Privacy, Interpretability, and Fairness in the Multilingual Space

Anonymous ACL submission

Abstract

Multilingual generalization or compression is an objective for cross-lingual models in natural language processing (NLP). We explore how the compression sought for in such models aligns with other common objectives in NLP such as performance, differential privacy, interpretability, and fairness. We show that compression, which can be quantified by, e.g., sentence retrieval or centered kernel alignment, is compatible with performance and privacy, but that performance and privacy are at odds, leading to non-linear interactions between compression, performance, and privacy. We also demonstrate that privacy is at odds with interpretability, leading to non-linear interactions between compression, privacy, and interpretability. Finally, while fairness and privacy are generally at odds, we show that in the multilingual space, fairness and privacy have common solutions. In sum, our study shows that if we want to learn multilingual models that exhibit good performance and good generalization properties, *and* are private, interpretable and fair (or any combination thereof), we need to jointly optimize for these inter-dependent objectives.[1]

1 Introduction

Multilingual NLP models facilitate transfer between closely related languages, but less so across typological divides, language families, or scripts (Singh et al., 2019). Cross-lingual generalization is the objective of multilingual models and a result of *multilingual compression* (Dufter and Schütze, 2020; Ravishankar and Søgaard, 2021), i.e., semantically equivalent expressions being encoded in the same way across languages. If multilingual models compartmentalize languages with different scripts, for example, compression is suboptimal.

NLP for low-resource and medium-resource languages today relies heavily on multilingual models. Language models (LMs) such as mBERT (Devlin et al., 2019) and XLM-RoBERTa (XLM-R; Conneau et al., 2020a) are used as pretrained models 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 and Banko, 2021).

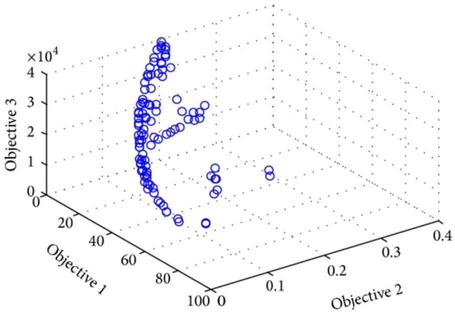


Figure 1: Non-linear interactions in multilingual learning with multiple objectives, e.g., minimizing Objective 1 may lead to high values w.r.t. Objective 3. If we want to optimize for all Objectives 1-3 in such a landscape, we need to do so *jointly*.

With the widespread adaptation of multilingual models comes responsibility. For NLP models to be trustworthy (Pruksachatkun et al., 2021), they must satisfy other requirements, including privacy, interpretability, and fairness. Privacy here means that individual data points cannot be derived from the final model (Dwork et al., 2006). Interpretability means that the training data points that had influence on a prediction can be identified (Koh and Liang, 2017). Fairness refers to equal performance across groups (Hansen and Søgaard, 2021).

Privacy, interpretability, and fairness have primarily been considered in a monolingual context, and it has been assumed that they are largely independent of another, enabling us to develop techniques for one at a time. Our paper presents *a preliminary exploration of the extent to which these objectives align or are at odds*. We explore this in a multilingual context and show how multilingual-

[1] Our code and models are publicly available at [URL].

ity presents options and challenges.[2] Our results indicate that privacy, interpretability, and fairness interact in non-linear ways. Such non-linear interactions mean we have to optimize for all dimensions jointly; see Figure 1 for an illustration.

Contributions In §2, we begin with a theoretical exploration of differential privacy, interpretability, and fairness in the context of multilingual NLP. We show that differential privacy and interpretability are fundamentally at odds,[3] a result which is not limited to the multilingual setting. We also show that differential privacy and fairness, often said to be at odds, 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 interpretability: 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? We compare four metrics for quantifying multilingual compression—sentence retrieval, centered-kernel alignment (CKA; Kornblith et al., 2019), IsoScore (Rudman et al., 2021), 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 interpretable in the sense that they prohibit identifying and tracing back influential examples. In sum, our work shows that *fair and private high-performance multilingual models are possible, even if learning them is challenging. Such models will not be interpretable, however.*

2 Theoretical Exploration

This paper considers 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, interpretability and fairness, with performance and compression in this setting.

Preliminaries We briefly introduce our formal definitions here: A model $\mathcal{M}_D$ induced from a dataset D is said to be ε_p-*differentially private* (Dwork, 2006) iff for all datasets D, D' s.t. $D = D' \cup \{x_{diff}\}$, it holds that $\Pr[\mathcal{M}_D(x_{test}) = y] \leq \exp(\varepsilon_p) \cdot \Pr[\mathcal{M}_{D'}(x_{test}) = y]$ for any x_{test} and y.[4] A model $\mathcal{M}_D$ is said to be *interpretable* iff for an unseen data point, x_{test}, it holds that the most influential training data point under leave-one-out influence, x_{diff} , s.t. $D = D' \cup \{x_{diff}\}$ and $D' = \arg\max_{D''} \Pr[\mathcal{M}_D(x_{test})] - \Pr[\mathcal{M}_{D''}(x_{test})]$,[5] had more influence on $\mathcal{M}_D$ than any other data point $x' \in D$ with $x_{diff} \neq x'$, by some margin ε_i, i.e., $(\Pr[\mathcal{M}_D(x_{test})] - \Pr[\mathcal{M}_{D'}(x_{test})]) > \exp(\varepsilon_i) \cdot (\Pr[\mathcal{M}_D(x_{test})] - \Pr[\mathcal{M}_{D''}(x_{test})])$ for $D = D'' \cup \{x'\}$. Finally, a model $\mathcal{M}$ is said to be fair if for a group partitioning $g(D) \rightarrow D_{g_1}, \ldots, D_{g_n}$ into smaller samples and for some loss function ℓ, e.g., 0-1 loss, $\ell(\mathcal{M}(D_{g_i})) = \ell(\mathcal{M}(D_{g_j}))$. In the following paragraphs, we discuss under what conditions differential privacy, interpretability, and fairness are at odds, and under what conditions they are compatible, and how these conditions align with common scenarios in multilingual NLP.

Differential Privacy and Interpretability We first derive the result that differential privacy and interpretability, as defined in the above, are fundamentally at odds. This result is not limited to the multilingual setting. To see this, recall a model $\mathcal{M}_D$ is differentially private when $\Pr[\mathcal{M}_D(x_{test}) = y] \leq \exp(\varepsilon_p) \cdot \Pr[\mathcal{M}_{D'}(x_{test}) = y]$. $\mathcal{M}_D$ is said to be interpretable if $(\Pr[\mathcal{M}_D(x_{test})] - \Pr[\mathcal{M}_{D'}(x_{test})]) > \exp(\varepsilon_i) \cdot (\Pr[\mathcal{M}_D(x_{test})] - \Pr[\mathcal{M}_{D''}(x_{test})])$ for $D = D'' \cup \{x'\}$. Assume $\mathcal{M}_D$ *is* interpretable. We show that the interpretability of $\mathcal{M}_D$ implies that $\mathcal{M}_D$ is not differentially private. This follows from the fact that since $\mathcal{M}_D$ is interpretable, by definition of interpretability, there is at least one data

[2] We are, to the best of our knowledge, also first to consider differential privacy in a multilingual setting.

[3] Naidu et al. (2021) show that differential privacy and GradCAM (Selvaraju et al., 2019), a feature attribution method, are compatible, but no one has, to the best of our knowledge, considered the interaction of differential privacy and instance-based interpretability.

[4] 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.

[5] It is not feasible to compute leave-one-out influence in practice, but we rely on a commonly used approximation (Pruthi et al., 2020).

point x_{diff} that is revealed by a margin ε_i when applying $\mathcal{M}_D$ to unseen data points. For any ε_i, there is thus a corresponding value ε_p such that $\mathcal{M}_D$ is *not* ε_p-differentially private.[6] Since the relation between ε_i and ε_p is monotonic, there is no optimal trade-off between the two.

Differential Privacy and Fairness Fairness and privacy are occasionally at odds, as shown in Bagdasaryan et al. (2019); Agarwal (2021),[7] but in the multilingual setting, fairness and privacy can be compatible. We first note that there is a trivial solution to obtaining differential privacy and fairness (a joint optimum), namely randomness. Next, imagine a perfectly compressed multilingual model $\mathcal{M}_D$ trained on parallel data from $|L|$ languages, $D = \{\ldots, i_1, \ldots, i_{|L|}, \ldots\}$ with i_j and i_k being translation equivalents. Since $\mathcal{M}_D$ is perfectly compressed, at any layer l, the representation of i_j is identical to i_k, i.e., $\mathcal{M}^l_D(i_j) = \mathcal{M}^l_D(i_k)$. This means, by definition, that $\mathcal{M}_D$ is ε-differentially private (for any ε). If fairness is defined in terms of groups that correspond to the set of languages L, as in Choudhury and Deshpande (2021); Anonymous (2022b); Wang et al. (2021), $\mathcal{M}_D$ is also fair. For any other group partitioning, this is not necessarily true, but it should be easy to see that nothing prevents $\mathcal{M}_D$ from being fair, just because it is private. In fact, if $\mathcal{M}_D$ is fair in just one of the source languages $s \in L$, it is globally fair.

