, k-anonymization meets differential privacy. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, pp. 32–33, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450316484. URL https://doi.org/10.1145/2414456.2414474.

Li, X., Tramer, F., Liang, P., and Hashimoto, T. Large language models can be strong differentially private learners. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Online, 2022. OpenReview.net. URL https://openreview.net/forum?id=bVuP3ltATMz.

Libovický, J., Rosa, R., and Fraser, A. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1663–1674, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.150.

Liu, B., Ding, M., Shaham, S., Rahayu, W., Farokhi, F., and Lin, Z. When machine learning meets privacy: A survey and outlook. *ACM Comput. Surv.*, 54(2), 2021a. ISSN 0360-0300. URL https://doi.org/10.1145/3436755.

Liu, J. and Talwar, K. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, pp. 298–309, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450367059. URL https://doi.org/10.1145/3313276.3316377.

Liu, W., Wang, X., Lu, X., Cheng, J., Jin, B., Wang, X., and Zha, H. Fair differential privacy can mitigate the disparate impact on model accuracy. *Submitted to the 9th International Conference on Learning Representations (ICLR)*, 2021b. URL https://openreview.net/forum?id=IqVB8e0DlUd.

Liu, Z., Winata, G. I., Madotto, A., and Fung, P. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pp. 64–71, Online, August 2021c. Association for Computational Linguistics. URL https://aclanthology.org/2021.repl4nlp-1.8.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA, 2019. OpenReview.net. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Lyu, L., He, X., and Li, Y. Differentially private representation for NLP: Formal guarantee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2355–2365, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.findings-emnlp.213.

Maronikolakis, A., Dufter, P., and Schütze, H. Wine is not v i n. on the compatibility of tokenizations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2382–2399, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.205.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018. OpenReview.net. URL https://openreview.net/forum?id=BJ0hF1Z0b.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), 2021. ISSN 0360-0300. URL https://doi.org/10.1145/3457607.

Miconi, T. The impossibility of "fairness": A generalized impossibility result for decisions. *arXiv preprint*, 2017. URL https://arxiv.org/abs/1707.01195.

Mironov, I. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, (CSF)*, pp. 263–275, Santa Barbara, CA, USA, 2017. IEEE Computer Society. URL https://doi.org/10.1109/CSF.2017.11.

Mironov, I., Talwar, K., and Zhang, L. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint*, 2019. URL https://arxiv.org/abs/1908.10530.

Muller, B., Elazar, Y., Sagot, B., and Seddah, D. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association*

for Computational Linguistics: Main Volume, pp. 2214–2231, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.189.

Naidu, R., Priyanshu, A., Kumar, A., Kotti, S., Wang, H., and Mireshghallah, F. When differential privacy meets interpretability: A case study. In *CVPR 2021 Workshop for Responsible Computer Vision (RCV)*, 2021. URL https://arxiv.org/abs/2106.13203.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., and Zeman, D. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4034–4043, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.497.

Pannekoek, M. and Spigler, G. Investigating trade-offs in utility, fairness and differential privacy in neural networks. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2102.05975.

Papernot, N. and Steinke, T. Hyperparameter tuning with renyi differential privacy. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Online, 2022. OpenReview.net. URL https://openreview.net/forum?id=-70L8lpp9DF.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 8024–8035, Vancouver, BC, Canada, 2019. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

Phang, J., Liu, H., and Bowman, S. R. Fine-tuned transformers show clusters of similar representations across layers. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 529–538, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.blackboxnlp-1.42.

Pires, T., Schlinger, E., and Garrette, D. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://aclanthology.org/P19-1493.

Pruksachatkun, Y., Ramakrishna, A., Chang, K.-W., Krishna, S., Dhamala, J., Guha, T., and Ren, X. (eds.). *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.trustnlp-1.0.

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, Online, 2020. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2020/hash/e6385d39ec9394f2f3a354d9d2b88eec-Abstract.html.

Rajaee, S. and Pilehvar, M. T. A cluster-based approach for improving isotropy in contextual embedding space. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 575–584, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-short.73.

Ravishankar, V. and Søgaard, A. The impact of positional encodings on multilingual compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 763–777, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.59.

Reimers, N. and Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525, Online, November 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-main.365.

Ruder, S., Vulić, I., and Søgaard, A. Square one bias in NLP: Towards a multi-dimensional exploration of the

research manifold. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2340–2354, Dublin, Ireland, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.184.

Rudman, W., Gillman, N., Rayne, T., and Eickhoff, C. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 3325–3339, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.findings-acl.262.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., Castagné, R., Luccioni, A. S., Yvon, F., Gallé, M., Tow, J., Rush, A. M., Biderman, S., Webson, A., Ammanamanchi, P. S., Wang, T., Sagot, B., Muennighoff, N., del Moral, A. V., Ruwase, O., Bawden, R., Bekman, S., McMillan-Major, A., Beltagy, I., Nguyen, H., Saulnier, L., Tan, S., Suarez, P. O., Sanh, V., Laurençon, H., Jernite, Y., Launay, J., Mitchell, M., Raffel, C., Gokaslan, A., Simhi, A., Soroa, A., Aji, A. F., Alfassy, A., Rogers, A., Nitzav, A. K., Xu, C., Mou, C., Emezue, C., Klamm, C., Leong, C., van Strien, D., Adelani, D. I., Radev, D., Ponferrada, E. G., Levkovizh, E., Kim, E., Natan, E. B., De Toni, F., Dupont, G., Kruszewski, G., Pistilli, G., Elsahar, H., Benyamina, H., Tran, H., Yu, I., Abdulmumin, I., Johnson, I., Gonzalez-Dios, I., de la Rosa, J., Chim, J., Dodge, J., Zhu, J., Chang, J., Frohberg, J., Tobing, J., Bhattacharjee, J., Almubarak, K., Chen, K., Lo, K., Von Werra, L., Weber, L., Phan, L., allal, L. B., Tanguy, L., Dey, M., Muñoz, M. R., Masoud, M., Grandury, M., Šaško, M., Huang, M., Coavoux, M., Singh, M., Jiang, M. T.-J., Vu, M. C., Jauhar, M. A., Ghaleb, M., Subramani, N., Kassner, N., Khamis, N., Nguyen, O., Espejel, O., de Gibert, O., Villegas, P., Henderson, P., Colombo, P., Amuok, P., Lhoest, Q., Harliman, R., Bommasani, R., López, R. L., Ribeiro, R., Osei, S., Pyysalo, S., Nagel, S., Bose, S., Muhammad, S. H., Sharma, S., Longpre, S., Nikpoor, S., Silberberg, S., Pai, S., Zink, S., Torrent, T. T., Schick, T., Thrush, T., Danchev, V., Nikoulina, V., Laippala, V., Lepercq, V., Prabhu, V., Alyafeai, Z., Talat, Z., Raja, A., Heinzerling, B., Si, C., Taşar, D. E., Salesky, E., Mielke, S. J., Lee, W. Y., Sharma, A., Santilli, A., Chaffin, A., Stiegler, A., Datta, D., Szczechla, E., Chhablani, G., Wang, H., Pandey, H., Strobelt, H., Fries, J. A., Rozen, J., Gao, L., Sutawika, L., Bari, M. S., Al-shaibani, M. S., Manica, M., Nayak, N., Teehan, R., Albanie, S., Shen, S., Ben-David, S., Bach, S. H., Kim, T., Bers, T., Fevry, T., Neeraj, T., Thakker, U., Raunak, V., Tang, X., Yong, Z.-X., Sun, Z., Brody, S., Uri, Y., Tojarieh, H., Roberts, A., Chung, H. W., Tae, J., Phang, J., Press, O., Li, C., Narayanan, D., Bourfoune, H., Casper, J., Rasley, J., Ryabinin, M., Mishra, M., Zhang, M., Shoeybi, M., Peyrounette, M., Patry, N., Tazi, N., Sanseviero, O., von Platen, P., Cornette, P., Lavallée, P. F., Lacroix, R., Rajbhandari, S., Gandhi, S., Smith, S., Requena, S., Patil, S., Dettmers, T., Baruwa, A., Singh, A., Cheveleva, A., Ligozat, A.-L., Subramonian, A., Névéol, A., Lovering, C., Garrette, D., Tunuguntla, D., Reiter, E., Taktasheva, E., Voloshina, E., Bogdanov, E., Winata, G. I., Schoelkopf, H., Kalo, J.-C., Novikova, J., Forde, J. Z., Clive, J., Kasai, J., Kawamura, K., Hazan, L., Carpuat, M., Clinciu, M., Kim, N., Cheng, N., Serikov, O., Antverg, O., van der Wal, O., Zhang, R., Zhang, R., Gehrmann, S., Mirkin, S., Pais, S., Shavrina, T., Scialom, T., Yun, T., Limisiewicz, T., Rieser, V., Protasov, V., Mikhailov, V., Pruksachatkun, Y., Belinkov, Y., Bamberger, Z., Kasner, Z., Rueda, A., Pestana, A., Feizpour, A., Khan, A., Faranak, A., Santos, A., Hevia, A., Unldreaj, A., Aghagol, A., Abdollahi, A., Tammour, A., HajiHosseini, A., Behroozi, B., Ajibade, B., Saxena, B., Ferrandis, C. M., Contractor, D., Lansky, D., David, D., Kiela, D., Nguyen, D. A., Tan, E., Baylor, E., Ozoani, E., Mirza, F., Ononiwu, F., Rezanejad, H., Jones, H., Bhattacharya, I., Solaiman, I., Sedenko, I., Nejadgholi, I., Passmore, J., Seltzer, J., Sanz, J. B., Dutra, L., Samagaio, M., Elbadri, M., Mieskes, M., Gerchick, M., Akinlolu, M., McKenna, M., Qiu, M., Ghauri, M., Burynok, M., Abrar, N., Rajani, N., Elkott, N., Fahmy, N., Samuel, O., An, R., Kromann, R., Hao, R., Alizadeh, S., Shubber, S., Wang, S., Roy, S., Viguier, S., Le, T., Oyebade, T., Le, T., Yang, Y., Nguyen, Z., Kashyap, A. R., Palasciano, A., Callahan, A., Shukla, A., Miranda-Escalada, A., Singh, A., Beilharz, B., Wang, B., Brito, C., Zhou, C., Jain, C., Xu, C., Fourrier, C., Periñán, D. L., Molano, D., Yu, D., Manjavacas, E., Barth, F., Fuhrimann, F., Altay, G., Bayrak, G., Burns, G., Vrabec, H. U., Bello, I., Dash, I., Kang, J., Giorgi, J., Golde, J., Posada, J. D., Sivaraman, K. R., Bulchandani, L., Liu, L., Shinzato, L., de Bykhovetz, M. H., Takeuchi, M., Pàmies, M., Castillo, M. A., Nezhurina, M., Sänger, M., Samwald, M., Cullan, M., Weinberg, M., De Wolf, M., Mihaljcic, M., Liu, M., Freidank, M., Kang, M., Seelam, N., Dahlberg, N., Broad, N. M., Muellner, N., Fung, P., Haller, P., Chandrasekhar, R., Eisenberg, R., Martin, R., Canalli, R., Su, R., Su, R., Cahyawijaya, S., Garda, S., Deshmukh, S. S., Mishra, S., Kiblawi, S., Ott, S., Sang-aroonsiri, S., Kumar, S., Schweter, S., Bharati, S., Laud, T., Gigant, T., Kainuma, T., Kusa, W., Labrak, Y., Bajaj, Y. S., Venkatraman, Y., Xu, Y., Xu, Y., Xu, Y., Tan, Z., Xie, Z., Ye, Z., Bras, M., Belkada, Y., and Wolf, T. Bloom: A 176b-parameter open-access multilingual language model, 2022. URL https://arxiv.org/abs/2211.05100.

