

LAGO: Few-shot Crosslingual Embedding Inversion Attacks via Language Similarity-Aware Graph Optimization

Wenrui Yu¹, Yiyi Chen², Johannes Bjerva², Sokol Kosta¹, Qiongxiu Li^{1*}

¹Department of Electronic Systems, ²Department of Computer Science

Aalborg University, Copenhagen, Denmark

wenyu@es.aau.dk, {yiyic, jbjerva}@cs.aau.dk, {sok, qili}@es.aau.dk

Abstract

We propose **LAGO** - *Language Similarity-Aware Graph Optimization* - a novel approach for few-shot cross-lingual embedding inversion attacks, addressing critical privacy vulnerabilities in multilingual NLP systems. Unlike prior work in embedding inversion attacks that treat languages independently, LAGO explicitly models linguistic relationships through a graph-based constrained distributed optimization framework. By integrating syntactic and lexical similarity as edge constraints, our method enables collaborative parameter learning across related languages. Theoretically, we show this formulation generalizes prior approaches, such as ALGEN, which emerges as a special case when similarity constraints are relaxed. Our framework uniquely combines Frobenius-norm regularization with linear inequality or total variation constraints, ensuring robust alignment of cross-lingual embedding spaces even with extremely limited data (as few as 10 samples per language). Extensive experiments across multiple languages and embedding models demonstrate that LAGO substantially improves the transferability of attacks with 10-20% increase in Rouge-L score over baselines. This work establishes language similarity as a critical factor in inversion attack transferability, urging renewed focus on language-aware privacy-preserving multilingual embeddings.

1 Introduction

Text embeddings, which encode semantic and syntactic information into dense vector representations, serve as the backbone of modern natural language processing (NLP) systems. They are also powering the Large Language Models, whose impact stretches far beyond NLP and is steadily shaping everyday lives and business operations. However, their widespread deployment in cloud-based ser-

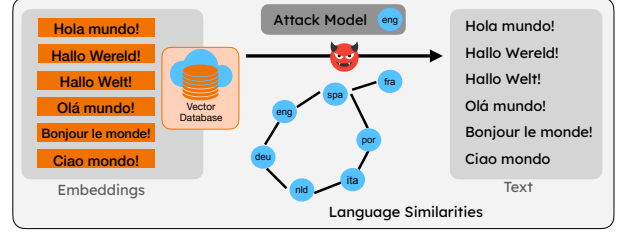


Figure 1: Few-shot Cross-lingual Textual Embedding Inversion Leveraging Language Similarities. Example: Attack model trained on English embeddings is used to attack embeddings in other languages, using language similarities as a prior.

vices introduces significant privacy risks. A particularly concerning threat is the embedding inversion attack (Song and Raghunathan, 2020; Chen et al., 2025b), where the adversaries can decode sensitive and private data directly from the embedding vectors. The security of the system can be compromised when malicious users abuse the embedding model API, collecting massive datasets to train attack models. Data leakage, whether accidental or deliberate, further exacerbates this vulnerability. As vector databases and generative-AI services proliferate across the globe, the embedding vectors offered as commodities are mostly multilingual. Yet, prior researches in this attack space mostly concentrate on inverting English embeddings (Song and Raghunathan, 2020; Li et al., 2023; Morris et al., 2023; Huang et al., 2024). While recent efforts (Chen et al., 2024a,b, 2025b) touch upon multilingual and cross-lingual inversion attacks, they *lack an explicit modeling of language similarities*, resulting in poor generalization across languages.

In real-world adversarial scenarios, such as specialized domains or low-resource languages, attackers may only have access to a handful of embedding-text pairs. Although ALGEN (Chen et al., 2025b) partially addresses the few-shot regime through decoder transfer, it lacks mech-

* Corresponding author.

anisms to exploit language similarity, which we hypothesize is a key factor in crosslingual generalization failure. Prior studies have shown that language similarities, simulated from typological features and lexical overlapping, correlate with structural variations in inversion outputs (Chen et al., 2024a, 2025a), providing empirical motivation for incorporating such relationships into attack models.

To address this, we propose **LAGO** (*Language Similarity-Aware Graph Optimization*) for few-shot crosslingual embedding inversion. LAGO explicitly models linguistic relationships by constructing a topological graph over languages, where nodes represent languages and edges encode similarity. As illustrated in Fig. 1, this graph is used to guide collaborative optimization of decoder alignment functions across languages, enabling knowledge transfer from typologically related neighbors. We formalize the attack objective as a distributed optimization problem, where each node corresponds to a language and constraints encode similarity-based consistency. We present two algorithmic variants: (1) an inequality-constrained approach based on IEQ-PDMM (Heusdens and Zhang, 2024b), and (2) a total variation regularized formulation (Peng et al., 2021) that softly penalizes parameter drift across similar languages. Our main contributions include:

- We propose LAGO, the first framework for few-shot crosslingual embedding inversion that incorporates language similarity as a structural prior in a graph-constrained optimization problem.
- We develop two algorithmic variants, one using inequality constraints and one using total variation penalties, that enable collaborative parameter learning across languages. Prior work, including ALGEN, emerges as a special case within our framework (cf. Section 4.3).
- Experiments across multiple embedding models and diverse languages show that language similarity strongly correlates with attack transferability, improving performance by 10 - 20% over prior methods.