3 Experimental Setup

In our experiments, we probe the relation between the performance and compression of fine-tuned multilingual language models, and their privacy and interpretability. We rely on a commonly used multilingual pretrained language model, which we fine-tune with different levels of (ε, δ)-differential privacy on two tasks and probe using metrics of compression and interpretability.[8] This section presents the pretrained language model, the tasks, the training protocol, the metrics of compression and interpretability, and the evaluation procedure.

Model We use a pretrained XLM-R Base (Conneau et al., 2020a), which has ~277M parameters.

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 1: 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.

Tasks and Data We fine-tune in a zero-shot cross-lingual transfer setting for POS tagging and 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, thereby eliminating potential confounding factors.[9]

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), exclude English,[10] 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

[6] This result generalizes to (ε, δ)-differential privacy if we assume $\mathcal{M}_D$ is influenced by more than δ training points.

[7] Several authors have considered practical trade-offs, including Jagielski et al. (2019), Lyu et al. (2020), Pannekoek and Spigler (2021), and Liu et al. (2021a).

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

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

[10] 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 1).

Russian-SynTagRus. Appendix Table 3 lists the treebanks' sizes.[11]

For NLI, we rely on the XNLI dataset (Conneau et al., 2018), which contains premise–hypothesis–label triples 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. (2021), we use DP-AdamW (i.e., the DP-SGD algorithm (Abadi et al., 2016) applied to the AdamW optimizer with default hyperparameters (Loshchilov and Hutter, 2019; Kingma and Ba, 2015)) to train with (ε, δ)-DP. We evaluate 6 different privacy budgets with $\varepsilon \in \{1, 3, 8, 15, 30, \infty\}$.[12] 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.[13] 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. (2021).[14,15] 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 and Talwar, 2019; Papernot and Steinke, 2021). 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. (2021) found to suit the NLI tasks in the GLUE benchmark (Wang et al., 2018) well.[16] Accordingly, we train with a maximum sequence length of 128 for 10 epochs with a total batch size of 96 for POS and 30 epochs with batch size 512 for XNLI.[17] 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 clipping 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 and Schwenk, 2019; Dufter and Schütze, 2020; Libovický et al., 2020; Ravishankar and Søgaard, 2021; Liu et al., 2021b; Maronikolakis et al., 2021). Dufter and Schütze (2020) quantify the degree of multilingual compression using bidirectional sentence retrieval precision as follows:[18]

[11] Regardless of test split size, each language contributes equally to the mean accuracy reported in Figure 2.

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

[13] 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.

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

[15] 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.

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

[17] 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., 2021) and Appendix D.4 in (Tramèr and Boneh, 2021)).

[18] Note that Dufter and 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.

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

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.[19]

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

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

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 and 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). Anonymous (2022a) also use CKA as a metric of multilingual compression.

IsoScore (Rudman et al., 2021) is an isotropy metric that quantifies the degree to which a point cloud uniformly utilizes the vector space. In our context, this point cloud corresponds to the n mean-pooled sub-word representations 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 downstream task performance (Zhou et al., 2019; Wang et al., 2020; Zhou et al., 2021; Rajaee and Pilehvar, 2021, *inter alia*). High isotropy also means languages are not compartmentalized and should therefore correlate with high compression. The IsoScore algorithm is outlined in Appendix D.

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 and Baroni, 2018; Chrupała, 2019; Chrupała and Alishahi, 2019; Lepori and 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/distance metrics, and similarity is typically measured using a rank-based correlation metric such as Spearman's ρ (Diedrichsen and Kriegeskorte, 2017).

Quantifying Instance-based Interpretability Instance-based interpretability metrics can help us gain an understanding of the inner workings of a model (Koh and Liang, 2017; Yeh et al., 2018; Charpiat et al., 2019; Koh et al., 2019; Pruthi et al., 2020; Basu et al., 2020; K and Søgaard, 2021; Zhang et al., 2021; Kong and 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 TracInCP,[20] 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:

$$\mathrm{TracInCP}(z, z') = \sum_{i=1}^{k} \eta_i \nabla \ell(\theta_i, z) \cdot \nabla \ell(\theta_i, z'), \tag{3}$$

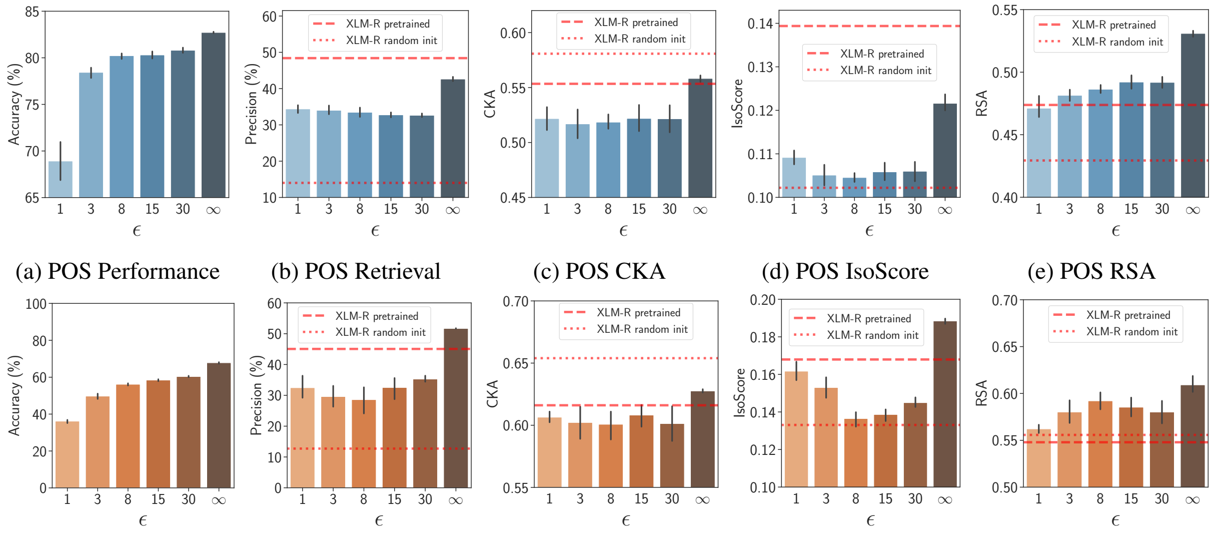
where η_i is the learning rate and $\nabla \ell(\theta_i, z)$ is the gradient of the loss w.r.t. the model parameters θ_i and inputs z for the i-th model checkpoint. We will use TracInCP as an approximation of interpretability in our experiments.

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

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

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



(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 2: 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.

computed using distinct evaluation corpora comprising parallel sentences for all possible 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 and 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,[21,22,23] numbers chosen based on availability and memory constraints for the IsoScore computation.

Following Dufter and 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 and Shutova, 2020; de Vries et al., 2020; Muller et al., 2021). We compute the sentence retrieval precision between language pairs and take the mean.[24] 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.[25] 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.[26] 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. For comparison, we also compute all metrics for the original pretrained XLM-R model and for XLM-R with randomly initialized weights.

4 Results

Privacy, Compression, Performance We now empirically investigate the relationship between privacy, multilingual compression, and cross-lingual transfer performance. We present aggregated results in Figure 2 and non-aggregated results in Appendix F. We observe that the zero-shot accuracy *decreases* as we fine-tune with stronger privacy guarantees (Figures 2a and 2f), which is expected due to the *privacy–utility tradeoff* (Geng et al.,

[21] https://tatoeba.org

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

[23] 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.

[24] Sentence retrieval is bidirectional (see Eq. 1). Given $|L|$ languages, we therefore average over the full $\mathbb{R}^{|L|\times|L|}$ language pair matrix, only excluding the main diagonal.

[25] 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.

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

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., 2021; Yu et al., 2021) 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 and Boneh, 2021; McMahan et al., 2018). Note 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, which shows that training with differential privacy, like other noise injection schemes (Bishop, 1995), is also a form of regularization.

Figures 2b and 2g 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 2d, 2i) and also similar, albeit less pronounced for CKA (Figures 2c, 2h).[27] RSA, in contrast, exhibits very low scores for highly private models; see Appendix E for an explanation.

Overall, 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., $\varepsilon = 8$). At first thought, it may seem counter-intuitive that multilingual compression and cross-lingual transfer performance are not strictly correlated. However, in the fine-tuning setting, it is possible to disregard all task-specific knowledge in favor of multilingual compression, which ultimately leads to poor performance. Vice-versa, a model may exploit 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 model, which scores highly in multilingual compression (as displayed in Figure 2) 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. (2021b) 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. 4), this finding clearly shows that there are exceptions. Future work may investigate this further.

Lastly, sentence retrieval and CKA scores are always highest between typologically similar languages and languages over-represented in pretraining (see Table 1 for a comparison across all languages) *across all levels of privacy*, as shown by the non-aggregated results in the Appendix Figures 5–12. This finding thus extends conclusions from prior work (Pires et al., 2019; Wu and Dredze, 2019; K et al., 2020; Lauscher et al., 2020) to private models.