Schwenk, H., Chaudhary, V., Sun, S., Gong, H., and Guzmán, F. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1351–1361, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.eacl-main.115.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, Oct 2019. ISSN 1573-1405. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

Senge, M., Igamberdiev, T., and Habernal, I. One size does not fit all: Investigating strategies for differentially-private learning across NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7340–7353, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.496.

Shokri, R., Strobel, M., and Zick, Y. On the privacy risks of model explanations. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, pp. 231–241, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384735. URL https://doi.org/10.1145/3461702.3462533.

Singh, J., McCann, B., Socher, R., and Xiong, C. BERT is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pp. 47–55, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-6106.

Søgaard, A. Explainable natural language processing. *Synthesis Lectures on Human Language Technologies*, 14(3):1–123, 2021. URL https://doi.org/10.2200/S01118ED1V01Y202107HLT051.

Strobel, M. and Shokri, R. Data privacy and trustworthy machine learning. *IEEE Security & Privacy*, 20(5): 44–49, 2022. URL https://doi.org/10.1109/MSEC.2022.3178187.

Tramèr, F. and Boneh, D. Differentially private learning needs better features (or much more data). In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Online, 2021. OpenReview.net. URL https://openreview.net/forum?id=YTWGvpFOQD-.

Verma, S. and Rubin, J. S. Fairness definitions explained. *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, pp. 1–7, 2018. URL https://doi.org/10.1145/3194770.3194776.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. URL https://aclanthology.org/W18-5446.

Wang, C. and Banko, M. Practical transformer-based multilingual text classification. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pp. 121–129, Online, June 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.naacl-industry.16.

Wang, L., Huang, J., Huang, K., Hu, Z., Wang, G., and Gu, Q. Improving neural language generation with spectrum control. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Online, 2020. OpenReview.net. URL https://openreview.net/forum?id=ByxY8CNtvr.

Williamson, R. C. and Menon, A. K. Fairness risk measures. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6786–6797, Long Beach, CA, USA, 2019. PMLR. URL http://proceedings.mlr.press/v97/williamson19a.html.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.emnlp-demos.6.

Wu, S. and Dredze, M. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings*

of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 833–844, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://aclanthology.org/D19-1077.

Yang, H., Chen, H., Zhou, H., and Li, L. Enhancing cross-lingual transfer by manifold mixup. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Online, 2022. OpenReview.net. URL https://openreview.net/forum?id=OjPmfr9GkVv.

Yeh, C., Kim, J. S., Yen, I. E., and Ravikumar, P. Representer point selection for explaining deep neural networks. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 9311–9321, Montréal, Canada, 2018. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper/2018/hash/8a7129b8f3edd95b7d969dfc2c8e9d9d-Abstract.html.

Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Ghosh, S., Bharadwaj, A., Zhao, J., Cormode, G., and Mironov, I. Opacus: User-friendly differential privacy library in pytorch. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, Online, 2021. URL https://openreview.net/forum?id=EopKEYBoI-.

Yu, D., Naik, S., Backurs, A., Gopi, S., Inan, H. A., Kamath, G., Kulkarni, J., Lee, Y. T., Manoel, A., Wutschitz, L., Yekhanin, S., and Zhang, H. Differentially private fine-tuning of language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, Online, 2022. OpenReview.net. URL https://openreview.net/forum?id=Q42f0dfjECO.

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Aleksandravičiūtė, G., Alfina, I., Antonsen, L., Aplonova, K., Aquino, A., Aragon, C., Aranzabe, M. J., Arıcan, B. N., Arnardóttir, Þ., Arutie, G., Arwidarasti, J. N., Asahara, M., Aslan, D. B., Ateyah, L., Atmaca, F., Attia, M., Atutxa, A., Augustinus, L., Badmaeva, E., Balasubramani, K., Ballesteros, M., Banerjee, E., Bank, S., Barbu Mititelu, V., Barkarson, S., Basmov, V., Batchelor, C., Bauer, J., Bedir, S. T., Bengoetxea, K., Berk, G., Berzak, Y., Bhat, I. A., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Bjarnadóttir, K., Blokland, R., Bobicev, V., Boizou, L., Borges Völker, E., Börstell, C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Braggaar, A., Brokaitė, K., Burchardt, A., Candito, M., Caron, B., Caron, G., Cassidy, L., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Čéplö, S., Cesur, N., Cetin, S., Çetinoğlu, Ö., Chalub, F., Chauhan, S., Chi, E., Chika, T., Cho, Y., Choi, J., Chun, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Courtin, M., Cristescu, M., Daniel, P., Davidson, E., de Marneffe, M.-C., de Paiva, V., Derin, M. O., de Souza, E., Diaz de Ilarraza, A., Dickerson, C., Dinakaramani, A., Di Nuovo, E., Dione, B., Dirix, P., Dobrovoljc, K., Dozat, T., Droganova, K., Dwivedi, P., Eckhoff, H., Eiche, S., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Etienne, A., Evelyn, W., Facundes, S., Farkas, R., Fernanda, M., Fernandez Alcalde, H., Foster, J., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Garcia, M., Gärdenfors, M., Garza, S., Gerardi, F. F., Gerdes, K., Ginter, F., Godoy, G., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grobol, L., Grūzītis, N., Guillaume, B., Guillot-Barbance, C., Güngör, T., Habash, N., Hafsteinsson, H., Hajič, J., Hajič jr., J., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Hanifmuti, M. Y., Hardwick, S., Harris, K., Haug, D., Heinecke, J., Hellwig, O., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hohle, P., Huber, E., Hwang, J., Ikeda, T., Ingason, A. K., Ion, R., Irimia, E., Ishola, Ọ., Ito, K., Jelínek, T., Jha, A., Johannsen, A., Jónsdóttir, H., Jørgensen, F., Juutinen, M., K, S., Kaşıkara, H., Kaasen, A., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Kara, N., Katz, B., Kayadelen, T., Kenney, J., Kettnerová, V., Kirchner, J., Klementieva, E., Köhn, A., Köksal, A., Kopacewicz, K., Korkiakangas, T., Kotsyba, N., Kovalevskaitė, J., Krek, S., Krishnamurthy, P., Kuyrukçu, O., Kuzgun, A., Kwak, S., Laippala, V., Lam, L., Lambertino, L., Lando, T., Larasati, S. D., Lavrentiev, A., Lee, J., Phương Lê Hồng, Lenci, A., Lertpradit, S., Leung, H., Levina, M., Li, C. Y., Li, J., Li, K., Li, Y., Lim, K., Lima Padovani, B., Lindén, K., Ljubešić, N., Loginova, O., Luthfi, A., Luukko, M., Lyashevskaya, O., Lynn, T., Macketanz, V., Makazhanov, A., Mandl, M., Manning, C., Manurung, R., Marşan, B., Mărănduc, C., Mareček, D., Marheinecke, K., Martínez Alonso, H., Martins, A., Mašek, J., Matsuda, H., Matsumoto, Y., Mazzei, A., McDonald, R., McGuinness, S., Mendonça, G., Miekka, N., Mischenkova, K., Misirpashayeva, M., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Mojiri Foroushani, A., Molnár, J., Moloodi, A., Montemagni, S., More, A., Moreno Romero, L., Moretti, G., Mori, K. S., Mori, S., Morioka, T., Moro, S., Mortensen, B., Moskalevskyi,

B., Muischnek, K., Munro, R., Murawaki, Y., Müürisep, K., Nainwani, P., Nakhlé, M., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nevaci, M., Lương Nguyễn Thị, Nguyễn Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Nourian, A., Nurmi, H., Ojala, S., Ojha, A. K., Olúòkun, A., Omura, M., Onwuegbuzia, E., Osenova, P., Östling, R., Øvrelid, L., Özateş, Ş. B., Özçelik, M., Özgür, A., Öztürk Başaran, B., Park, H. H., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Peljak-Łapińska, A., Peng, S., Perez, C.-A., Perkova, N., Perrier, G., Petrov, S., Petrova, D., Phelan, J., Piitulainen, J., Pirinen, T. A., Pitler, E., Plank, B., Poibeau, T., Ponomareva, L., Popel, M., Pretkalniņa, L., Prévost, S., Prokopidis, P., Przepiórkowski, A., Puolakainen, T., Pyysalo, S., Qi, P., Rääbis, A., Rademaker, A., Rama, T., Ramasamy, L., Ramisch, C., Rashel, F., Rasooli, M. S., Ravishankar, V., Real, L., Rebeja, P., Reddy, S., Rehm, G., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rocha, L., Rögnvaldsson, E., Romanenko, M., Rosa, R., Roșca, V., Rovati, D., Rudina, O., Rueter, J., Rúnarsson, K., Sadde, S., Safari, P., Sagot, B., Sahala, A., Saleh, S., Salomoni, A., Samardžić, T., Samson, S., Sanguinetti, M., Sanıyar, E., Särg, D., Saulīte, B., Sawanakunanon, Y., Saxena, S., Scannell, K., Scarlata, S., Schneider, N., Schuster, S., Schwartz, L., Seddah, D., Seeker, W., Seraji, M., Shen, M., Shimada, A., Shirasu, H., Shishkina, Y., Shohibussirri, M., Sichinava, D., Siewert, J., Einar Freyr Sigurðsson, Silveira, A., Silveira, N., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Simov, K., Skachedubova, M., Smith, A., Soares-Bastos, I., Spadine, C., Sprugnoli, R., Steingrímsson, S., Stella, A., Straka, M., Strickland, E., Strnadová, J., Suhr, A., Sulestio, Y. L., Sulubacak, U., Suzuki, S., Szántó, Z., Taji, D., Takahashi, Y., Tamburini, F., Tan, M. A. C., Tanaka, T., Tella, S., Tellier, I., Testori, M., Thomas, G., Torga, L., Toska, M., Trosterud, T., Trukhina, A., Tsarfaty, R., Türk, U., Tyers, F., Uematsu, S., Untilov, R., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vajjala, S., van der Goot, R., Vanhove, M., van Niekerk, D., van Noord, G., Varga, V., Villemonte de la Clergerie, E., Vincze, V., Vlasova, N., Wakasa, A., Wallenberg, J. C., Wallin, L., Walsh, A., Wang, J. X., Washington, J. N., Wendt, M., Widmer, P., Williams, S., Wirén, M., Wittern, C., Woldemariam, T., Wong, T.-s., Wróblewska, A., Yako, M., Yamashita, K., Yamazaki, N., Yan, C., Yasuoka, K., Yavrumyan, M. M., Yenice, A. B., Yıldız, O. T., Yu, Z., Žabokrtský, Z., Zahra, S., Zeldes, A., Zhu, H., Zhuravleva, A., and Ziane, R. Universal dependencies 2.8, 2021. URL http://hdl.handle.net/11234/1-3683. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zhang, W., Huang, Z., Zhu, Y., Ye, G., Cui, X., and Zhang, F. On sample based explanation methods for NLP: Faithfulness, efficiency and semantic evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5399–5411, Online, August 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.419.