By exposing overlooked vulnerabilities in multilingual embedding systems and demonstrating effective inversion under realistic low-resource conditions, our work underscores the urgent need for stronger privacy protections in cross-lingual NLP

deployments. While differential privacy offers some protection against our attack, it also significantly degrades downstream utility (Chen et al., 2025b), highlighting the need for more targeted and efficient defense mechanisms.

2 Related Work

2.1 Embedding Inversion Attacks

Early work on embedding inversion framed the task as classification over fixed vocabularies. For example, Song and Raghunathan (2020) aim to recover input tokens directly from embeddings, achieving up to 70% reconstruction. Subsequent advances recast the task as generation: Li et al. (2023) introduce a decoder-based approach to produce fluent text, while Morris et al. (2023) further improve accuracy through iterative refinement. Several work have since then extended inversion attacks to multilingual scenarios (Chen et al., 2024b,a). Moreover, Huang et al. (2024) trains a surrogate model to conduct transfer attack on victim embeddings under black-box access.

These methods, however, typically rely on massive training samples (8k to 5 million victim embeddings) and are primarily evaluated in monolingual or well-resourced settings. In practice, attackers often face few-shot scenarios. For example, reconstructing text in low-resource languages or specialized domains with only a handful of available samples. ALGEN (Chen et al., 2025b) introduces a linear alignment technique, allowing a decoder trained in one domain to be reused in another. While effective in few-shot transfer, ALGEN does not explicitly model language similarity or structural relationships between languages (cf. Section 3.2). Our framework improves ALGEN by directly incorporating linguistic knowledge to achieve stronger few-shot cross-lingual inversion.

2.2 Cross-lingual Transferability

Crosslingual transferability is an central research topic in multilingual NLP. Prior researches leverage crosslingual transferability to improve downstream task performances in target languages, mainly through fine-tuning LLMs on related source languages (Choenni et al., 2023), or using zero-shot transfer (Adelani et al., 2022; de Vries et al., 2022; Blaschke et al., 2025) or few-shot transfer with pre-trained MLLMs (Lauscher et al., 2020). Language similarity based on linguistic data, such as typological features (Littell et al., 2017) and lexi-

cal databases (Wichmann et al., 2022), have been used extensively in facilitating crosslingual transfer (Philippy et al., 2023). In this work, we leverage language similarity generated from both syntactic features and lexical overlap to provide alternative perspectives on constructing graphs, to assist crosslingual inversion attacks.

2.3 Distributed Optimization

Distributed optimization decomposes a global objective into smaller local problems that are solved collaboratively across networked nodes. Owing to its scalability and efficiency, it has become a foundational tool in large-scale machine learning and signal processing. Applications span domains such as federated learning (McMahan et al., 2017), sensor networks (Rabbat and Nowak, 2004), and privacy-preserving systems (Li et al., 2020; Yu et al., 2024). Classical distributed optimization algorithms include the Alternating Direction Method of Multipliers (ADMM, (Boyd, 2010)) and Primal-Dual Method of Multipliers (PDMM, (Zhang and Heusdens, 2017)) and their variants (Wang and Banerjee, 2014; Ouyang et al., 2015; Heusdens and Zhang, 2024a,b). To the best of our knowledge, their application to inversion attacks remains unexplored. In this work, we present, for the first time, a novel migration of distributed optimization techniques to inversion attacks.

3 Preliminaries

3.1 Embedding Inversion attack

Let $x \in \mathbb{V}^s$ denote a sequence of text tokens, and the text encoder $\phi = \text{enc}(\cdot) : \mathbb{V}^s \rightarrow \mathbb{R}^n$ be an embedding function that maps text x to a fixed-length vector $\phi(x) \in \mathbb{R}^n$. s is the sequence length and n the embedding dimension respectively. An embedding inversion attack is formally defined as the process of learning an approximate inverse function $g = \text{dec}(\cdot)$ such that:

$$g(\phi(x)) \approx x.$$

3.2 ALGEN

ALGEN enables cross-domain and cross-lingual sentence-level inversion through a framework combining embedding alignment and sequence generation. The framework consists of three parts:

1) Training a local attack model $\text{dec}_A(\cdot)$ by fine-tuning a pre-trained decoder to function as an embedding-to-text generator.