5 Are more multilingual models less interpretable?

Metric To answer this question, we introduce a new metric, termed InfU (**Inf**luence **U**niformity), which is based on the TracInCP influence scores for each training example in the (multi-)parallel dataset $D = \{\ldots, i_1, \ldots, i_{|L|}, \ldots\}$, where i_j and i_k are translation equivalents. As its name suggests, InfU is a measure of uniformity, based on the idea that for a perfectly multilingual model, the following equation should hold $\forall j, k, q, r \in L$:

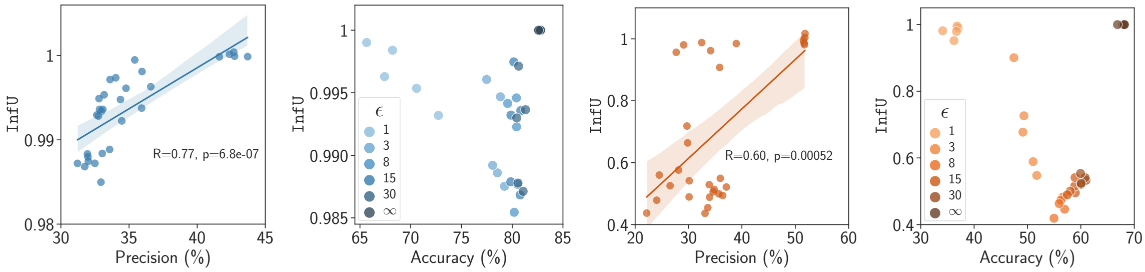
$$\mathrm{TracInCP}(i_j, i_k) = \mathrm{TracInCP}(i_q, i_r) \qquad (4)$$

We compute InfU for a model $\mathcal{M}$ and the set of translation equivalent examples $i = \{i_1, \ldots, i_{|L|}\}$ as

$$\mathrm{InfU}_{\mathcal{M}}(i) = \frac{1}{|L|}\sum_{k}^{|L|} \mathrm{H}(\sigma(\mathrm{TracInCP}(i_k, i))), \qquad (5)$$

where H is the entropy with $log_{|L|}$ and σ is a softmax used to obtain a probability distribution over the list of influence scores. InfU is maximized ($\mathrm{InfU} = 1$) for uniform influence scores, fulfilling Eq. 4. In this scenario of maximum uncertainty our model is also the least interpretable because we do not know the languages of the most or least influential examples for another example's prediction.

[27] Note the randomly initialized XLM-R model scores particularly highly in CKA. This phenomenon 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 InfU – Retrieval (b) POS InfU – Perf. (c) XNLI InfU – Retrieval (d) XNLI InfU – Perf.

Figure 3: Linear fit and Pearson correlation between the influence uniformity InfU and sentence retrieval precision (3a, 3c) and InfU versus downstream performance for different levels of privacy (3b, 3d). We see significant positive correlations between retrieval precision and InfU, suggesting a negative correlation between multilingual compression and interpretability. For task performance, we see the trade-off between interpretability (InfU) and privacy, which aligns with our theoretical expectations (§2).

While it essentially follows from the above definition that multilingual compression and interpretability are at odds, we use it to study to what extent interpretability aligns with privacy, our metrics used in §4, and cross-lingual transfer performance.

Setup We use 1000 examples from the respective training dataset for both POS and XNLI. We use the last 3 model checkpoints with their corresponding learning rates to compute TracInCP scores.[28] Checkpoints were stored every 100 training steps. We use the gradients w.r.t. all model parameters.

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 3a and 3c. For both tasks, there is a significant ($p < 0.05$) strong positive correlation between the InfU score and the multilingual compression as determined through sentence retrieval. This result further supports the idea that *multilingual compression is at odds with interpretability*.

We also see that highly private and correspondingly low-performing models score highly in InfU (Figures 3b, 3d), which suggests that they are not interpretable. The same applies to the non-private and correspondingly high-performing models. For medium levels of privacy we, again, see a trade-off characterized by lower InfU, i.e., better interpretability, and medium performance. It thus also becomes clear that, unless optimizing for them jointly, *high privacy, interpretability, and performance are not compatible*, because the high-performing models are strictly low in privacy and interpretability and the models high in privacy are strictly low in performance and interpretability. We can at best achieve a satisfactory trade-off between these three objectives.

6 Conclusion

In this work, we conducted a preliminary investigation of the interactions between multilingual compression, privacy, interpretability, and fairness in the context of multilingual language models. We first found, through theoretical exploration, that privacy and interpretability are generally incompatible, both inside and outside the multilingual space. We also established that privacy and fairness, although often thought to be fundamentally at odds, are theoretically compatible in the multilingual space.

We further explored the space empirically. Our results overall support the theoretical expectations we laid out; we found that high multilingual compression can be achieved either by optimizing for performance or by optimizing for privacy. Likewise, by trading off privacy and performance, we compromise multilingual compression. In other words, the interactions between these objectives are non-linear. Ideally, however, we want models to do well in all of these dimensions, which remains an open problem. We hope that our study will spark future research in this direction.

Finally, we introduced a new metric, the influence uniformity, to empirically validate the theoretical idea that privacy and interpretability are incompatible and that the interactions between privacy, interpretability, and multilingual compression are, therefore, also non-linear.

[28] Since the learning rate changes constantly, we use the learning rate from the end of each checkpointing interval.

7 Ethical Aspects and Broader Impact

Beyond performance metrics, it is crucial to study objectives such as privacy, interpretability, and fairness in (multilingual) NLP. 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 see practically no potential for misuse of the results of our work.

References

Martín Abadi, Andy Chu, Ian J. Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, Vienna, Austria. ACM.

Sushant Agarwal. 2021. Trade-offs between fairness and privacy in machine learning. In *IJCAI 2021 Workshop on AI for Social Good*. International Joint Conference of Artificial Intelligence (IJCAI).

Anonymous. 2022a. Enhancing cross-lingual transfer by manifold mixup. In *Submitted to the 10th International Conference on Learning Representations (ICLR)*. Under review.

Anonymous. 2022b. Fairness in representation for multilingual NLP: Insights from controlled experiments on conditional language modeling. In *Submitted to the 10th International Conference on Learning Representations (ICLR)*. Under review.

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

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 15453–15462, Vancouver, BC, Canada. Curran Associates, Inc.

Samyadeep Basu, Xuchen You, and Soheil Feizi. 2020. 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*, pages 715–724, Online. PMLR.

Chris M. Bishop. 1995. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116.

Diane Bouchacourt and Marco Baroni. 2018. 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*, pages 981–985, Brussels, Belgium. Association for Computational Linguistics.

Clément L. Canonne, Gautam Kamath, and Thomas Steinke. 2020. The discrete gaussian for differential privacy. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 15676–15688, Online. Curran Associates, Inc.

Guillaume Charpiat, Nicolas Girard, Loris Felardos, and Yuliya Tarabalka. 2019. Input similarity from the neural network perspective. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 5343–5352, Vancouver, BC, Canada. Curran Associates, Inc.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *arXiv preprint*.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 12710–12718, Online. AAAI Press.

Grzegorz Chrupała. 2019. Symbolic inductive bias for visually grounded learning of spoken language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6452–6462, Florence, Italy. Association for Computational Linguistics.

Grzegorz Chrupała and Afra Alishahi. 2019. Correlating neural and symbolic representations of language. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2952–2962, Florence, Italy. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, and Malvina Nissim. 2020. What's so special about BERT's layers? a closer look at the NLP pipeline in monolingual and multilingual models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4339–4350, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörn Diedrichsen and Nikolaus Kriegeskorte. 2017. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):1–33.

Philipp Dufter and Hinrich Schütze. 2020. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4423–4437, Online. Association for Computational Linguistics.

Cynthia Dwork. 2006. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP), Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12, Venice, Italy. Springer.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography (TCC)*, page 265–284, Berlin / Heidelberg, Germany. Springer-Verlag.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407.

Shimon Edelman. 1998. Representation is representation of similarities. *Behavioral and Brain Sciences*, 21(4):449–467.

Kawin Ethayarajh. 2019. 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)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation degeneration problem in training natural language generation models. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA. OpenReview.net.

Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. 2020. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 108 of *Proceedings of Machine Learning Research*, pages 89–99, Online. PMLR.

Goran Glavaš and Ivan Vulić. 2021. 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*, pages 3090–3104, Online. Association for Computational Linguistics.

Ivan Habernal. 2021. When differential privacy meets NLP: The devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Petrén Bach Hansen and Anders Søgaard. 2021. Is the lottery fair? evaluating winning tickets across demographics. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3214–3224, Online. Association for Computational Linguistics.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021. 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)*, pages 2208–2222, Online. Association for Computational Linguistics.

Matthew Jagielski, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi Malvajerdi, and Jonathan Ullman. 2019. Differentially private fair learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3000–3008, Long Beach, CA, USA. PMLR.

Karthikeyan K and Anders Søgaard. 2021. Revisiting methods for finding influential examples. *arXiv preprint*.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual BERT: an empirical study. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Online. OpenReview.net.

Gavin Kerrigan, Dylan Slack, and Jens Tuyls. 2020. Differentially private language models benefit from public pre-training. In *Proceedings of the Second Workshop on Privacy in NLP*, pages 39–45, Online. Association for Computational Linguistics.