Zhelezniak, V., Savkov, A., Shen, A., and Hammerla, N. Correlation coefficients and semantic textual similarity. 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)*, pp. 951–962, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1100.

Zhou, T., Sedoc, J., and Rodu, J. Getting in shape: Word embedding subspaces. In Kraus, S. (ed.), *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 5478–5484, Macao, China, 2019. ijcai.org. URL https://doi.org/10.24963/ijcai.2019/761.

Zhou, W., Lin, B. Y., and Ren, X. Isobn: Fine-tuning BERT with isotropic batch normalization. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 14621–14629, Online, 2021a. AAAI Press. URL https://ojs.aaai.org/index.php/AAAI/article/view/17718.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. Do feature attribution methods correctly attribute features? In *XAI 4 Debugging Workshop at NeurIPS 2021*, Online, 2021b. OpenReview. URL https://openreview.net/forum?id=h4J41lQqaJ3.

A. Reproducibility

We make our code available at `https://github.com/xplip/multilingual-lm-objectives`.

Implementation Our implementation is written in PyTorch v1.10.0 (Paszke et al., 2019) for Python 3.9.5 and builds on code from the following repositories:

- https://github.com/huggingface/transformers v4.9.2 (Wolf et al., 2020) for model training and evaluation
- https://github.com/lxuechen/private-transformers v0.1.0 (Li et al., 2022) for DP-training
- https://github.com/pdufter/minimult (Dufter & Schütze, 2020) for computing sentence retrieval precision
- https://github.com/jayroxis/CKA-similarity for computing CKA scores
- https://github.com/mlepori1/Picking_BERTs_Brain (Lepori & McCoy, 2020) for computing RSA scores
- https://github.com/bcbi-edu/p_eickhoff_isoscore (Rudman et al., 2022) for computing IsoScores
- https://github.com/FengNiMa/VAE-TracIn-pytorch (Kong & Chaudhuri, 2021) for computing $\mathrm{TracInCP}$ scores.

Models We primarily use the pretrained XLM-RoBERTa (XLM-R; Conneau et al., 2020a) base model and tokenizer from https://huggingface.co/xlm-roberta-base. XLM-R (base) is a 12-layer encoder-only transformer with a vocabulary size of 250k and ∼277M total parameters pretrained via masked language modeling on the 100-language CC-100 dataset.

In Appendix F, we further conduct experiments with multilingual BERT (mBERT; Devlin et al., 2019), using the base model and tokenizer from https://huggingface.co/bert-base-multilingual-cased. mBERT is a 12-layer encoder-only transformer with a vocabulary size of 120k and ∼177M total parameters pretrained via masked language modeling on Wikipedia data in 104 languages.

Data We provide download links and references for the various datasets we used in Table 3.

Hardware We train on single Nvidia Titan RTX, A100 (both with CUDA version 11.0), and RTX 3090 (with CUDA version 11.5) GPUs. All machines have at least 64GB of RAM, which is required to compute the IsoScore for our larger evaluation sets (e.g., TED 2020 for POS).

Runtime Fine-tuning with evaluation during training on the Titan RTX, which is the slowest of the GPUs used, takes 2–3 hours for POS and 5–6 hours for XNLI. Computing $\mathrm{TracInCP}$ influence scores for one fine-tuned model takes about 30–45 minutes.

Carbon Footprint Our fine-tuning runs accumulated ∼36 compute days on the hardware mentioned above (most experiments were conducted on the less powerful Titan RTX GPUs) according to Weights & Biases[32], where we logged our experiments. Although we do not have precise numbers, a highly conservative estimate of the total compute spent including prototyping, hyper-parameter search, and all our evaluations is ∼75 compute days.

B. (ε, δ)-Differential Privacy

In §2, we provide the definition of ε-differential privacy (DP), also called pure DP, as the basis for our theoretical exploration. In our experiments, we rely on (ε, δ)-DP (Dwork & Roth, 2014), also called approximate-DP, which is typically used in practice and relaxes the privacy guarantees by a (small) δ as follows:

A randomized algorithm $\mathcal{M} : \mathcal{D} \to \mathcal{Y}$ is *(ε, δ)-differentially private* (Dwork, 2006) iff for all adjacent datasets $D, D' \in \mathcal{D}$ and all $Y \subset \mathcal{Y}$, $\mathbb{P}(\mathcal{M}(D) \in Y) \leq \exp(\varepsilon_p) \cdot \mathbb{P}(\mathcal{M}(D') \in Y) + \delta$.

[32] https://wandb.ai/

C. Best Fine-Tuning Settings

As mentioned in §3, we pre-selected a set of suitable learning rates (LRs) for each task and ran 3 random initializations each. Based on the validation performance, we then selected the following 5 best settings for each privacy budget and task:

Table 1: Best 5 settings for each task and privacy budget. Includes LR and the corresponding number of random initializations (# seeds).

ε	POS LR (# Seeds)	XNLI LR (# Seeds)
1	$5e{-}4$ (2); $7e{-}4$ (3)	$3e{-}4$ (1); $4e{-}4$ (2); $5e{-}4$ (2)
3	$5e{-}4$ (2); $7e{-}4$ (3)	$3e{-}4$ (1); $4e{-}4$ (2); $5e{-}4$ (2)
8	$5e{-}4$ (3); $7e{-}4$ (2)	$4e{-}4$ (2); $5e{-}4$ (3)
15	$3e{-}4$ (1); $5e{-}4$ (2); $7e{-}4$ (2)	$3e{-}4$ (1); $4e{-}4$ (2); $5e{-}4$ (2)
30	$3e{-}4$ (1); $5e{-}4$ (2); $7e{-}4$ (2)	$3e{-}4$ (1); $4e{-}4$ (2); $5e{-}4$ (2)
∞	$5e{-}5$ (2); $7e{-}5$ (2); $1e{-}4$ (1)	$9e{-}5$ (2); $1e{-}4$ (3)

D. IsoScore Algorithm

Algorithm 1 describes the IsoScore algorithm (Rudman et al., 2022).

Algorithm 1 IsoScore (Rudman et al., 2022)

begin Let $X \subset \mathbb{R}^n$ be a finite collection of points.
  Let X^{PCA} denote the points in X transformed by the first n principal components.
  Define $\Sigma_D \in \mathbb{R}^n$ as the diagonal of the covariance matrix of X^{PCA}.
  Normalize diagonal to $\hat{\Sigma}_D := \sqrt{n} \cdot \Sigma_D / \|\Sigma_D\|$, where $\|\cdot\|$ is the standard Euclidean norm.
  The isotropy defect is $\delta(X) := \|\hat{\Sigma}_D - \mathbf{1}\| / \sqrt{2(n - \sqrt{n})}$, where $\mathbf{1} = (1, \ldots, 1)^{\mathrm{T}} \in \mathbb{R}^n$
  X uniformly occupies $\phi(X) := (n - \delta(X)^2(n - \sqrt{n}))^2 / n^2$ percent of ambient dimensions.
  Transform $\phi(X)$ so it can take values in $[0, 1]$, via $\iota(X) := (n \cdot \phi(X) - 1)/(n - 1)$.
  return: $\iota(X)$
end

E. Further Analysis of RSA Results

As we see in §4, RSA aligns with sentence retrieval precision, CKA, and IsoScore in producing higher scores for non-private models. However, there is a mismatch between RSA and the other metrics in highly private regimes, where our most private models ($\varepsilon = 1$) do not exhibit high RSA scores. Instead, the aggregated RSA scores peak at medium levels of privacy ($\varepsilon \in \{8, 15\}$) and for the non-private ($\varepsilon = \infty$) models. Unlike for the other metrics, there is also no clear trend among our two tasks in terms of whether the pretrained or a randomly initialized XLM-R model scores higher in RSA.

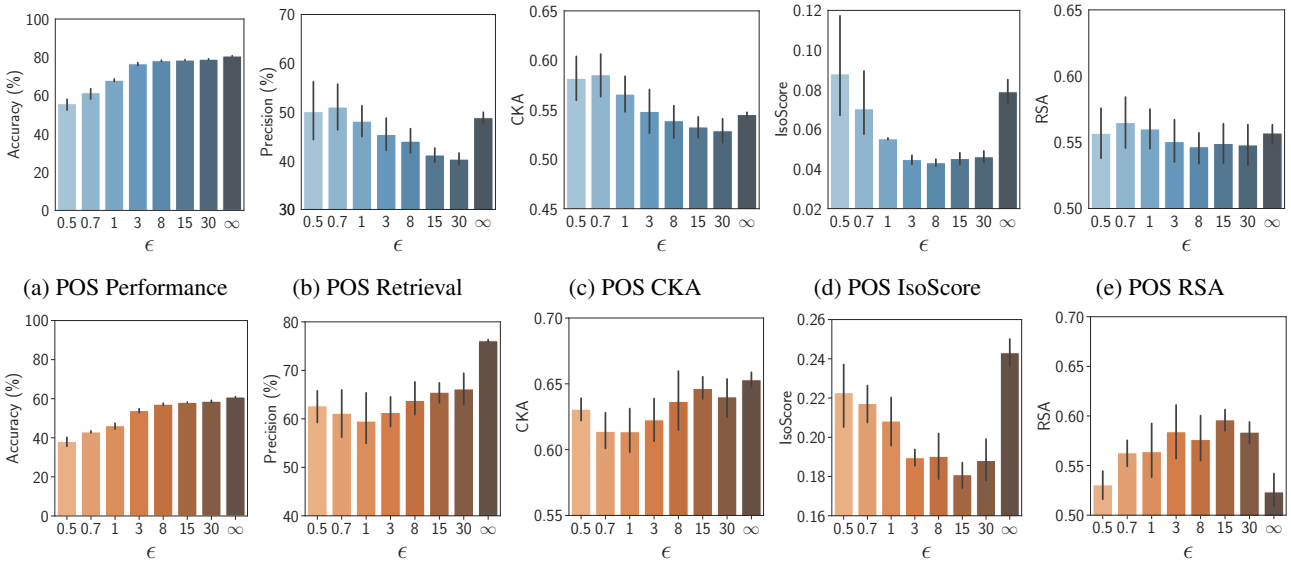
A closer look at the non-aggregated results (Appendix Figures 10, 11, and 14) shows how the similarity patterns obtained from RSA are often unexpected. For instance, the similarities between the typologically distant languages FR and ZH are consistently high for the TED 2020 corpus whereas scores for typologically closer languages are lower (Fig. 10). Based on prior work by, for example, (Pires et al., 2019), (Wu & Dredze, 2019), and (Lauscher et al., 2020), we would expect the model to first compress similar languages before achieving compression for distant ones. Sometimes, we also observe extreme jumps in similarity between layers 0 and 8, for instance, between IT and TR in the Tatoeba corpus (Fig. 11). We do not find these jumps in CKA and sentence retrieval.