2) Embedding Alignment To bridge the discrepancy between the victim $e_V \in \mathbb{R}^m$ and the attack $e_A \in \mathbb{R}^n$ embedding spaces, a linear mapping matrix $W \in \mathbb{R}^{m \times n}$ is learned:

$$\hat{e}_A = e_V W.$$

The optimal alignment matrix W is obtained by solving the following least-squares minimization:

$$\min_W \|E_A - E_V W\|_F^2,$$

where $\|\cdot\|_F$ denotes the Frobenius norm, $E_V = [e_V^{1\top}, \dots, e_V^{b\top}]^\top \in \mathbb{R}^{b \times m}$ is the victim model’s embedding matrix, and $E_A = [e_A^{1\top}, \dots, e_A^{b\top}]^\top \in \mathbb{R}^{b \times n}$ is the attacker’s embedding matrix, and b is the number of training samples. m and n are the embedding dimensions of the victim and attack models, respectively. This optimization problem admits a closed-form solution via the normal equation (see the derivation in Appendix A):

$$W = (E_V^\top E_V)^{-1} E_V^\top E_A,$$

which minimizes the reconstruction error between the aligned victim embeddings $E_V W$ and the attacker’s reference embeddings E_A .

3) Text Reconstruction The aligned embeddings \hat{e}_A are decoded into text via dec_A , i.e.,

$$\hat{x} = \text{dec}_A(\hat{e}_A) = \text{dec}_A(e_V W).$$

ALGEN achieves inversion without retraining the victim model, requiring only fine-tuning of dec_A and estimation of W .

3.3 Fundamentals of Distributed Optimization

Distributed optimization addresses global optimization problems through a unified objective function while incorporating constraints derived from inter-node relationships within the network. Formally, this approach can be expressed as

$$\begin{aligned} \min_{\{w_i : i \in \mathcal{V}\}} \quad & \sum_{i \in \mathcal{V}} f_i(w_i), \\ \text{s.t.} \quad & h_{ij}(w_i, w_j) \leq 0, \quad (i, j) \in \mathcal{E}, \end{aligned}$$

where f_i denotes the local objectives on node i and h_{ij} encodes the constraints between adjacent nodes i and j .

Specifically, a fundamental formulation in distributed optimization employs linear inequality constraints to couple decision variables across network nodes. Such formulation can be efficiently solved

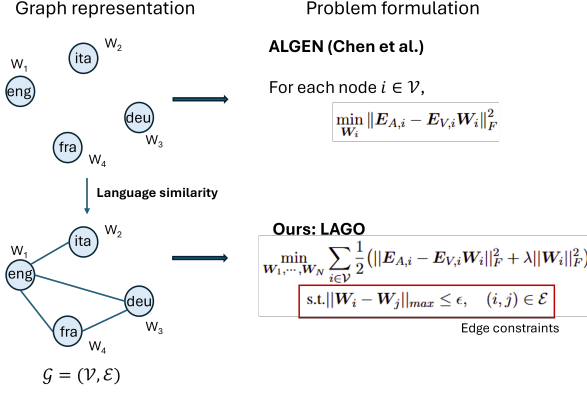


Figure 2: Illustration of LAGO vs. ALGEN (Chen et al., 2025b). **Top:** ALGEN treats each language independently. **Bottom:** LAGO leverages language similarity by introducing edge constraints in a joint distributed optimization framework.

using IEQ-PDMM method (Heusdens and Zhang, 2024b) and this can be formally expressed as

$$\begin{aligned} \min_{\{w_i: i \in \mathcal{V}\}} \quad & \sum_{i \in \mathcal{V}} f_i(w_i), \\ \text{s.t.} \quad & A_{i|j} w_i + A_{j|i} w_j \leq b_{i,j}, \quad (i, j) \in \mathcal{E}. \end{aligned} \quad (1)$$

Constraints between entries are defined by $A_{i|j}$, $A_{j|i}$ and $b_{i,j}$.

4 LAGO: Language Similarity-Aware Graph Optimization Framework

Building upon the ALGEN paradigm and grounded in distributed optimization, we propose **LAGO** - a general framework for few-shot cross-lingual embedding inversion. LAGO operates in two stages: (1) constructing a language similarity graph to capture topological relationships between languages, and (2) solving a graph-constrained optimization problem to jointly estimate transformation matrices across languages. This section details both components and introduces two algorithmic variants that implement our optimization framework.