Muhammad Khalifa, Muhammad Abdul-Mageed, and Khaled Shaalan. 2021. 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*, pages 769–782, Online. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Pang Wei Koh, Kai-Siang Ang, Hubert H. K. Teo, and Percy Liang. 2019. On the accuracy of influence functions for measuring group effects. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 5255–5265, Vancouver, BC, Canada. Curran Associates, Inc.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, Sydney, NSW, Australia. PMLR.

Zhifeng Kong and Kamalika Chaudhuri. 2021. 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. Curran Associates, Inc.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529, Long Beach, CA, USA. PMLR.

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

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. 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)*, pages 4483–4499, Online. Association for Computational Linguistics.

Michael Lepori and R. Thomas McCoy. 2020. 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*, pages 3637–3651, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. 2021. Large language models can be strong differentially private learners. *arXiv preprint*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674, Online. Association for Computational Linguistics.

Jingcheng Liu and Kunal Talwar. 2019. Private selection from private candidates. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing (STOC)*, page 298–309, New York, NY, USA. Association for Computing Machinery.

Wenyan Liu, Xiangfeng Wang, Xingjian Lu, Junhong Cheng, Bo Jin, Xiaoling Wang, and Hongyuan Zha. 2021a. Fair differential privacy can mitigate the disparate impact on model accuracy. *Submitted to the 9th International Conference on Learning Representations (ICLR)*.

Zihan Liu, Genta Indra Winata, Andrea Madotto, and Pascale Fung. 2021b. Preserving cross-linguality of pre-trained models via continual learning. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 64–71, Online. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, New Orleans, LA, USA. OpenReview.net.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. 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*, pages 2355–2365, Online. Association for Computational Linguistics.

Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. Wine is not v i n. on the compatibility of tokenizations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2382–2399, Punta Cana, Dominican Republic. Association for Computational Linguistics.

H. Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning differentially private recurrent language models. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada. OpenReview.net.

Ilya Mironov. 2017. Rényi differential privacy. In *30th IEEE Computer Security Foundations Symposium, (CSF)*, pages 263–275, Santa Barbara, CA, USA. IEEE Computer Society.

Ilya Mironov, Kunal Talwar, and Li Zhang. 2019. Rényi differential privacy of the sampled gaussian mechanism. *arXiv preprint*.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. 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*, pages 2214–2231, Online. Association for Computational Linguistics.

Rakshit Naidu, Aman Priyanshu, Aadith Kumar, Sasikanth Kotti, Haofan Wang, and Fatemehsadat Mireshghallah. 2021. When differential privacy meets interpretability: A case study. In *CVPR 2021 Workshop for Responsible Computer Vision (RCV)*.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Marlotte Pannekoek and Giacomo Spigler. 2021. Investigating trade-offs in utility, fairness and differential privacy in neural networks. *arXiv preprint*.

Nicolas Papernot and Thomas Steinke. 2021. Hyperparameter tuning with renyi differential privacy. *arXiv preprint*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, Vancouver, BC, Canada. Curran Associates, Inc.

Jason Phang, Haokun Liu, and Samuel R. Bowman. 2021. 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*, pages 529–538, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Yada Pruksachatkun, Anil Ramakrishna, Kai-Wei Chang, Satyapriya Krishna, Jwala Dhamala, Tanaya Guha, and Xiang Ren, editors. 2021. *Proceedings of the First Workshop on Trustworthy Natural Language Processing*. Association for Computational Linguistics, Online.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems (NeurIPS)*, Online. Curran Associates, Inc.

Sara Rajaee and Mohammad Taher Pilehvar. 2021. 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)*, pages 575–584, Online. Association for Computational Linguistics.

Vinit Ravishankar and Anders Søgaard. 2021. The impact of positional encodings on multilingual compression. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 763–777, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2021. IsoScore: Measuring the uniformity of vector space utilization. *arXiv preprint*.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. 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*, pages 1351–1361, Online. Association for Computational Linguistics.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2019. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359.

Manuel Senge, Timour Igamberdiev, and Ivan Habernal. 2021. One size does not fit all: Investigating strategies for differentially-private learning across nlp tasks. *arXiv preprint*.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. 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)*, pages 47–55, Hong Kong, China. Association for Computational Linguistics.

Florian Tramèr and Dan Boneh. 2021. Differentially private learning needs better features (or much more data). In *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, Online. OpenReview.net.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. 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*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Cindy Wang and Michele Banko. 2021. 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*, pages 121–129, Online. Association for Computational Linguistics.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Assessing multilingual fairness in pre-trained multimodal representations. *arXiv preprint*.

Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020. Improving neural language generation with spectrum control. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, Online. OpenReview.net.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. 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)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Chih-Kuan Yeh, Joon Sik Kim, Ian En-Hsu Yen, and Pradeep Ravikumar. 2018. Representer point selection for explaining deep neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 9311–9321, Montréal, Canada. Curran Associates, Inc.

Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, Graham Cormode, and Ilya Mironov. 2021. Opacus: User-friendly differential privacy library in pytorch. In *NeurIPS 2021 Workshop Privacy in Machine Learning*, Online.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A. Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. 2021. Differentially private fine-tuning of language models. *arXiv preprint*.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Gabrielė Aleksandravičiūtė, Ika Alfina, Lene Antonsen, Katya Aplonova, Angelina Aquino, Carolina Aragon, Maria Jesus Aranzabe, Bilge Nas Arıcan, H͡órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Deniz Baran Aslan, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Kepa Bengoetxea, Gözde Berk, Yevgeni Berzak, Irshad Ahmad Bhat, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, Kristina Brokaitė, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Lauren

Cassidy, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Mihaela Cristescu, Philemon. Daniel, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Hanne Eckhoff, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Marília Fernanda, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Sebastian Garza, Fabrício Ferraz Gerardi, Kim Gerdes, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Eva Huber, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Kaoru Ito, Tomáš Jelínek, Apoorva Jha, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Sarveswaran K, Hüner Kaşıkara, Andre Kaasen, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Boris Katz, Tolga Kayadelen, Jessica Kenney, Václava Kettnerová, Jesse Kirchner, Elena Klementieva, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Veronika Laippala, Lucia Lam, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phương Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Cheuk Ying Li, Josie Li, Keying Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Krister Lindén, Nikola Ljubešić, Olga Loginova, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Niko Miekka, Karina Mischenkova, Margarita Misirpashayeva, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Keiko Sophie Mori, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Lương Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Adédayọ̀ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Angelika Peljak-Łapińska, Siyao Peng, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Jason Phelan, Jussi Piitulainen, Tommi A Pirinen, Emily Pitler, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andriela Rääbis, Alexandre Rademaker, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Georg Rehm, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Eiríkur Rögnvaldsson, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Benoît Sagot, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Baiba Saulīte, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Dmitry Sichinava, Janine Siewert, Einar Freyr Sigurðsson, Aline Silveira, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Carolyn Spadine, Rachele Sprugnoli, Steinh͡ór Steingrímsson, Antonio Stella, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Zsolt Szántó, Dima Taji, Yuta Takahashi, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Sowmya Vajjala, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Eric Villemonte de la Clergerie, Veronika

Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jing Xian Wang, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, Hanzhi Zhu, Anna Zhuravleva, and Rayan Ziane. 2021. Universal dependencies 2.8. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Wei Zhang, Ziming Huang, Yada Zhu, Guangnan Ye, Xiaodong Cui, and Fan Zhang. 2021. 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)*, pages 5399–5411, Online. Association for Computational Linguistics.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Hammerla. 2019. 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)*, pages 951–962, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyuan Zhou, João Sedoc, and Jordan Rodu. 2019. Getting in shape: Word embedding subspaces. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5478–5484, Macao, China. ijcai.org.

Wenxuan Zhou, Bill Yuchen Lin, and Xiang Ren. 2021. Isobn: Fine-tuning BERT with isotropic batch normalization. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pages 14621–14629, Online. AAAI Press.

A Reproducibility

Our code and trained models are openly available at [URL].

Implementation Our implementation is written in PyTorch version 1.10.0 (Paszke et al., 2019) for Python 3.9.5 and builds on code from the following repositories:

- https://github.com/huggingface/transformers version 4.9.2 (Wolf et al., 2020) for downloading, training, and evaluating models
- https://github.com/lxuechen/private-transformers version 0.1.0 (Li et al., 2021) for DP-training
- https://github.com/pdufter/minimult (Dufter and 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 and McCoy, 2020) for computing RSA scores
- https://github.com/bcbi-edu/p_eickhoff_isoscore (Rudman et al., 2021) for computing IsoScores
- https://github.com/FengNiMa/VAE-TracIn-pytorch (Kong and Chaudhuri, 2021) for computing TracInCP influence scores.

Model We use the pretrained XLM-RoBERTa (Conneau et al., 2020a) base model and tokenizer from https://huggingface.co/xlm-roberta-base.

Data We provide download links and references for the various datasets we used in Table 2.

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 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[29], where we logged our experiments. Although we do not have precise numbers, a highly conservative estimate of the total compute spent including prototyping, hyperparameter 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 and Roth, 2014), also called approximate-DP, which is typically used in practice and relaxes the privacy guarantees by a (small) δ as follows:

A model $\mathcal{M}_D$ induced from a dataset D is said to be *(ε, δ)-differentially private* iff for all datasets D, D' s.t. $D = D' \cup \{x_{diff}\}$, it holds that $\Pr[\mathcal{M}_D(x_{test}) = y] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{M}_{D'}(x_{test}) = y] + \delta$ for any x_{test} and y.