One reason why RSA scores may be more sensitive to stricter privacy guarantees (e.g., $\varepsilon = 1$) is that the correlation between sentence vector distances is very sensitive to outliers. Differential privacy reduces the number of such outliers, effectively regularizing the correlation coefficients.

F. Multilingual BERT Results

In Figures 3 and 4, we present results from re-running the experiments from §4 and §5 with mBERT. We make two changes to the experimental setup outlined above: We use representations extracted at layer 8, which showed to be more meaningful than layer 0 in the XLM-R experiments, to compute the multilinguality metrics. We also include two additional privacy settings, $\varepsilon = 0.5$ and $\varepsilon = 0.7$, as we found mBERT to be easier to finetune with strong privacy guarantees than XLM-R.

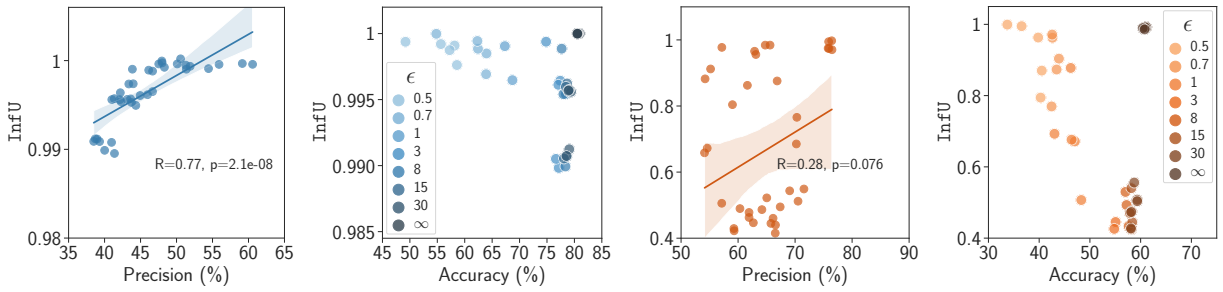
We see the same trends as for XLM-R: performance strictly increases with decreasing privacy while the multilinguality metrics tend to follow a U-shape,[33] i.e., they are high for strong privacy settings (small ε) and low privacy settings (large ε) and decrease towards medium privacy. Likewise, we find a positive correlation between InfU and cross-lingual sentence retrieval precision. The correlation is strong for part-of-speech tagging (POS) but it is mild for XNLI. We believe this may be due to mBERT being less sensitive to the privacy parameter (Figure 3g is not symmetrical; considering even stronger privacy settings would likely even out the U-shape). Overall, these results further support our finding that there is a negative correlation between multilingual compression and training data influence sparsity.



(a) POS Performance (b) POS Retrieval (c) POS CKA (d) POS IsoScore (e) POS RSA

(f) XNLI Performance (g) XNLI Retrieval (h) XNLI CKA (i) XNLI IsoScore (j) XNLI RSA

Figure 3: Aggregated **mBERT** results, analogous to Figure 1.



(a) POS InfU – Retrieval (b) POS InfU – Perf. (c) XNLI InfU – Retrieval (d) XNLI InfU – Perf.

Figure 4: Aggregated **mBERT** results, analogous to Figure 2.

[33] We again refer to Appendix E for a discussion of the RSA results.

G. Detailed Results for Experiments in §4

Figure 5 shows the development of the mean sentence retrieval precision at layer 8 for POS and XNLI over the course of fine-tuning with different privacy budgets.

We further present non-aggregated results for

- POS performance in Table 5
- XNLI performance in Table 6
- Sentence retrieval for POS in Figures 6 and 7
- Sentence retrieval for XNLI in Figure 12
- CKA for POS in Figures 8 and 9
- CKA for XNLI in Figure 13
- IsoScore for POS in Table 7
- IsoScore for XNLI in Table 8
- RSA for POS in Figures 10 and 11
- RSA for XNLI in Figure 14.

Table 2: Overview of languages used in our experiments. Tokens (in millions) and size (in Gibibytes) refer to the respective monolingual corpora in XLM-R's pretraining corpus. Numbers taken from (Conneau et al., 2020a). *: includes romanized variants also used in pretraining.

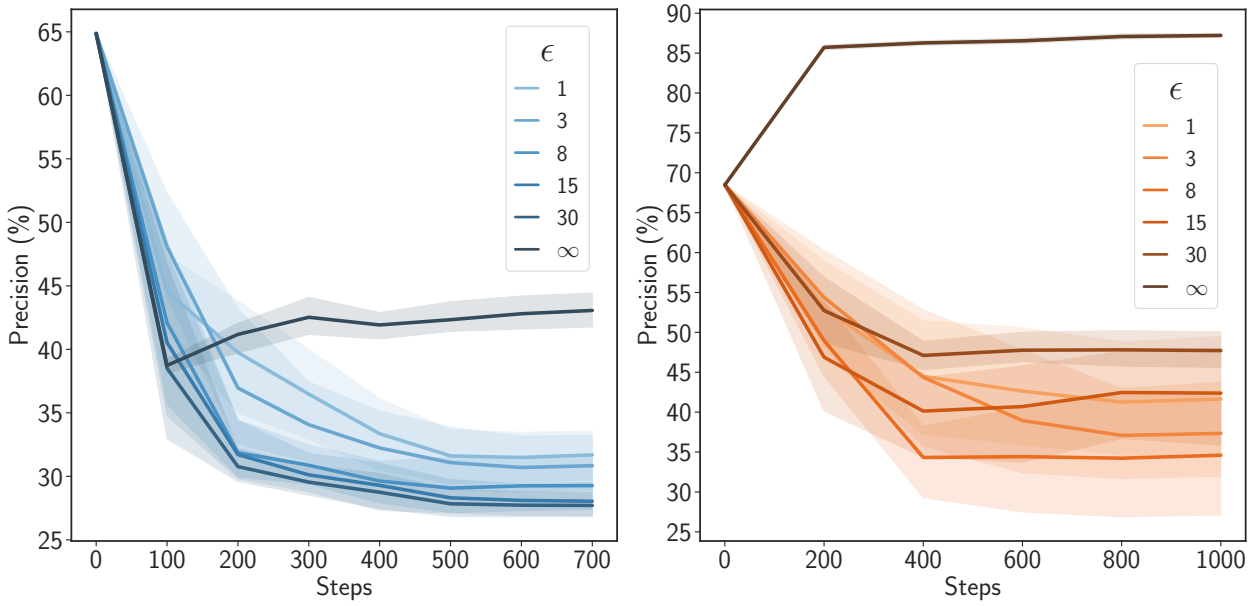
Language	**ISO**	**Family**	**Script**	**Tokens** (M)	**Size** (GiB)
Arabic	AR	Afro-Asiatic	Arabic	2869	28.0
Bulgarian	BG	Indo-European	Cyrillic	5487	57.5
Chinese	ZH	Sino-Tibetan	Chinese	435	63.5
French	FR	Indo-European	Latin	9780	56.8
German	DE	Indo-European	Latin	10297	66.6
Greek	EL	Indo-European	Greek	4285	46.9
Hindi	HI	Indo-European	Devanagari	1803*	20.7*
Indonesian	ID	Austronesian	Latin	22704	148.3
Italian	IT	Indo-European	Latin	4983	30.2
Japanese	JA	Japonic	Japanese	530	69.3
Kiswahili	SW	Niger-Congo	Latin	275	1.6
Korean	KO	Koreanic	Korean	5644	54.2
Portuguese	PT	Indo-European	Latin	8405	49.1
Russian	RU	Indo-European	Cyrillic	23408	278.0
Thai	TH	Kra-Dai	Thai	1834	71.7
Turkish	TR	Turkic	Latin	2736	20.9
Urdu	UR	Indo-European	Arabic	815*	6.2*
Vietnamese	VI	Austro-Asiatic	Latin	24757	137.3

Table 3: Links and references to the datasets we used in our experiments. License information are also available via these links. We ensure that we comply with respective license conditions and only use the data within their intended use policy where applicable.

Dataset	**Download Link**	**Reference**
UD v2.8 (POS)	https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3683	(Nivre et al., 2020; Zeman et al., 2021)
XNLI	https://huggingface.co/datasets/xnli	(Conneau et al., 2018; Lhoest et al., 2021)
TED 2020	https://github.com/UKPLab/sentence-transformers/blob/master/docs/datasets/TED2020.md	(Reimers & Gurevych, 2020)
WikiMatrix	https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix	(Schwenk et al., 2021)
Tatoeba	https://github.com/LBeaudoux/tatoebatools	

Table 4: Overview of the UD v2.8 (Nivre et al., 2020; Zeman et al., 2021) treebanks (test splits only) that we use as test sets in our POS tagging experiments (§3,4) including their respective sizes (number of sentences).

Language	Treebank	# Sentences
AR	Arabic-PADT	680
DE	German-GSD	977
ES	Spanish-GSD	426
HI	Hindi-HDTB	1684
ID	Indonesian-GSD	557
KO	Korean-Kaist	2287
RU	Russian-SynTagRus	6491



(a) POS

(b) XNLI

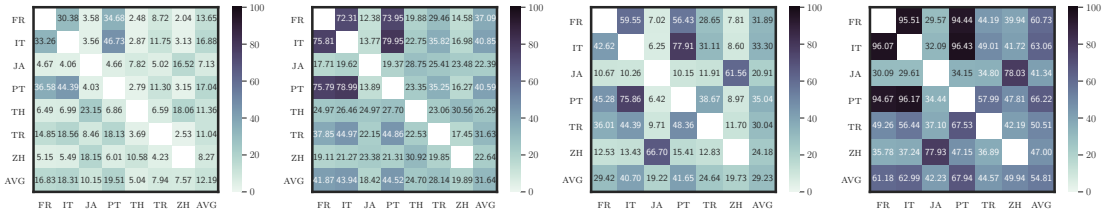
Figure 5: Mean sentence retrieval precision for our TED 2020 splits (different languages/data for POS and XNLI) at layer 8 over the course of fine-tuning with different privacy budgets (ε). $\varepsilon = \infty$ denotes non-private models. Error bands show variation around the mean over 5 random seeds. At $\text{Steps} = 0$, all models are equivalent to the pretrained XLM-R Base. We see that the non-private models can retain (and for XNLI even improve) their multilingual compression much better than the private models and have less variation.

Table 5: **POS** Performance (validation / test accuracy) when fine-tuning XLM-R Base with different privacy budgets (ε). We show results averaged over 5 random seeds each. $\varepsilon = \infty$ denotes non-private models. AVG is the average over the 7 languages. See §3 for our experimental setup. We see that performance increases with decreased privacy across all languages.