4.1 Step I: Language Similarity-Aware Graph Construction

To formalize cross-lingual relationships, we propose to construct a linguistic topological graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the set of nodes \mathcal{V} represents languages and the set of undirected edges \mathcal{E} encodes pairwise similarity. Language similarity is quantified using established metrics such as AJSP (Wichmann et al., 2022) and Lang2vec (Littell et al., 2017). For a predefined threshold r , an edge

is established between two languages i and j if their distance metric $D_{ij} < r$. Mathematically, given a distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ over N languages, the adjacency matrix $\mathbf{A} \in \{0, 1\}^{N \times N}$ of the resulting topology is derived as:

$$\mathbf{A} = \frac{1 - \text{sign}(\mathbf{D} - r)}{2}.$$

See Appendix B for a concrete example of graph construction.

4.2 Step II: Graph-Constrained Optimization Algorithms

Using the constructed graph, we reformulate the optimization objective, leveraging cross-lingual relationships, thereby enhancing embedding inversion attacks through knowledge transfer from linguistically related languages. In few-shot settings where local data is scarce, this formulation improves transferability by leveraging cross-lingual regularities. We introduce two optimization strategies: one enforcing hard constraints and one applying soft penalties. Let \mathbf{W}_i denote the transformation matrix at node i (language i). To ensure stability in underdetermined settings (e.g., $b < m$), we incorporate Frobenius norm regularization to mitigate rank deficiency and enhance convergence.

Variant 1: Linear Inequality Constraints The first algorithm variant introduces topological constraints to enforce consistency between adjacent nodes' transformation matrices. Formally, we formulate the objective as minimizing the sum of reconstruction errors across all nodes while imposing ϵ -bounded constraints on the pairwise differences between neighbors' mapping matrices:

$$\begin{aligned} \min_{\mathbf{W}_1, \dots, \mathbf{W}_N} \quad & \sum_{i \in \mathcal{V}} \frac{1}{2} (\|\mathbf{E}_{A,i} - \mathbf{E}_{V,i} \mathbf{W}_i\|_F^2 + \lambda \|\mathbf{W}_i\|_F^2), \\ \text{s.t.} \quad & \|\mathbf{W}_i - \mathbf{W}_j\|_{\max} \leq \epsilon, \quad (i, j) \in \mathcal{E}, \end{aligned}$$

where $\|\cdot\|_{\max}$ denotes the entry-wise ℓ_∞ norm.

This formulation corresponds to the general inequality-constrained form in Eq. (1), where $A_{i|j} = -A_{j|i} = [1 \ -1]^T$ and $b_{i,j} = [\epsilon \ \epsilon]^T$. As such, it is compatible with the IEQ-PDMM optimization framework (Heusdens and Zhang, 2024b). The update equations used in this framework are

given below¹.

$$\begin{aligned}
\mathbf{W}_i^{(t)} &= [\mathbf{E}_{V,i}^\top \mathbf{E}_{V,i} + (2cd_i + \lambda)\mathbf{I}]^{-1} \\
&\quad (\mathbf{E}_{V,i}^\top \mathbf{E}_{A,i} - \sum_{j \in \mathcal{N}_i} A_{i|j} \mathbf{Z}_{i|j}^{(t)}) \\
\mathbf{Y}_{i|j}^{(t)} &= \mathbf{Z}_{i|j}^{(t)} + 2cA_{i|j} \mathbf{W}_i^{(t)} - cb_{i,j} \\
\mathbf{Z}_{i|j}^{(t+1)} &= \begin{cases} \mathbf{Y}_{j|i}^{(t+1)}, & \mathbf{Y}_{i|j}^{(t+1)} + \mathbf{Y}_{j|i}^{(t+1)} > 0, \\ -\mathbf{Y}_{i|j}^{(t+1)}, & \text{otherwise,} \end{cases}
\end{aligned} \tag{2}$$

where $d_i = |\mathcal{N}_i|$ is the degree of node i .

Variant 2: Total Variation Regularization The second variant introduces soft penalties using total variation across edges, a technique originally proposed for Byzantine-robust decentralized learning systems (Peng et al., 2021). The optimization objective is formulated as follows:

$$\begin{aligned}
\min_{\mathbf{W}_1, \dots, \mathbf{W}_N} \sum_{i \in \mathcal{V}} & \left(\frac{1}{2} \|\mathbf{E}_{A,i} - \mathbf{E}_{V,i} \mathbf{W}_i\|_F^2 + \frac{\lambda}{2} \|\mathbf{W}_i\|_F^2 \right. \\
& \left. + \eta \sum_{j \in \mathcal{N}_i} \|\mathbf{W}_i - \mathbf{W}_j\|_{sum} \right),
\end{aligned}$$

where $\|\cdot\|_{sum}$ denotes the entry-wise ℓ_1 norm. At time t , each node updates its \mathbf{W} with $\mathbf{W}_i^{(t+1)} = \mathbf{W}_i^{(t)} - \frac{\alpha}{\sqrt{t+1}} \left[-\mathbf{E}_{V,i}^\top (\mathbf{E}_{A,i} - \mathbf{E}_{V,i} \mathbf{W}_i^{(t)}) + \lambda \mathbf{W}_i^{(t)} + \eta \sum_{j \in \mathcal{N}_i} \text{sign}(\mathbf{W}_i^{(t)} - \mathbf{W}_j^{(t)}) \right]$,

where α is the learning rate.