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:

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)

Table 4: Best 5 settings for each task and privacy budget. Includes LR and the corresponding number of random initializations (# seeds).

D IsoScore Algorithm

Algorithm 1 describes the IsoScore algorithm (Rudman et al., 2021).

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 9, 10, and 13) 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. 9). Based on prior work by, for example, Pires et al. (2019), Wu and 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. 10). 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 Detailed Results for Experiments in §4

Figure 4 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.

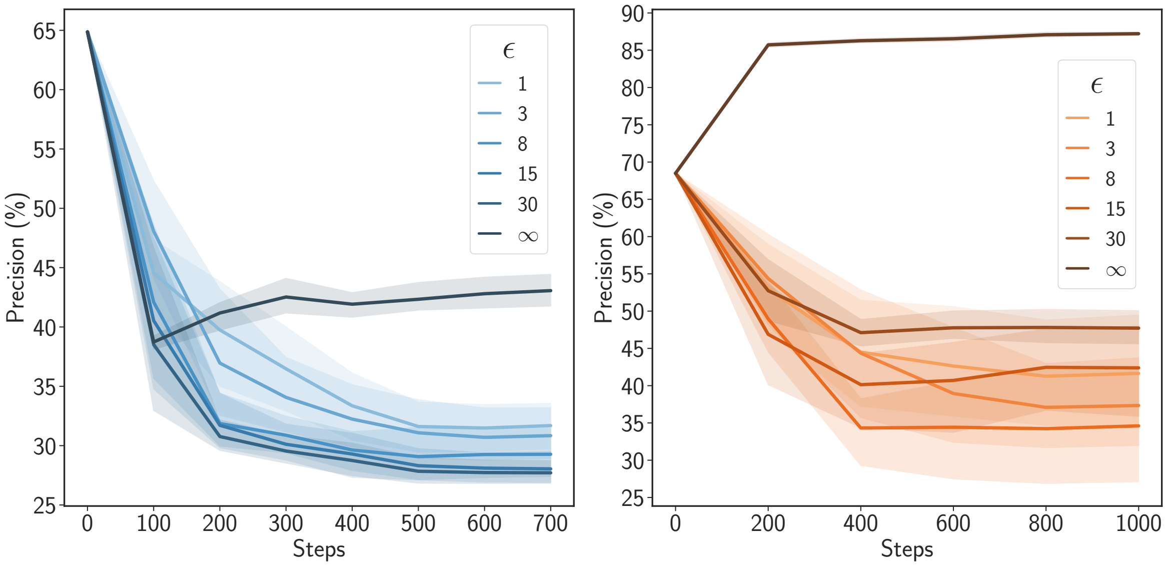
We further present non-aggregated results for

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

[29] https://wandb.ai/

Algorithm 1 IsoScore (Rudman et al., 2021)

```
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
```



(a) POS

(b) XNLI

Figure 4: 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 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.

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 and Gurevych, 2020)
WikiMatrix	https://github.com/facebookresearch/LASER/tree/main/tasks/WikiMatrix	(Schwenk et al., 2021)
Tatoeba	https://github.com/LBeaudoux/tatoebatools	

Table 2: 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.

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

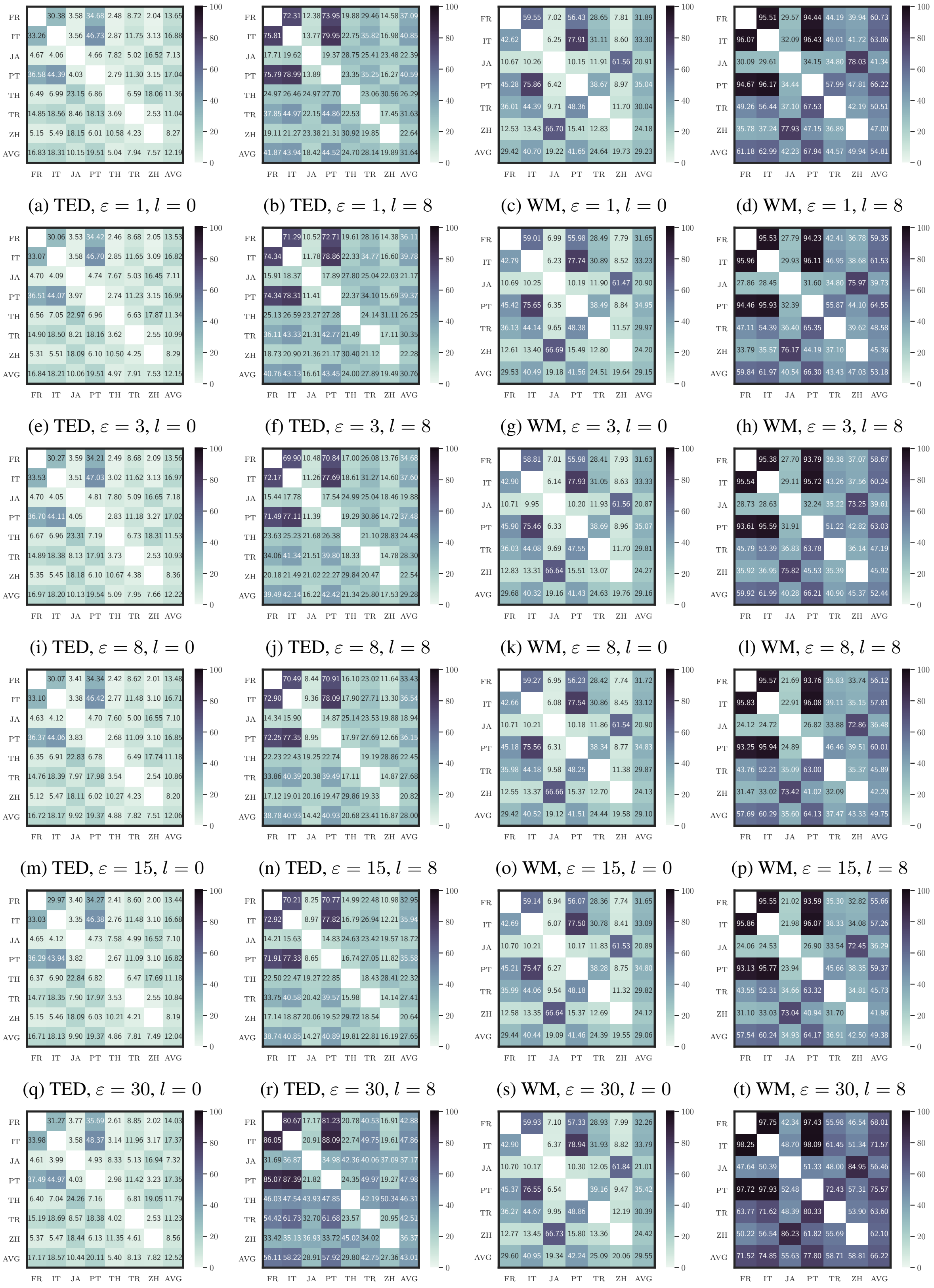
Table 3: Overview of the UD v2.8 (Nivre et al., 2020; Zeman et al., 2021) treebanks (test splits only) that we use as test sets in our POS tagging experiments (§3,4) including their respective sizes (number of sentences).

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 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	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**

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.



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

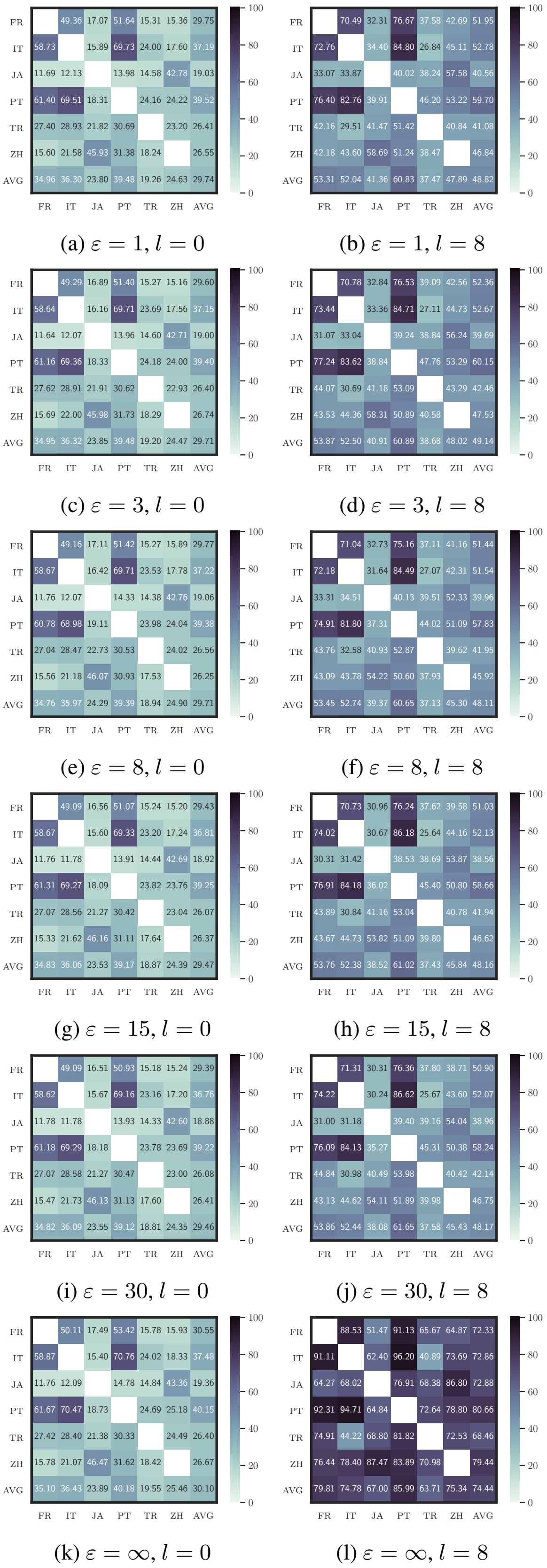
(i) TED, $\varepsilon = 8$, $l = 0$ (j) TED, $\varepsilon = 8$, $l = 8$ (k) WM, $\varepsilon = 8$, $l = 0$ (l) WM, $\varepsilon = 8$, $l = 8$

(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 5: **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 and particularly at layer 0.