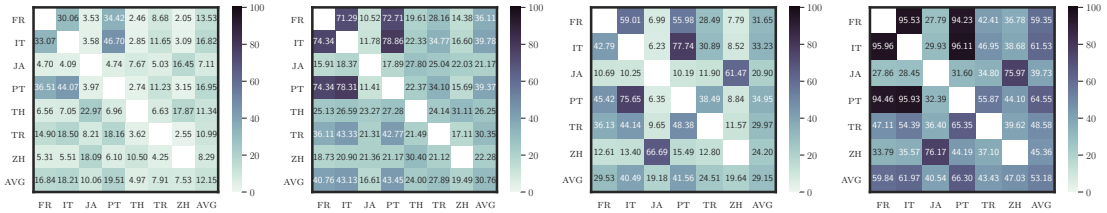
ε	AR	DE	ES	HI	ID	KO	RU	AVG
1	68.3 / 64.6	75.5 / 75.1	79.8 / 79.0	65.0 / 63.3	73.8 / 71.9	66.1 / 54.2	74.8 / 74.0	71.9 / 68.9
3	79.1 / 76.6	86.6 / 86.8	90.3 / 89.3	74.4 / 70.9	82.6 / 79.4	71.1 / 59.4	86.1 / 86.3	81.4 / 78.4
8	81.0 / 77.6	88.4 / 88.3	91.6 / 90.2	78.2 / 75.6	84.2 / 81.2	70.8 / 60.9	87.1 / 87.4	83.0 / 80.2
15	81.3 / 78.4	88.8 / 89.0	92.4 / 90.9	77.0 / 73.2	83.9 / 80.7	71.9 / 61.8	87.7 / 87.8	83.3 / 80.3
30	81.8 / 78.7	89.4 / 89.6	92.9 / 91.5	77.6 / 74.0	84.3 / 81.1	72.3 / 62.2	88.2 / 88.4	83.8 / 80.8
∞	**83.8 / 79.7**	**91.5 / 91.2**	**95.0 / 93.2**	**82.8 / 80.2**	**86.2 / 81.3**	**74.2 / 62.9**	**89.9 / 90.2**	**86.2 / 82.7**

Table 6: **XNLI** Performance (validation / test accuracy) when fine-tuning XLM-R Base with different privacy budgets (ε). We show results averaged over 5 random seeds each. $\varepsilon = \infty$ denotes non-private models. AVG is the average over the 7 languages. See §3 for our experimental setup. We see that performance increases with decreased privacy across all languages. Here, we also particularly observe that the gap between validation and test performance is substantially lower for private models, which shows the strong regularization effect of training with differential privacy.

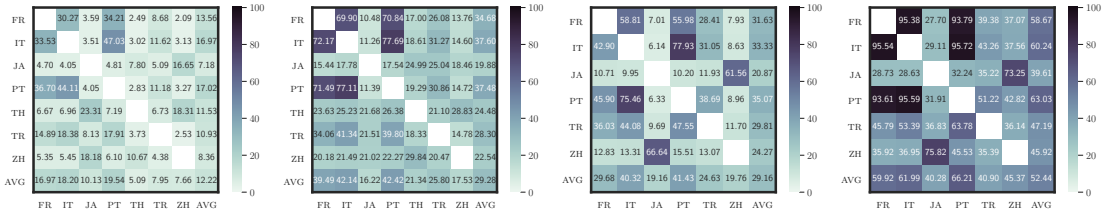
ε	AR	DE	EL	RU	SW	TH	UR	AVG
1	37.3 / 37.4	36.8 / 37.0	36.6 / 36.5	36.3 / 36.2	34.3 / 34.5	35.6 / 35.7	35.6 / 35.6	36.1 / 36.1
3	49.6 / 50.3	49.3 / 51.0	50.8 / 51.5	49.7 / 50.2	45.9 / 47.2	48.8 / 49.5	47.6 / 48.2	48.8 / 49.7
8	55.9 / 56.4	56.8 / 58.5	58.2 / 58.1	56.3 / 57.1	52.0 / 53.2	55.6 / 55.7	53.3 / 53.7	55.5 / 56.1
15	59.1 / 58.3	60.4 / 60.8	61.5 / 60.9	59.7 / 59.5	54.4 / 54.8	58.9 / 58.2	56.4 / 56.1	58.6 / 58.4
30	61.6 / 60.8	63.6 / 63.1	64.8 / 62.0	62.0 / 61.1	56.5 / 57.3	61.2 / 60.2	58.6 / 57.8	61.2 / 60.3
∞	**90.9 / 67.8**	**96.2 / 70.5**	**95.5 / 70.1**	**93.4 / 69.7**	**79.0 / 62.5**	**91.6 / 68.5**	**86.8 / 65.4**	**90.5 / 67.8**



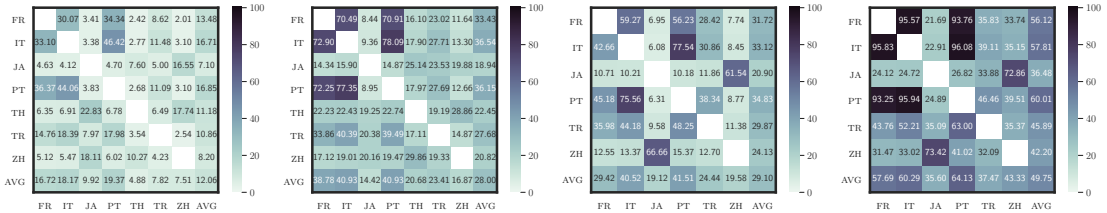
(a) TED, $\varepsilon = 1, l = 0$ (b) TED, $\varepsilon = 1, l = 8$ (c) WM, $\varepsilon = 1, l = 0$ (d) WM, $\varepsilon = 1, l = 8$



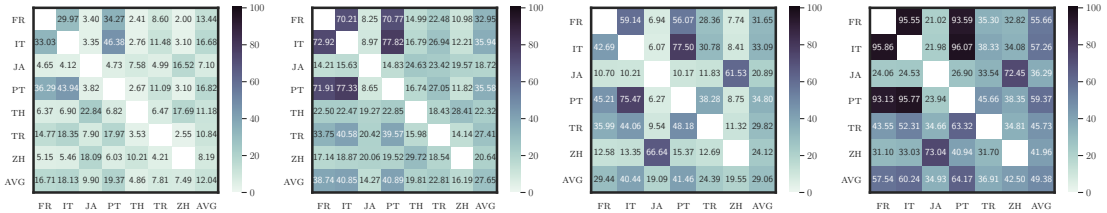
(e) TED, $\varepsilon = 3, l = 0$ (f) TED, $\varepsilon = 3, l = 8$ (g) WM, $\varepsilon = 3, l = 0$ (h) WM, $\varepsilon = 3, l = 8$



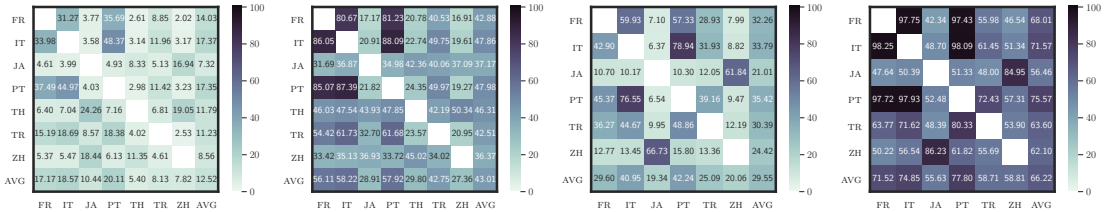
(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$



(m) TED, $\varepsilon = 15, l = 0$ (n) TED, $\varepsilon = 15, l = 8$ (o) WM, $\varepsilon = 15, l = 0$ (p) WM, $\varepsilon = 15, l = 8$

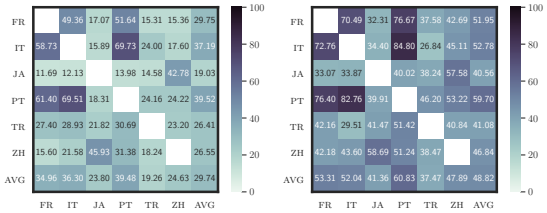


(q) TED, $\varepsilon = 30, l = 0$ (r) TED, $\varepsilon = 30, l = 8$ (s) WM, $\varepsilon = 30, l = 0$ (t) WM, $\varepsilon = 30, l = 8$



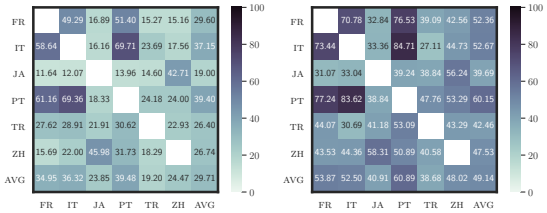
(u) TED, $\varepsilon = \infty, l = 0$ (v) TED, $\varepsilon = \infty, l = 8$ (w) WM, $\varepsilon = \infty, l = 0$ (x) WM, $\varepsilon = \infty, l = 8$

Figure 6: **POS** Sentence retrieval results for the TED 2020 (TED) and WikiMatrix (WM) datasets and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.