4.3 Generalization: ALGEN as a Special Case

Our proposed LAGO is general and subsumes prior method ALGEN as a special case. Specifically, in the inequality-constrained variant, when $\epsilon \rightarrow \infty$, cross-node constraints vanish, and each language node solves an independent alignment problem. Similarly, in the total variation setting, setting $\eta = 0$ decouples all nodes. In both cases, the optimization reduces to ALGEN’s per-language formulation with no cross-lingual structure. This highlights the flexibility of LAGO: by adjusting constraint strength, it interpolates between isolated optimization (as in ALGEN) and fully collaborative cross-lingual inversion. Our approach thus provides a principled, generalizable framework for multilingual attack design.

¹The comparison in Eq. (2) is applied element-wise.

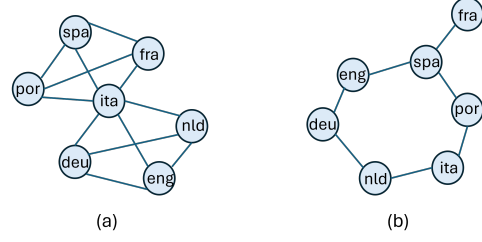


Figure 3: Example graphs using two Language Similarities: (a) AJSP model with $r = 0.9$; (b) Lang2vec model with $r = 0.45$.

5 Experimental Setup

Models and Dataset Our attack framework is initialized using a pre-trained FLAN-T5 model. To evaluate the robustness of our approach, we conduct experiments with two distinct victim model encoders, MT5, E5-SMALL-V2 (E5) and OpenAI’s TEXT-EMBEDDING-ADA-002 (ADA-2) (see the details in Tabel 3). The decoder $dec_A(\cdot)$, fine-tuned on the MMARCO English dataset (Bonifacio et al., 2021), serves in this paper as *the attack model* for simulating few-shot inversion attack scenarios. We employ the current state-of-the-art ALGEN method as the baseline for the few-shot scenario, maintaining identical training and testing configurations for the decoder as those used in ALGEN. To assess cross-lingual transferability, we select a subset of seven syntactically and lexically related languages: English, German, French, Dutch, Spanish, Italian and Portuguese.

Language Graphs We evaluate two distinct topologies derived from language similarities: AJSP and Lang2vec. The tested topologies are illustrated in Fig. 3.

Regularization Parameters To accomplish substantial convergence, the number of iterations is fixed at 500. For the linear inequality constraints method, the convergence parameter is set to $c = 0.4$, while for the TV penalty term method, the learning rate is chosen as $\alpha = 0.01$. The computational cost of the attack is relatively low, using the topology of 7 languages as an example, it takes approximately five minutes to compute a set of matrix $\{\mathbf{W}_i : i \in \mathcal{V}\}$ with the inequality constrained formulation, while the total variation method is faster, completing the attack in about two minutes.

Evaluation Metrics We use Cosine similarity to measure the semantic alignment between the adversarial embeddings of the victim model $\mathbf{E}_V \mathbf{W}$

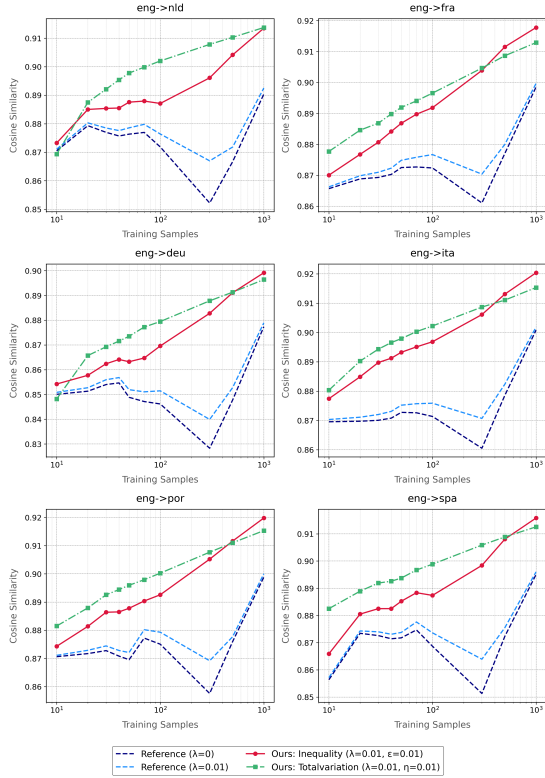


Figure 4: Cross-lingual Inversion Performances with AJSP Graph in Cosine Similarities across Training Samples.

and the target attack embeddings E_A . Meanwhile, Rouge-L (Lin, 2004) evaluates the lexical overlap between the reconstructed text and the ground truth by computing the length of their longest common subsequence, serving as a proxy for assessing the fidelity of the generated output at the lexical level.