(a) $\varepsilon = 1, l = 0$ (b) $\varepsilon = 1, l = 8$

(c) $\varepsilon = 3, l = 0$ (d) $\varepsilon = 3, l = 8$

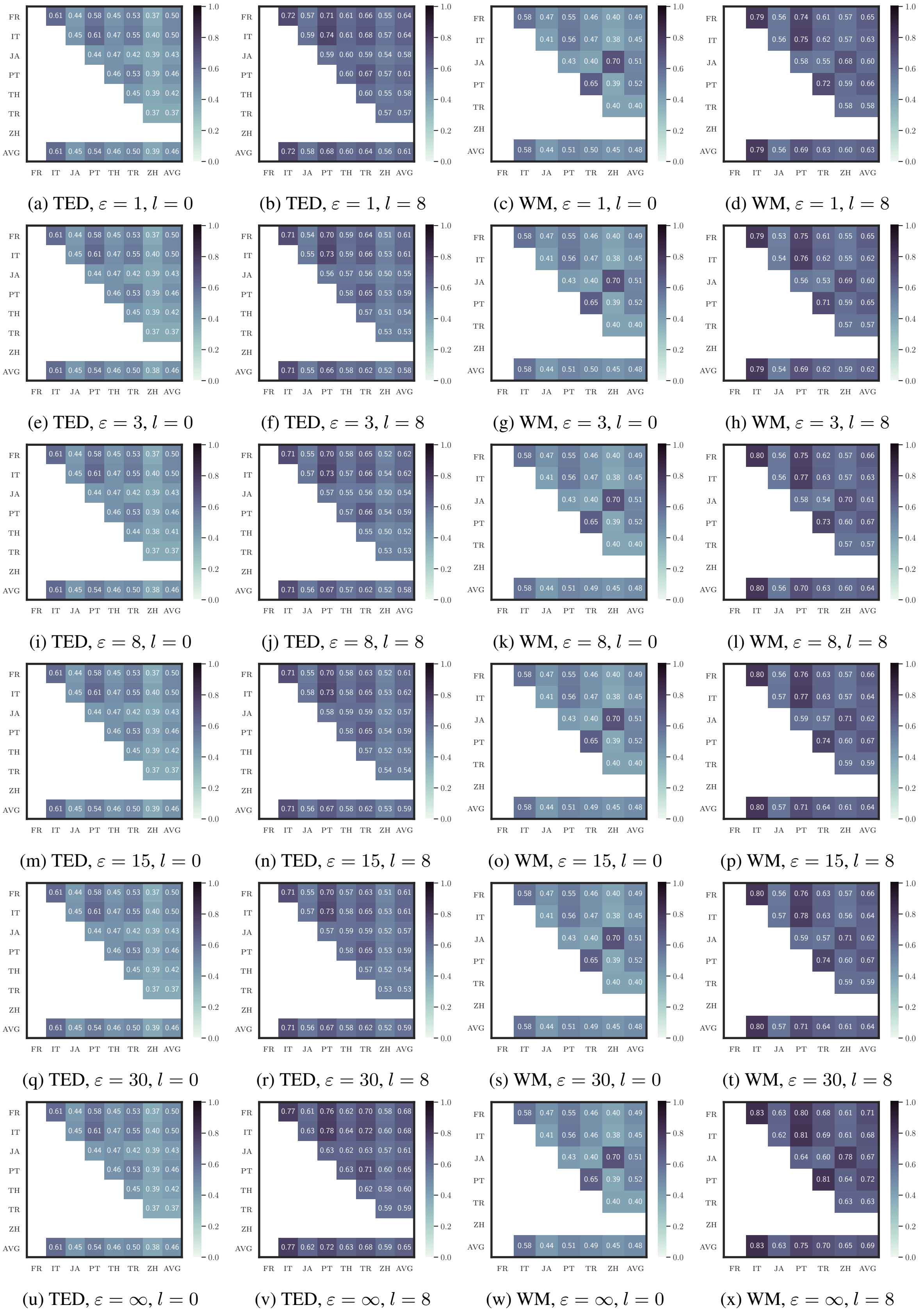
(e) $\varepsilon = 8, l = 0$ (f) $\varepsilon = 8, l = 8$

(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 6: **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 and 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, $\varepsilon = 3$, $l = 8$

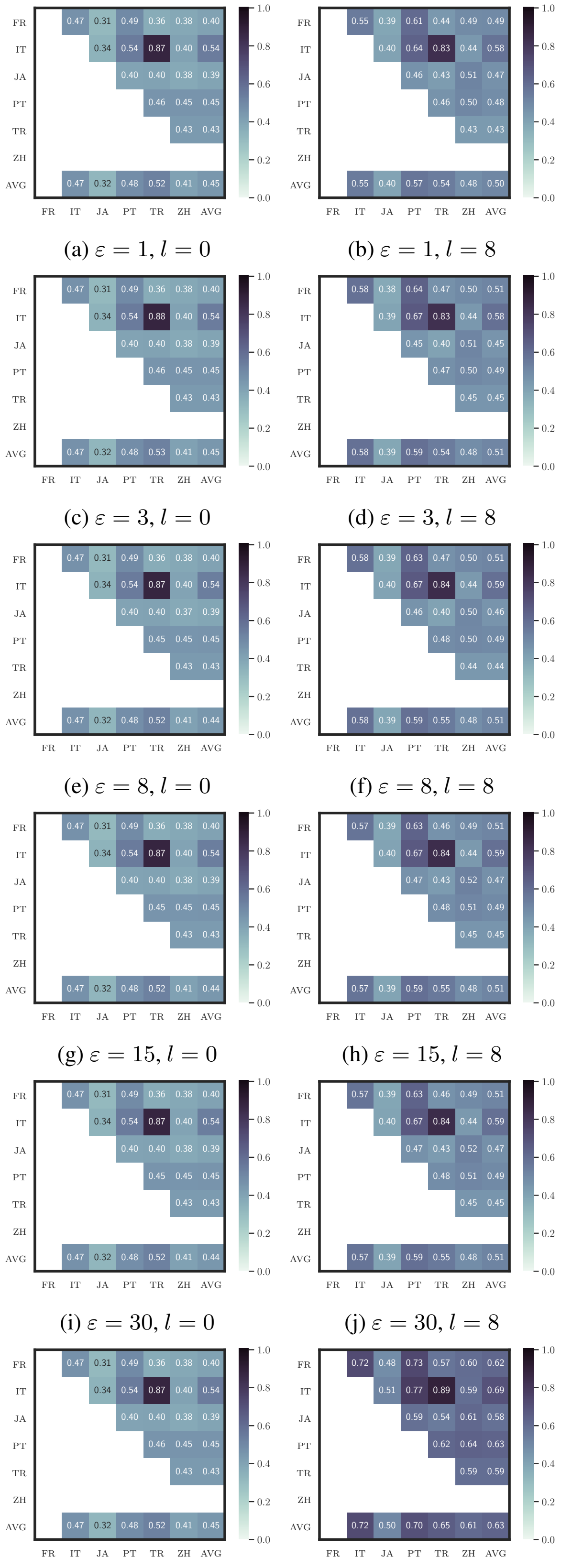
(i) TED, $\varepsilon = 8$, $l = 0$
(j) TED, $\varepsilon = 8$, $l = 8$
(k) WM, $\varepsilon = 8$, $l = 0$
(l) WM, $\varepsilon = 8$, $l = 8$

(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 7: **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 and 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) $\varepsilon = 15, l = 8$