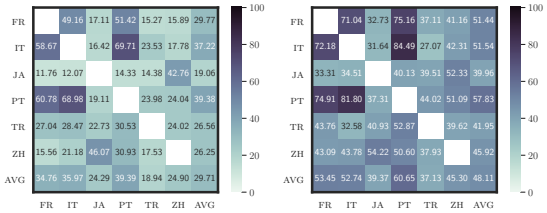
(a) $\varepsilon = 1, l = 0$

(b) $\varepsilon = 1, l = 8$



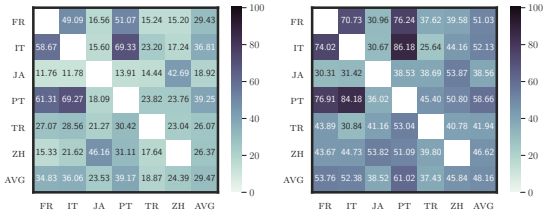
(c) $\varepsilon = 3, l = 0$

(d) $\varepsilon = 3, l = 8$



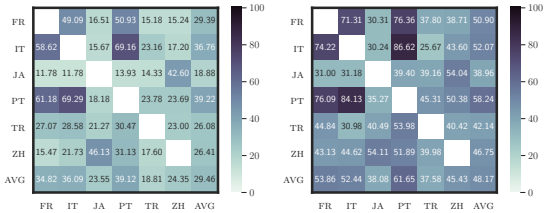
(e) $\varepsilon = 8, l = 0$

(f) $\varepsilon = 8, l = 8$



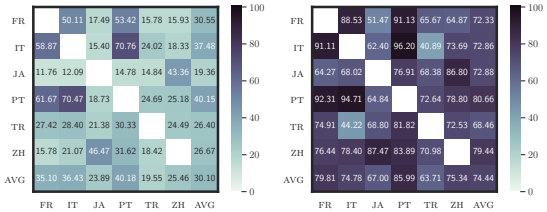
(g) $\varepsilon = 15, l = 0$

(h) $\varepsilon = 15, l = 8$



(i) $\varepsilon = 30, l = 0$

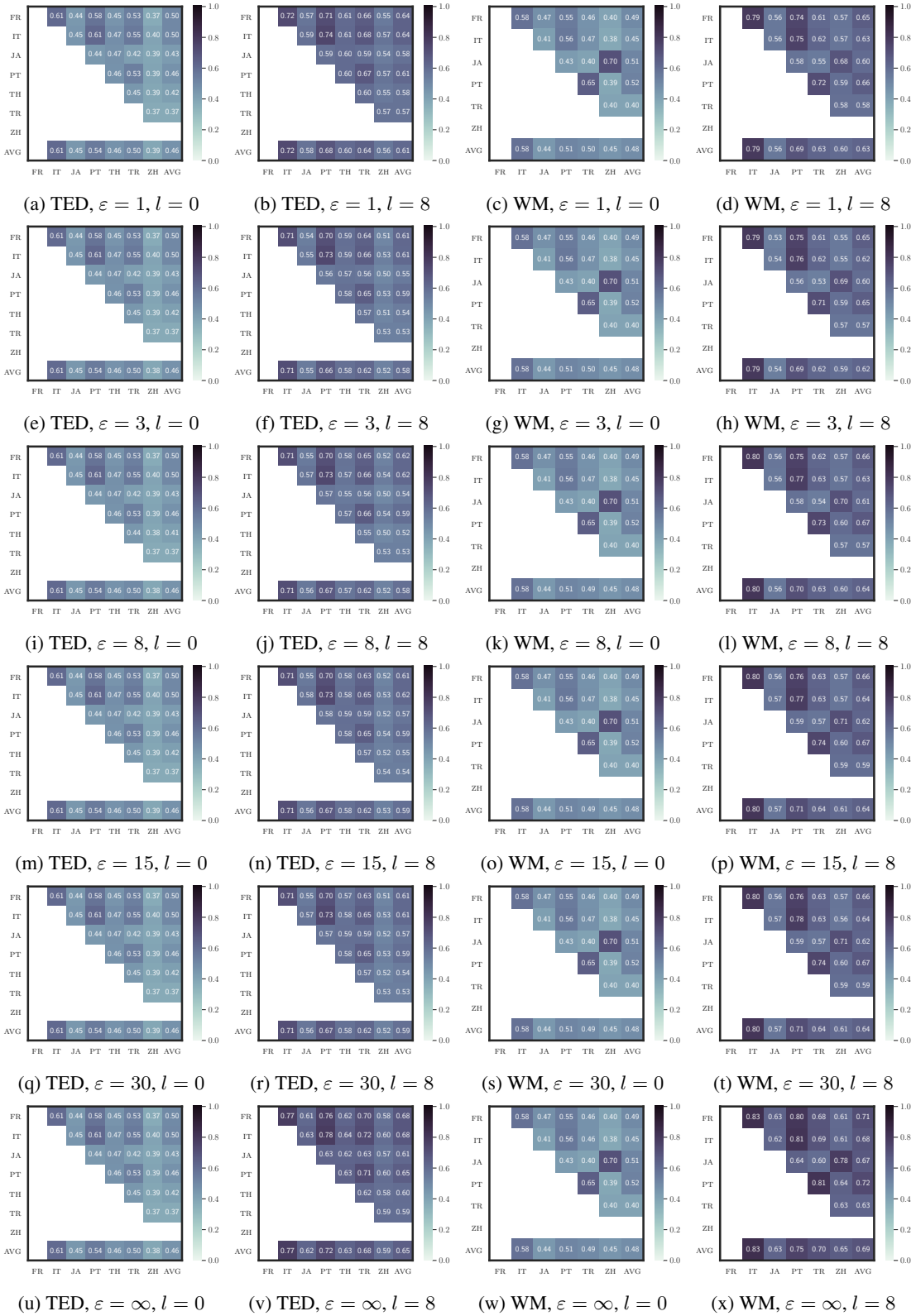
(j) $\varepsilon = 30, l = 8$



(k) $\varepsilon = \infty, l = 0$

(l) $\varepsilon = \infty, l = 8$

Figure 7: **POS** sentence retrieval results for the Tatoeba dataset and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.



(a) TED, $\varepsilon = 1, l = 0$ (b) TED, $\varepsilon = 1, l = 8$ (c) WM, $\varepsilon = 1, l = 0$ (d) WM, $\varepsilon = 1, l = 8$

(e) TED, $\varepsilon = 3, l = 0$ (f) TED, $\varepsilon = 3, l = 8$ (g) WM, $\varepsilon = 3, l = 0$ (h) WM, $\varepsilon = 3, l = 8$

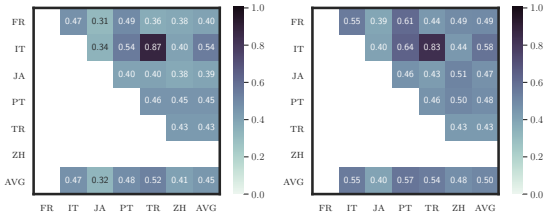
(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$

(m) TED, $\varepsilon = 15, l = 0$ (n) TED, $\varepsilon = 15, l = 8$ (o) WM, $\varepsilon = 15, l = 0$ (p) WM, $\varepsilon = 15, l = 8$

(q) TED, $\varepsilon = 30, l = 0$ (r) TED, $\varepsilon = 30, l = 8$ (s) WM, $\varepsilon = 30, l = 0$ (t) WM, $\varepsilon = 30, l = 8$

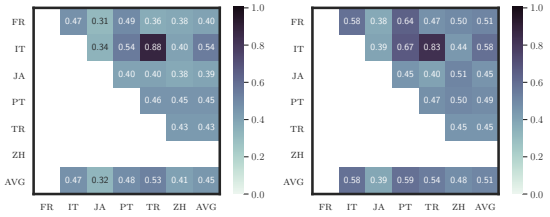
(u) TED, $\varepsilon = \infty, l = 0$ (v) TED, $\varepsilon = \infty, l = 8$ (w) WM, $\varepsilon = \infty, l = 0$ (x) WM, $\varepsilon = \infty, l = 8$

Figure 8: **POS** CKA results for the TED 2020 (TED) and WikiMatrix (WM) datasets and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.



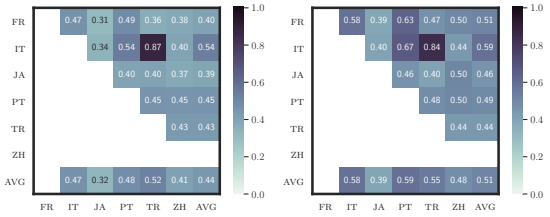
(a) $\varepsilon = 1, l = 0$

(b) $\varepsilon = 1, l = 8$



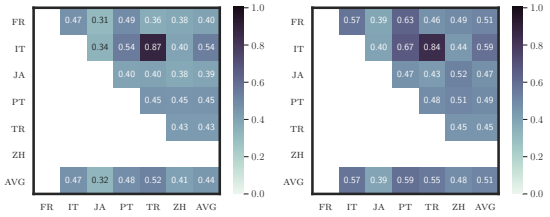
(c) $\varepsilon = 3, l = 0$

(d) $\varepsilon = 3, l = 8$



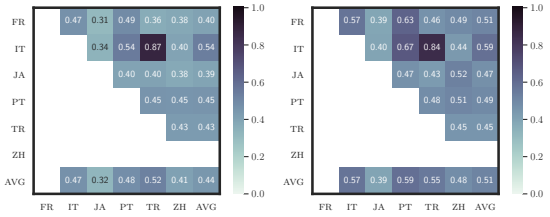
(e) $\varepsilon = 8, l = 0$

(f) $\varepsilon = 8, l = 8$



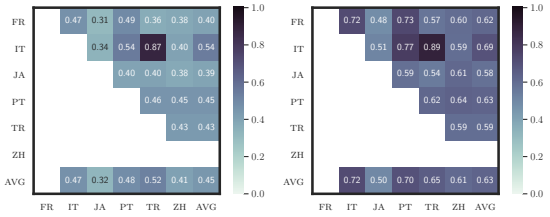
(g) $\varepsilon = 15, l = 0$

(h) $\varepsilon = 15, l = 8$



(i) $\varepsilon = 30, l = 0$

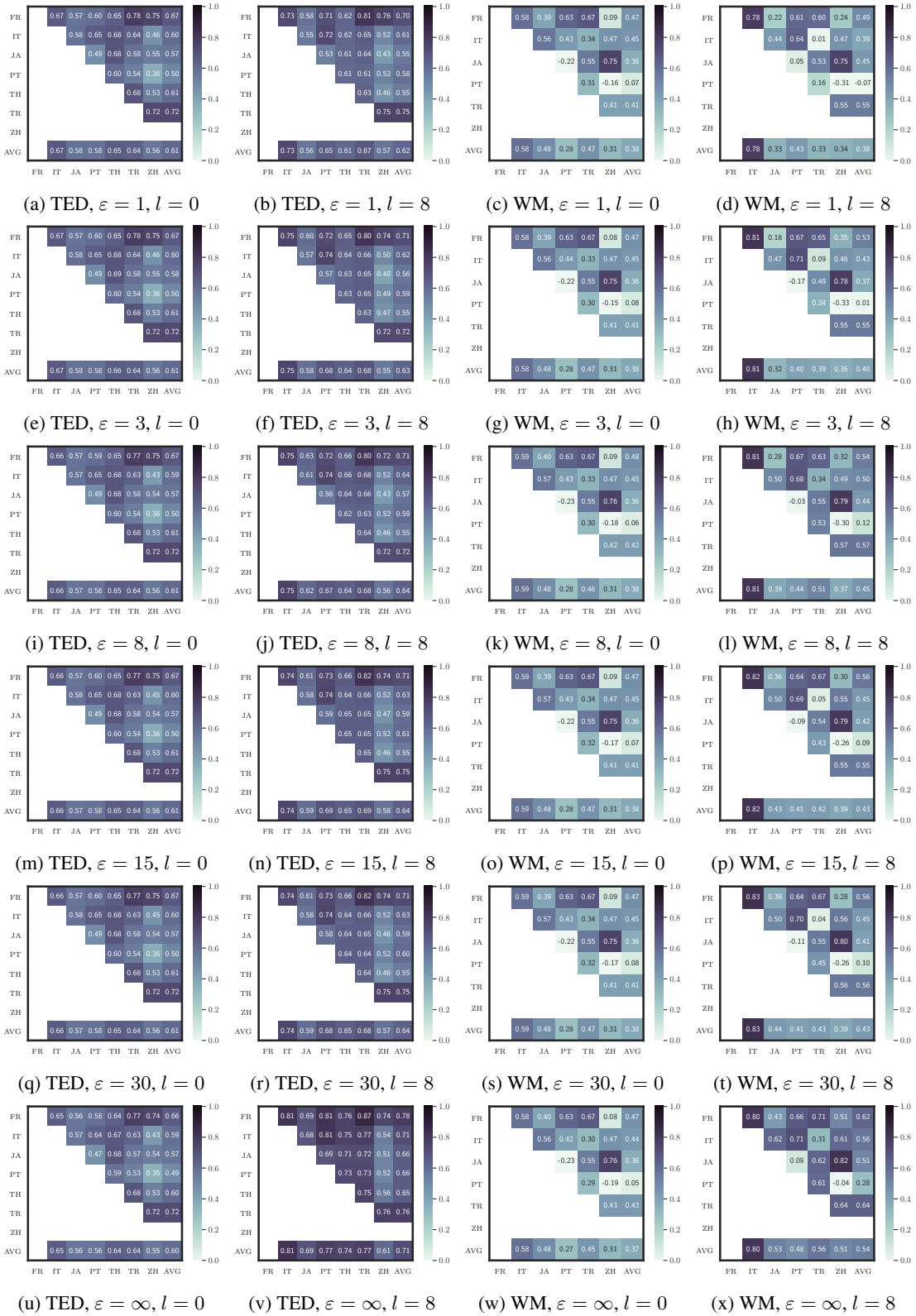
(j) $\varepsilon = 30, l = 8$



(k) $\varepsilon = \infty, l = 0$

(l) $\varepsilon = \infty, l = 8$

Figure 9: **POS** CKA results for the Tatoeba dataset and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.