6 Analysis and Results

To validate the effectiveness of our proposed LAGO framework, we experiment across a range of settings and tasks. Each subsection addresses one research question, probing key aspects of cross-lingual transferability, generalization, and robustness to defense mechanisms².

6.1 Do Similar Languages Transfer Vulnerabilities?

To assess whether language similarity aids attack transfer, we use an attack model trained on English data attack embeddings in other languages. We compare LAGO (with both optimization variants) to ALGEN baselines with and without Frobenius norm regularization ($\lambda = 0.01$), using 10 to 1000

²We open-source our code https://anonymous.4open.science/r/ALGO_anonymous.

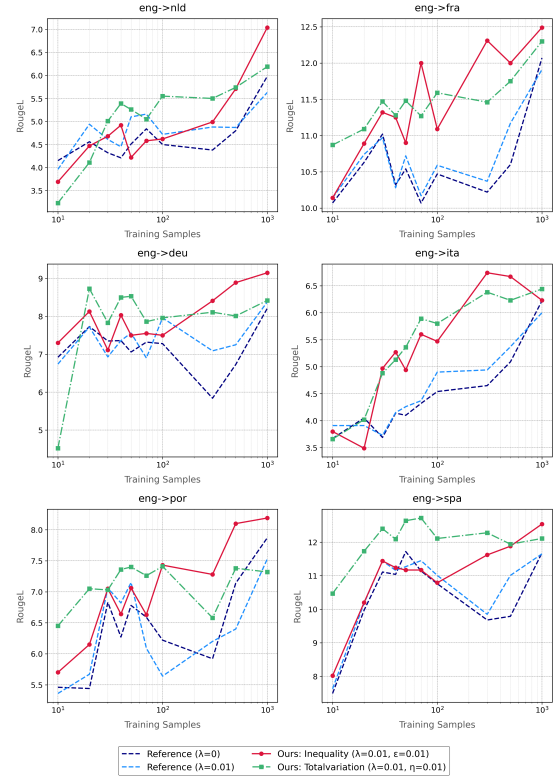


Figure 5: Cross-lingual Inversion Performances with AJSP Graph in Rouge-L Scores across Training Samples.

training samples. Notice that the training sample is used exclusively for alignment. For LAGO, we set $\epsilon = 0.01$ and $\eta = 0.01$. As shown in Table 1, LAGO consistently improves both cosine similarity and Rouge-L scores on inverting French embeddings across all training sizes. In low-resource settings (e.g., 10 samples), our method yields a 10–20% boost in Rouge-L over ALGEN. This trend generalizes to other languages, such as Dutch, German, Italian, Portuguese and Spanish, as demonstrated in Fig. 4; 5 and Fig. 6. These findings suggest that leveraging language similarity both mitigates data scarcity and optimizes cross-lingual generalization in low-resource settings.

6.2 Does the choice of Language Similarity Metric Impact the Attack Effectiveness?

To test the sensitivity of LAGO to the choice of language similarity measures, we compare performance under two topologies: ASJP (lexical similarity) and Lang2vec (syntactic similarity). The results, demonstrated in Fig. 6 and 9 in Appendix C, confirm that LAGO is robust to the choice of similarity metric. Lang2vec shows slightly better performance in moderate-data settings in terms

Method		Cosine Similarities				Rouge-L			
Train Samples		10	100	300	1000	10	100	300	1000
ALGEN	-	0.8657	0.8723	0.8610	0.8986	10.07	10.47	10.22	12.07
	Reg. ($\lambda = 0.01$)	0.8663	0.8767	0.8703	0.8997	10.14	10.59	10.37	11.91
Ours	Inequality	0.8701	0.8919	0.9039	<u>0.9178</u>	10.14	11.09	12.31	12.49
	Total Variation	<u>0.8777</u>	<u>0.8966</u>	<u>0.9046</u>	0.9129	10.87	11.59	11.46	12.30

Table 1: Cross-lingual Inversion Performances of French embeddings with Attack Model trained in English in Cosine Similarities and Rouge-L scores across Training Samples. The best Rouge-L scores are **bold**, and the maximum cosine similarities are underlined.