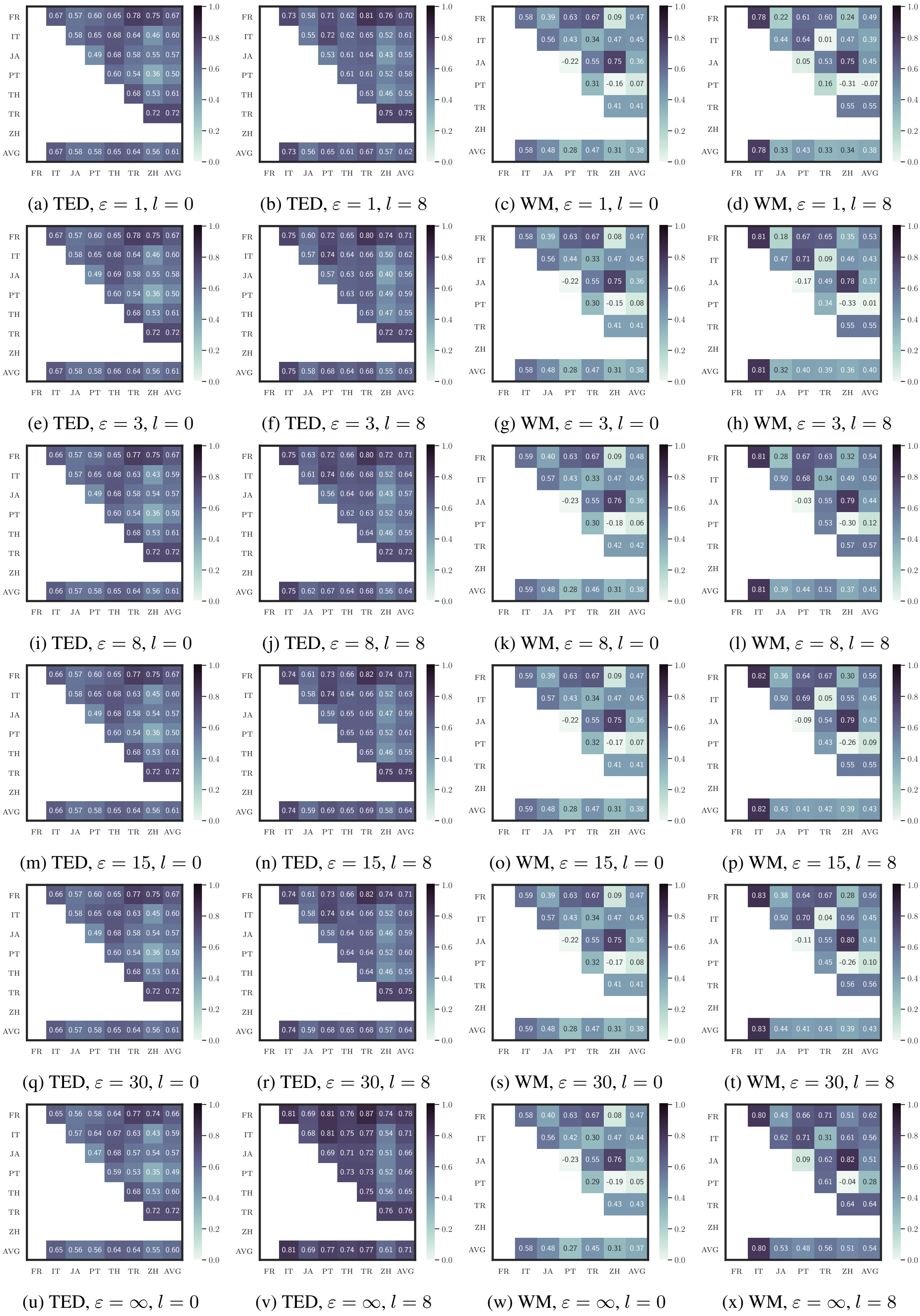
(i) $\varepsilon = 30, l = 0$

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

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

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

Figure 8: **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 and 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, $\varepsilon = 3, l = 8$

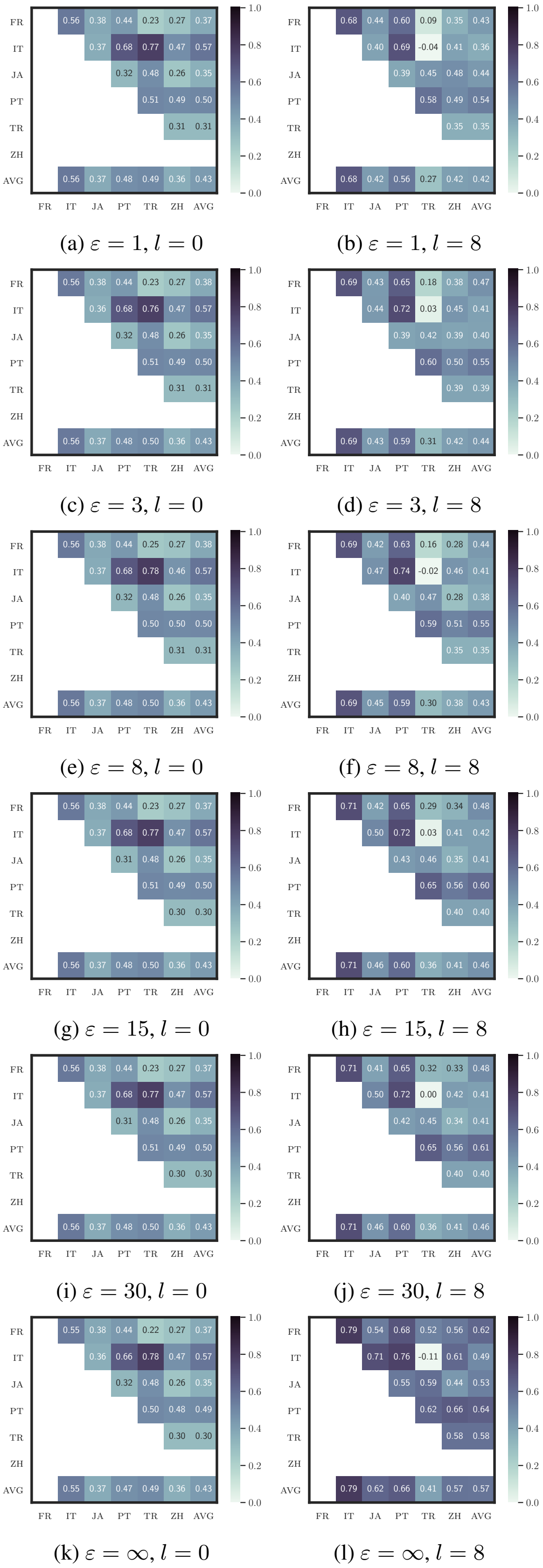
(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$

(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 9: **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 and 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) $\varepsilon = 15, l = 8$

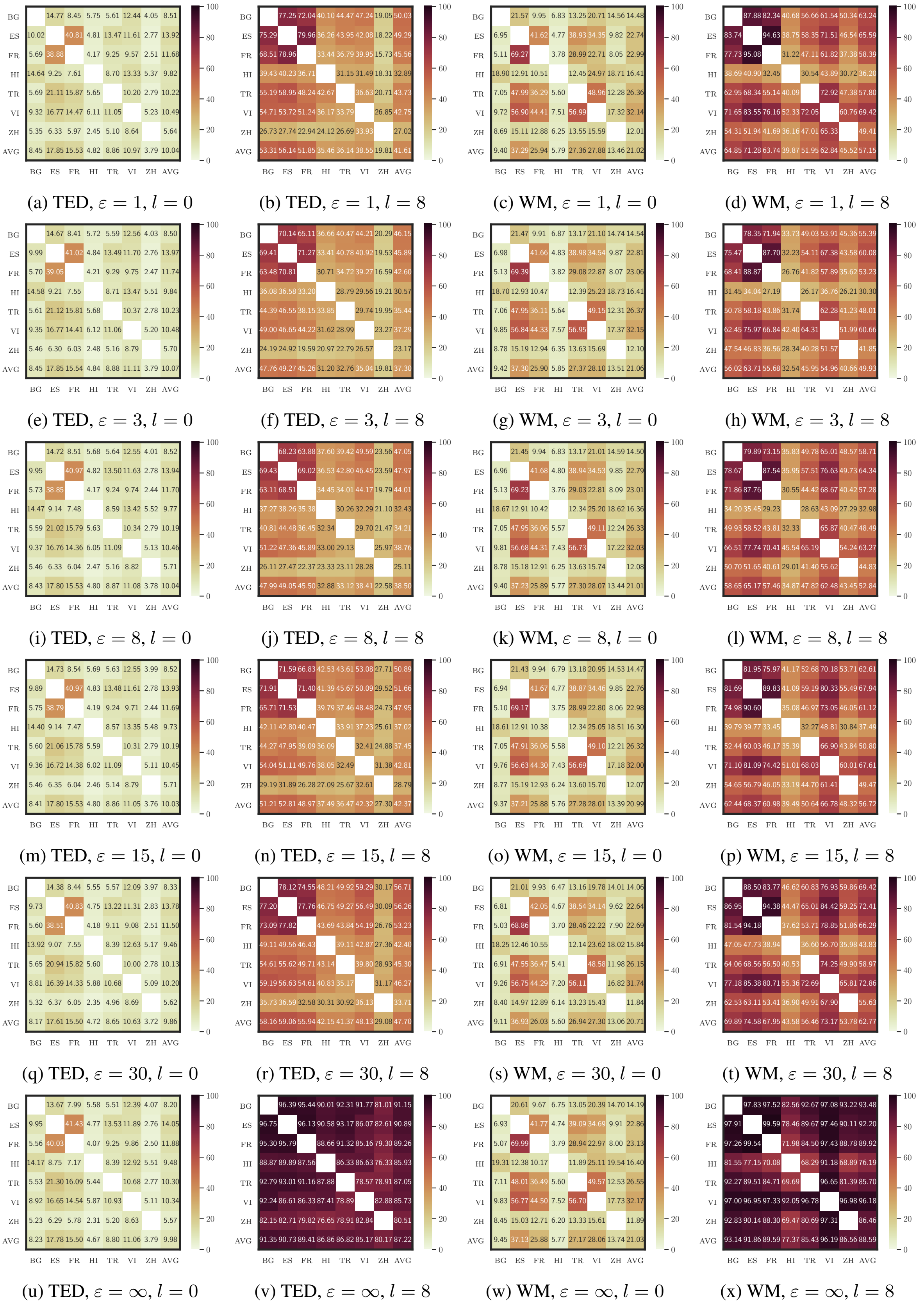
(i) $\varepsilon = 30, l = 0$

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

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

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

Figure 10: **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 and particularly at layer 0. Also note that, unlike in CKA (Figure 8), 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, $\varepsilon = 3, l = 8$