(a) TED, $\varepsilon = 1, l = 0$ (b) TED, $\varepsilon = 1, l = 8$ (c) WM, $\varepsilon = 1, l = 0$ (d) WM, $\varepsilon = 1, l = 8$

(e) TED, $\varepsilon = 3, l = 0$ (f) TED, $\varepsilon = 3, l = 8$ (g) WM, $\varepsilon = 3, l = 0$ (h) WM, $\varepsilon = 3, l = 8$

(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$

(m) TED, $\varepsilon = 15, l = 0$ (n) TED, $\varepsilon = 15, l = 8$ (o) WM, $\varepsilon = 15, l = 0$ (p) WM, $\varepsilon = 15, l = 8$

(q) TED, $\varepsilon = 30, l = 0$ (r) TED, $\varepsilon = 30, l = 8$ (s) WM, $\varepsilon = 30, l = 0$ (t) WM, $\varepsilon = 30, l = 8$

(u) TED, $\varepsilon = \infty, l = 0$ (v) TED, $\varepsilon = \infty, l = 8$ (w) WM, $\varepsilon = \infty, l = 0$ (x) WM, $\varepsilon = \infty, l = 8$

Figure 10: **POS** RSA results for the TED 2020 (TED) and WikiMatrix (WM) datasets and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.

(a) $\varepsilon = 1, l = 0$

(b) $\varepsilon = 1, l = 8$

(c) $\varepsilon = 3, l = 0$

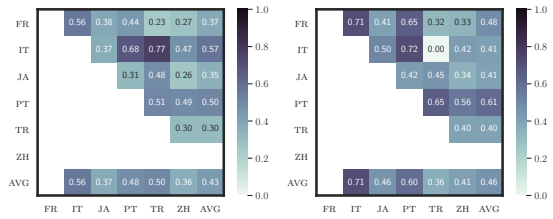
(d) $\varepsilon = 3, l = 8$

(e) $\varepsilon = 8, l = 0$

(f) $\varepsilon = 8, l = 8$

(g) $\varepsilon = 15, l = 0$

(h) $\varepsilon = 15, l = 8$



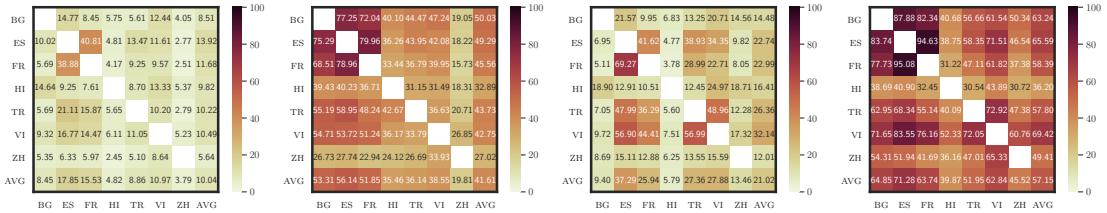
(i) $\varepsilon = 30, l = 0$

(j) $\varepsilon = 30, l = 8$

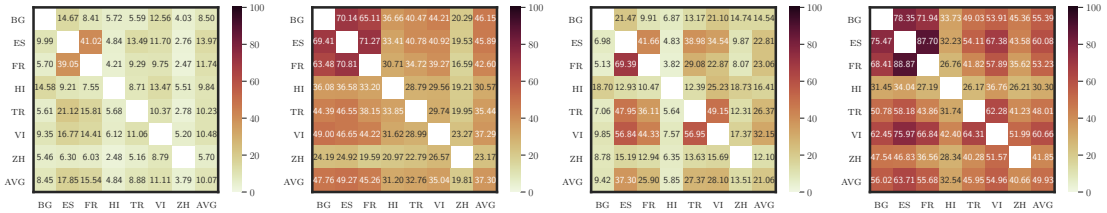
(k) $\varepsilon = \infty, l = 0$

(l) $\varepsilon = \infty, l = 8$

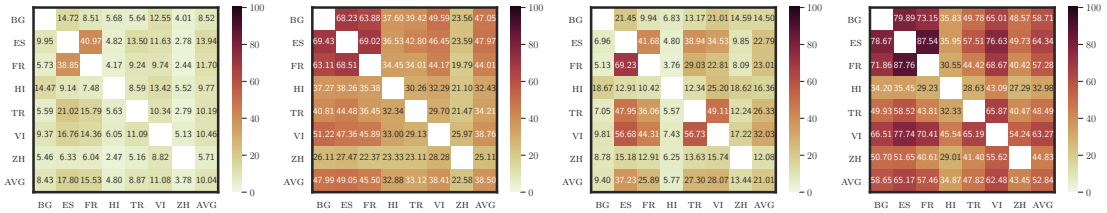
Figure 11: **POS** RSA results for the Tatoeba dataset and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0. Also note that, unlike in CKA (Figure 9), the similarity between IT and TR is high at layer 0 but low at layer 8.



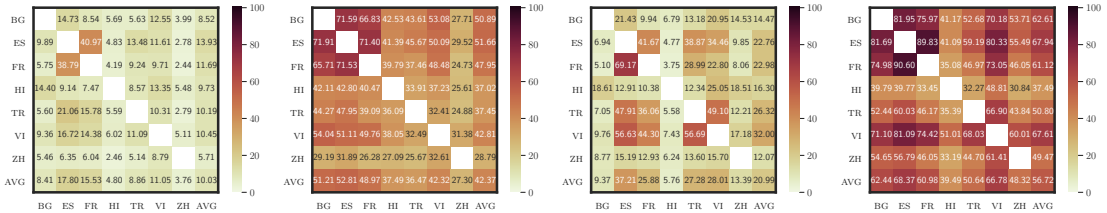
(a) TED, $\varepsilon = 1, l = 0$ (b) TED, $\varepsilon = 1, l = 8$ (c) WM, $\varepsilon = 1, l = 0$ (d) WM, $\varepsilon = 1, l = 8$



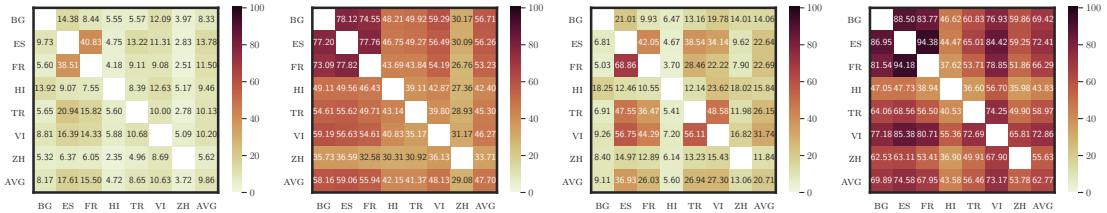
(e) TED, $\varepsilon = 3, l = 0$ (f) TED, $\varepsilon = 3, l = 8$ (g) WM, $\varepsilon = 3, l = 0$ (h) WM, $\varepsilon = 3, l = 8$



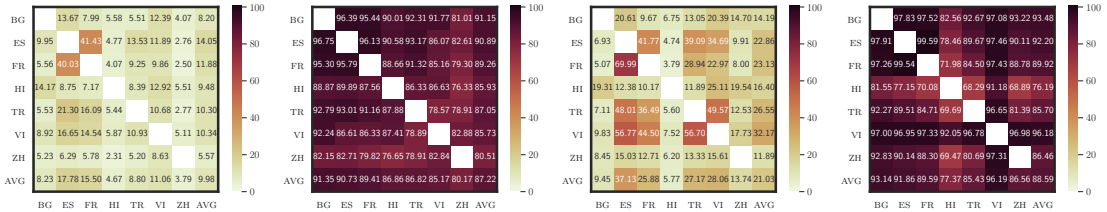
(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$



(m) TED, $\varepsilon = 15, l = 0$ (n) TED, $\varepsilon = 15, l = 8$ (o) WM, $\varepsilon = 15, l = 0$ (p) WM, $\varepsilon = 15, l = 8$