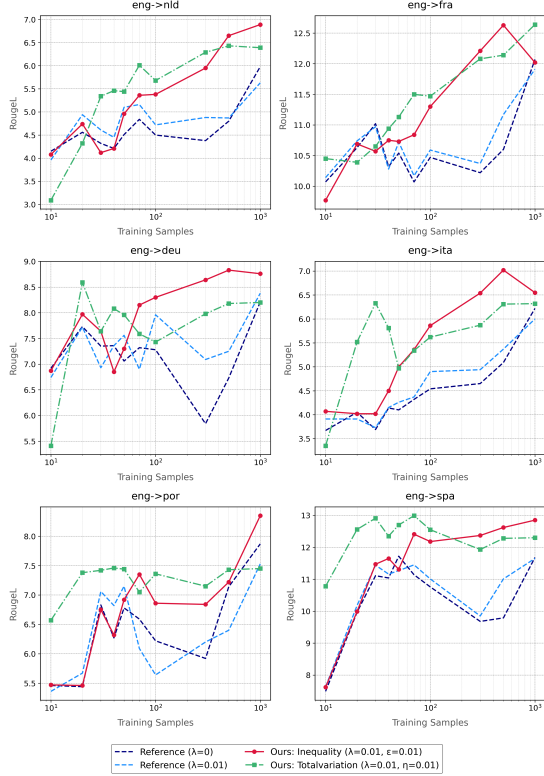


Figure 6: Cross-lingual Inversion Performances with Lang2vec Graph in Rouge-L Scores across Training Samples.

of Rouge-L scores for moderately larger training sample sizes (>300). For instance, Dutch, with training samples of $|D_V| = 500$, exhibits an increase from 5.71 to 6.65. Overall, our approach consistently outperforms the baseline in terms of attack efficacy, irrespective of the similarity metric. This suggests that LAGO is not contingent upon a specific language similarity framework but instead exhibits robust generalizability across diverse language structures. Furthermore, the observed improvements in attack effectiveness indicate that our methodology is particularly advantageous for languages with shared linguistic features. Whether the similarity is lexical or syntactic, the attack remains effective, reinforcing its versatility.

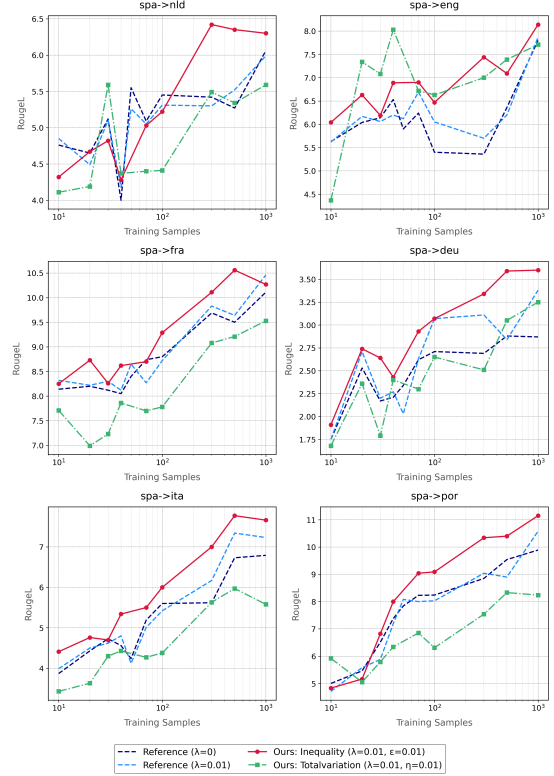


Figure 7: Cross-lingual Inversion Performances with Attack Model trained in Spanish in Rouge-L Scores across Training Samples.

6.3 Is the Inversion Generalizable to Different Victim Models?

We assess generalizability by evaluating our method on embeddings from ADA-2 and E5 encoders. As shown in Appendix Fig. 10 - 13, LAGO consistently outperforms ALGEN in both cosine similarity and Rouge-L across these models.

Relatively, under the Rouge-L metric, the inequality constraint demonstrates stronger performance with larger sample sizes, whereas total variation proves more effective in extremely few-shot scenarios with fewer than 300 training samples. We attribute this to the flexibility of inequality constraints, a smaller sample size provides \mathbf{W} with

greater degrees of freedom, thereby imposing relatively weaker restrictions on \mathbf{W} under the same ϵ . Consequently, the performance of inequality constraints under smaller sample sizes aligns more closely with the ALGEN method.

6.4 Can other Languages assume the Source of Transfer?

English as the most represented language in the pretrained LLMs, serves as an obvious choice for training the attack model to facilitate the inversion of other languages. We demonstrate that the proposed scheme remains robust even when the attack model is trained in an alternative language. As shown in Fig. 7 and Fig. 14 in Appendix C, when Spanish is used as the attack language, LAGO continues to yield consistent improvements over the baseline. The cosine similarity increases across target languages, and the inequality-constrained variant shows stronger gains in Rouge-L, particularly under low-resource conditions.