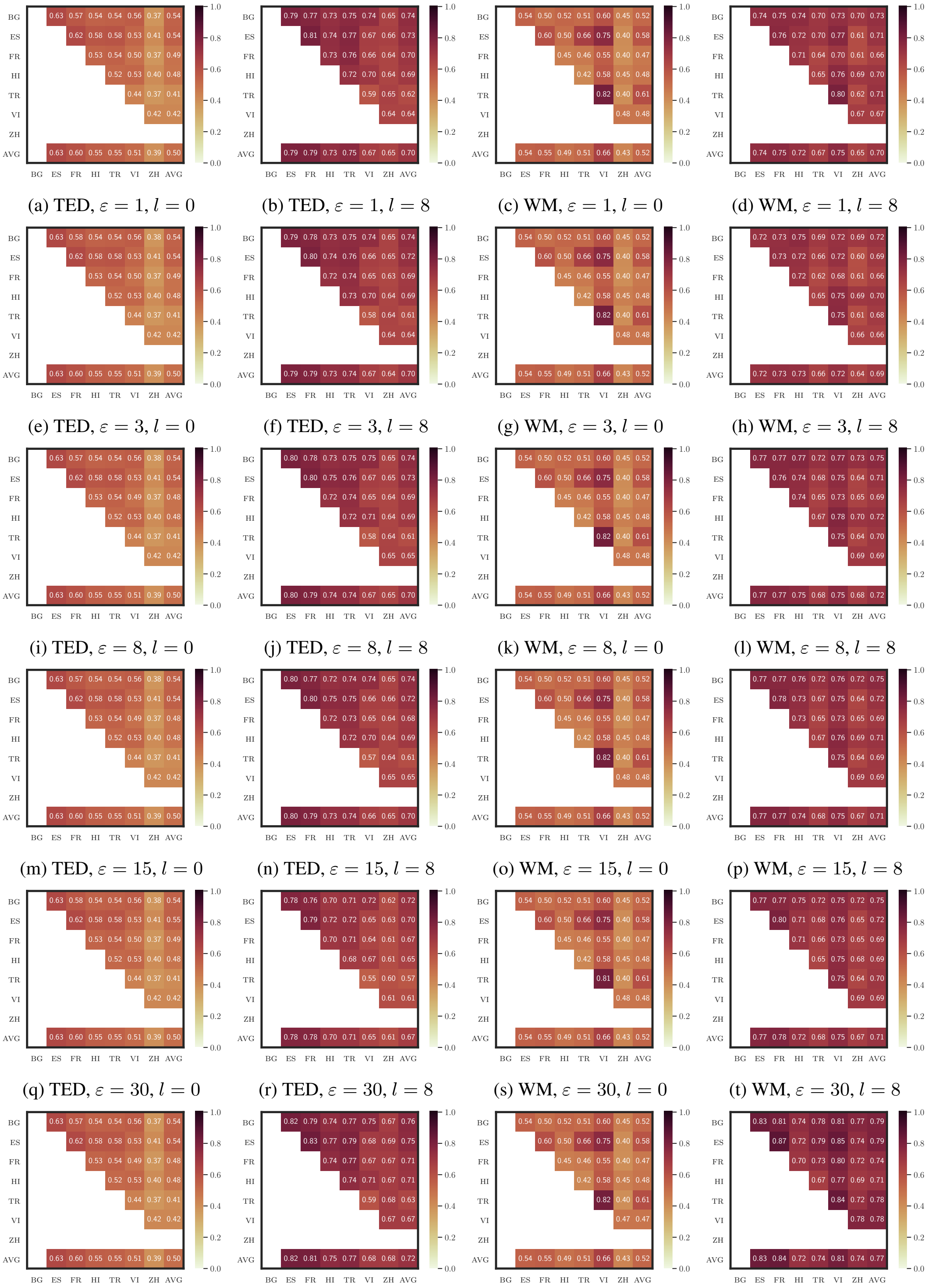
(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$

(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 11: **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 and 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, $\varepsilon = 3, l = 8$

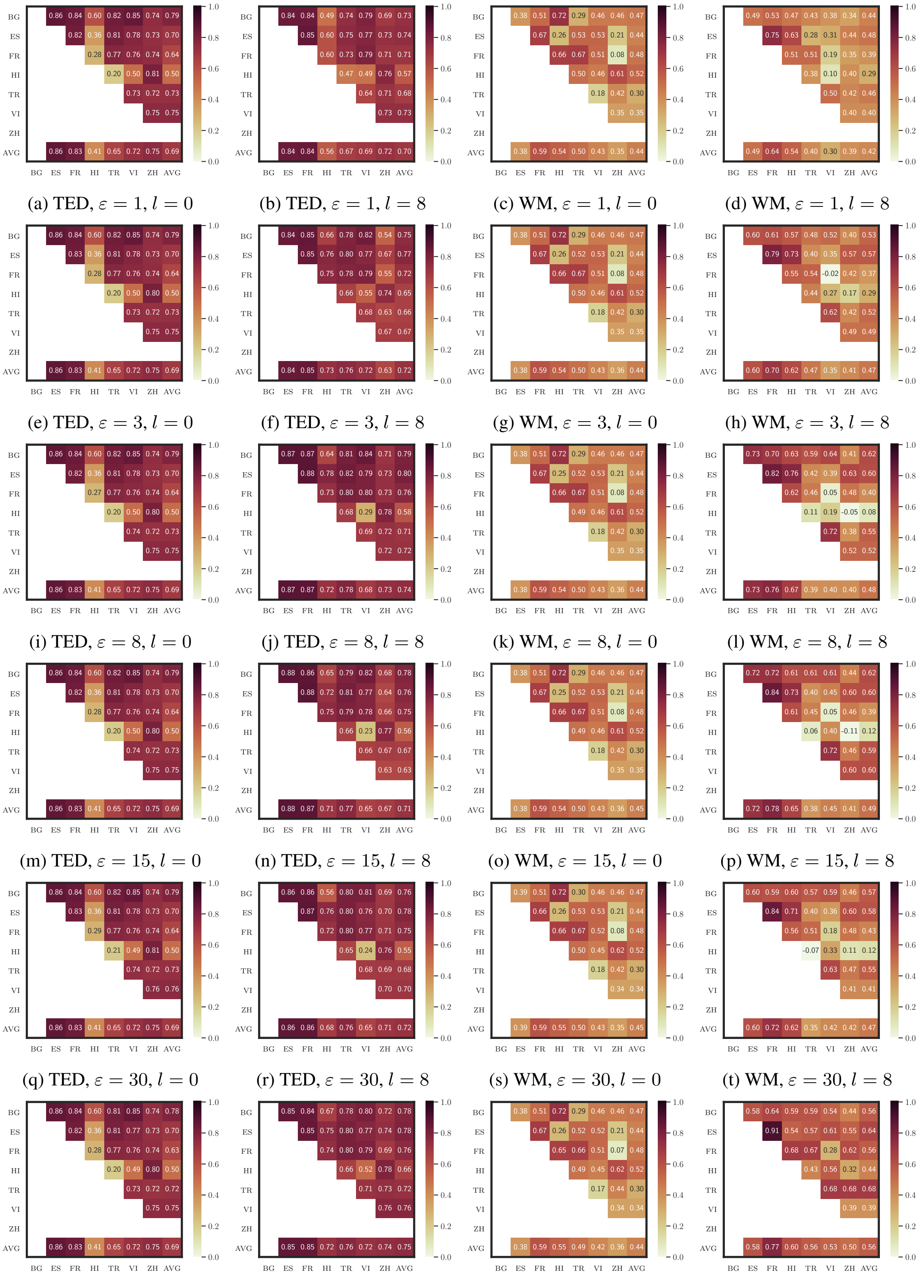
(i) TED, $\varepsilon = 8, l = 0$ (j) TED, $\varepsilon = 8, l = 8$ (k) WM, $\varepsilon = 8, l = 0$ (l) WM, $\varepsilon = 8, l = 8$

(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 12: **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 and 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** 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 and particularly at layer 0.

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 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	
	$l=0$	$l=8$	$l=0$	$l=8$
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

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.