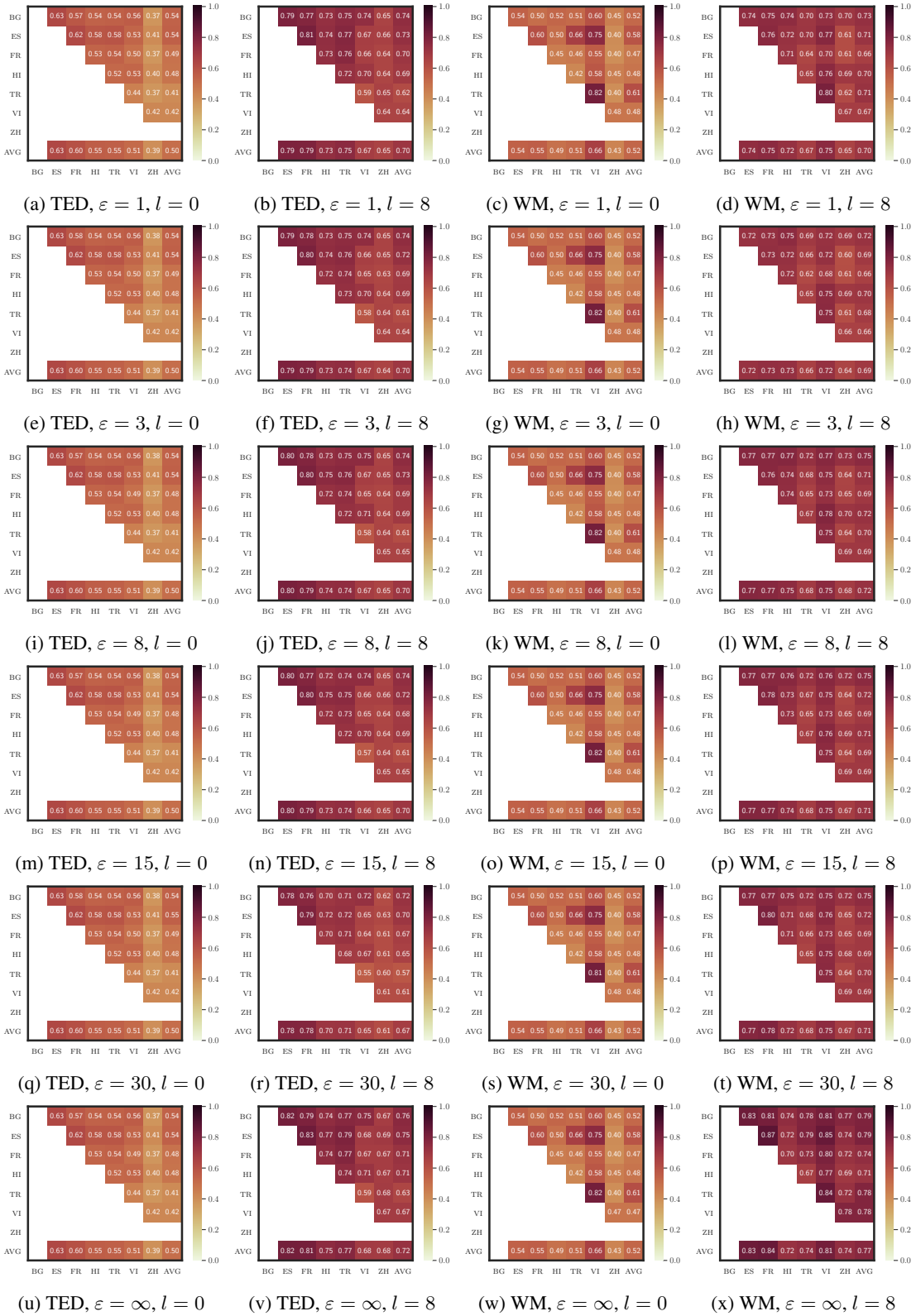


(q) TED, $\varepsilon = 30, l = 0$ (r) TED, $\varepsilon = 30, l = 8$ (s) WM, $\varepsilon = 30, l = 0$ (t) WM, $\varepsilon = 30, l = 8$



(u) TED, $\varepsilon = \infty, l = 0$ (v) TED, $\varepsilon = \infty, l = 8$ (w) WM, $\varepsilon = \infty, l = 0$ (x) WM, $\varepsilon = \infty, l = 8$

Figure 12: **XNLI** Sentence retrieval results for the TED 2020 (TED) and WikiMatrix (WM) datasets and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.



(a) TED, $\varepsilon = 1, l = 0$ (b) TED, $\varepsilon = 1, l = 8$ (c) WM, $\varepsilon = 1, l = 0$ (d) WM, $\varepsilon = 1, l = 8$

(e) TED, $\varepsilon = 3, l = 0$ (f) TED, $\varepsilon = 3, l = 8$ (g) WM, $\varepsilon = 3, l = 0$ (h) WM, $\varepsilon = 3, l = 8$

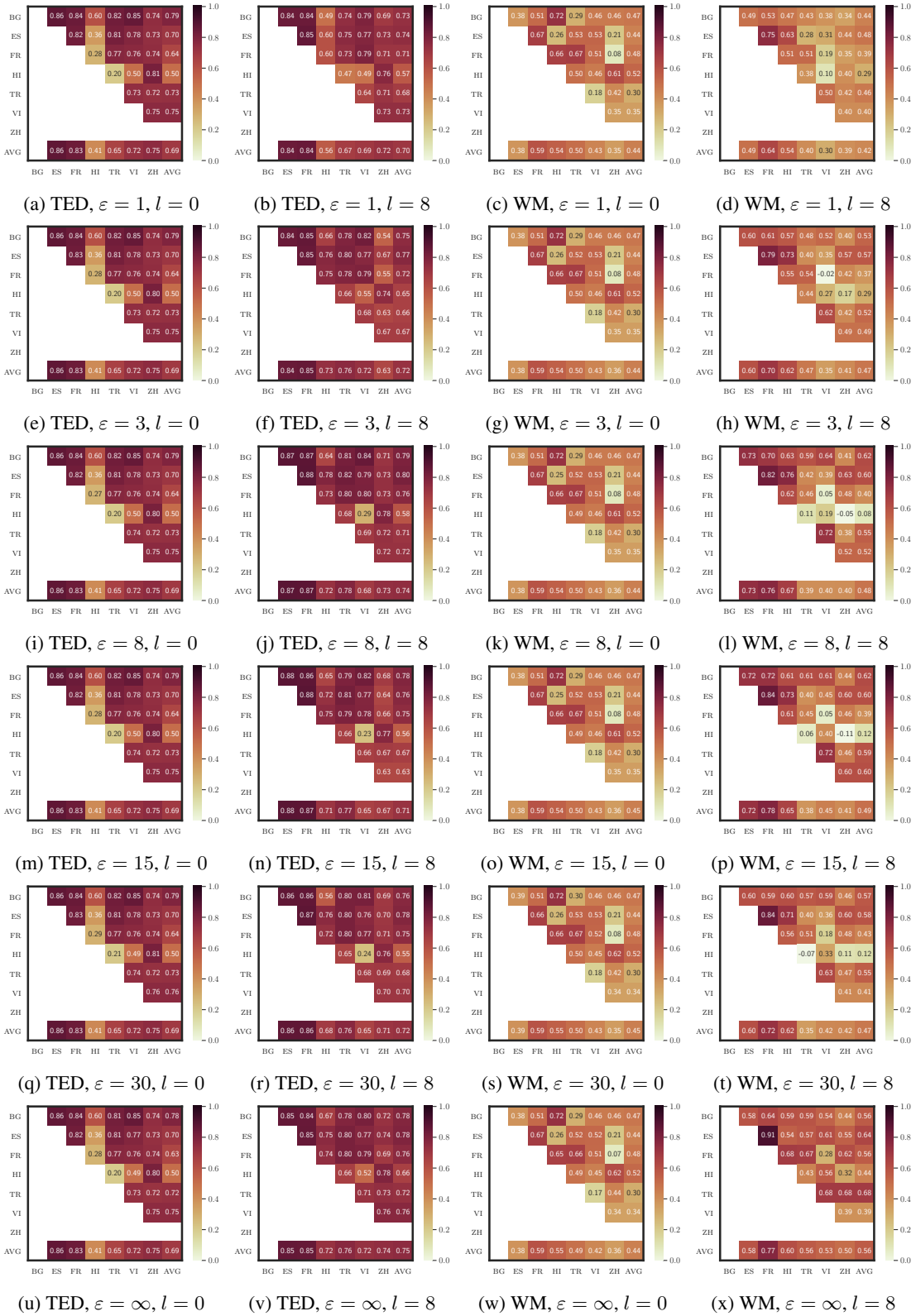
(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$

(m) TED, $\varepsilon = 15, l = 0$ (n) TED, $\varepsilon = 15, l = 8$ (o) WM, $\varepsilon = 15, l = 0$ (p) WM, $\varepsilon = 15, l = 8$

(q) TED, $\varepsilon = 30, l = 0$ (r) TED, $\varepsilon = 30, l = 8$ (s) WM, $\varepsilon = 30, l = 0$ (t) WM, $\varepsilon = 30, l = 8$

(u) TED, $\varepsilon = \infty, l = 0$ (v) TED, $\varepsilon = \infty, l = 8$ (w) WM, $\varepsilon = \infty, l = 0$ (x) WM, $\varepsilon = \infty, l = 8$

Figure 13: **XNLI** CKA results for the TED 2020 (TED) and WikiMatrix (WM) datasets and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.



(a) TED, $\varepsilon = 1, l = 0$ (b) TED, $\varepsilon = 1, l = 8$ (c) WM, $\varepsilon = 1, l = 0$ (d) WM, $\varepsilon = 1, l = 8$

(e) TED, $\varepsilon = 3, l = 0$ (f) TED, $\varepsilon = 3, l = 8$ (g) WM, $\varepsilon = 3, l = 0$ (h) WM, $\varepsilon = 3, l = 8$

(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$

(m) TED, $\varepsilon = 15, l = 0$ (n) TED, $\varepsilon = 15, l = 8$ (o) WM, $\varepsilon = 15, l = 0$ (p) WM, $\varepsilon = 15, l = 8$

(q) TED, $\varepsilon = 30, l = 0$ (r) TED, $\varepsilon = 30, l = 8$ (s) WM, $\varepsilon = 30, l = 0$ (t) WM, $\varepsilon = 30, l = 8$

(u) TED, $\varepsilon = \infty, l = 0$ (v) TED, $\varepsilon = \infty, l = 8$ (w) WM, $\varepsilon = \infty, l = 0$ (x) WM, $\varepsilon = \infty, l = 8$

Figure 14: **XNLI** RSA results for the TED 2020 (TED) and WikiMatrix (WM) datasets and different combinations of privacy budgets (ε) and layers (l). Each heatmap cell corresponds to the average over 5 random seeds. We observe that the overall patterns are highly similar across all levels of privacy, particularly at layer 0.

Table 7: **POS** IsoScores for different combinations of privacy budgets (ε) and layers (l). We show results averaged over 5 random seeds, except for RND and PRE. RND and PRE (added for comparison) denote XLM-R with randomly initialized weights and the original pretrained XLM-R, respectively. We see that the isotropy is fairly uniform across privacy budgets at layer 0 and generally higher at layer 0 than at layer 8. At layer 8, it peaks for non-private ($\varepsilon = \infty$) and our most private ($\varepsilon = 1$) models.

ε	**TED 2020**		**WikiMatrix**		**Tatoeba**	
	$l = 0$	$l = 8$	$l = 0$	$l = 8$	$l = 0$	$l = 8$
RND	0.141	0.132	0.114	0.111	0.054	0.061
PRE	0.187	0.130	0.198	0.112	0.134	0.075
1	0.188	0.054	0.199	0.046	0.135	0.033
3	0.188	0.044	0.199	0.038	0.135	0.027
8	0.187	0.045	0.197	0.038	0.133	0.027
15	0.187	0.047	0.199	0.040	0.135	0.028
30	0.187	0.047	0.199	0.040	0.135	0.028
∞	0.188	0.087	0.199	0.070	0.135	0.051

Table 8: **XNLI** IsoScores for different combinations of privacy budgets (ε) and layers (l). We show results averaged over 5 random seeds, except for RND and PRE. RND and PRE (added for comparison) denote XLM-R with randomly initialized weights and the original pretrained XLM-R, respectively. We see that the isotropy is fairly uniform across privacy budgets at layer 0 and generally higher at layer 0 than at layer 8. At layer 8, it peaks for non-private ($\varepsilon = \infty$) and our most private ($\varepsilon = 1$) models.

ε	**TED 2020**		**WikiMatrix**	
	$l = 0$	$l = 8$	$l = 0$	$l = 8$
RND	0.144	0.134	0.130	0.124
PRE	0.195	0.138	0.210	0.129
1	0.195	0.121	0.211	0.120
3	0.196	0.101	0.211	0.104
8	0.196	0.074	0.212	0.079
15	0.196	0.071	0.212	0.077
30	0.194	0.087	0.210	0.089
∞	0.195	0.182	0.211	0.166