We also observe performance disparities across specific language pairs. For example, the inversion performance from English to German is notably higher than that from Spanish to German - a pattern already present in the ALGEN baseline. This disparity may be attributed to two factors: differences in decoder training quality and variations in language similarity. In our constructed graph, English and German are directly connected (one-hop neighbors), whereas Spanish and German are two hops apart. The increased topological distance may weaken the effectiveness of parameter transfer, as similarity constraints exert less influence.

These observations suggest that the relative position of languages in the similarity graph - and not just data size or encoder choice - can influence transfer strength. Understanding the dynamics of language topology in transfer-based attacks presents an important direction for future work.

6.5 Defenses

We further investigate the effectiveness of differential privacy (DP) in mitigating embedding inversions. We employ the SNLI dataset (Bowman et al., 2015) to fine-tune the decoder and subsequently transfer the adversarial attack framework to German, French and Spanish using the XNLI dataset (Conneau et al., 2018). While the SNLI dataset is widely utilized for downstream tasks like text classification, Chen et al. (2025b) has demonstrated that with a strong privacy guarantee $\epsilon_{dp} = 1$,

model accuracy drops to 40%, which is a significant reduction from the 60% accuracy achieved at $\epsilon_{dp} = 12$ where DP defenses show limited impact on utility and inversion performance.

In our setup, we apply two DP mechanisms: the Purkayastha Mechanism (PurMech) and the Normalized Planar Laplace Mechanism (LapMech) proposed by Du et al. (2023) in sentence embeddings. The privacy budget parameter is evaluated across $\epsilon_{dp} \in [1, 4, 8, 12]$. As shown in Table 2, Table 4, and Fig. 15, 16 in Appendix C, inversion attacks in cross-lingual settings are highly sensitive to DP perturbations. Specifically, Rouge-L scores are consistently suppressed to below 2 across tested configurations. These results are consistent with theoretical expectations: more challenging examples, such as those in cross-lingual or low-resource settings, tend to be more sensitive to DP noise (Carlini et al., 2019; Feldman, 2020). While DP mechanisms provide meaningful protection against inversion, they incur a non-trivial utility cost, underscoring the need for more efficient, structure-aware defenses in multilingual NLP applications.

Lang	ϵ_{dp}	Rouge-L↓		COS↓	
		LapMech		PurMech	
eng → eng	1	14.11	0.0017	14.05	0.0156
	4	13.58	0.0087	13.94	0.0348
	8	13.38	0.0249	13.45	0.0185
	12	13.90	0.0345	12.77	-0.0076
eng → fra	1	1.66	-0.0013	1.31	0.0136
	4	1.70	-0.0043	1.58	0.0140
	8	1.42	0.0364	1.24	0.0166
	12	1.60	0.0411	1.44	0.0113
eng → deu	1	0.52	-0.0119	0.49	0.0090
	4	0.32	0.0065	0.54	0.0127
	8	0.62	0.0187	0.53	0.0418
	12	0.44	0.0327	0.43	0.0367
eng → spa	1	1.47	0.0062	1.55	-0.0090
	4	1.43	-0.0006	1.32	0.0208
	8	1.70	0.0384	1.35	0.0266
	12	1.52	0.0160	1.41	0.0389

Table 2: Cross-lingual Inversion Performance with $|D_V|=100$ on Classification Tasks on SNLI dataset with Local DP (Inequality). From a defender’s perspective, ↓ means lower is better.

7 Conclusion

We proposed two optimization-based paradigms for enhancing few-shot crosslingual embedding inversions. Both are grounded in distributed optimization and operate over a topological network of languages constructed via language similarity. This graph structure enables collaborative alignment of embedding decoders, facilitating effective

knowledge transfer even with extremely limited supervision. Our experimental results show that both variants - linear inequality constraints and total variation penalties - consistently outperform existing methods, including ALGEN. In particular, the total variation approach demonstrates superior robustness in extremely few-shot settings, validating the importance of smooth cross-lingual parameter sharing. These findings establish language similarity as a key enabler of transferable inversion attacks, and underscore the need for privacy-preserving defenses that account for structural relationships among languages in multilingual NLP systems.

Limitations

While our approach outperforms prior methods, few-shot crosslingual embedding inversion remains a challenging task with substantial room for improvement. One limiting factor appears to be the decoder itself: even in the monolingual (original language) setting, inversion accuracy remains moderate, achieving approximately a 25 Rouge-L score on the MMARCO English dataset with $|D_V|=1k$, and further declines under cross-lingual transfer. This suggests that the current attack decoder may struggle to generalize across languages, particularly when signal supervision is limited.

Interestingly, we observe that cross-lingual settings exhibit higher sensitivity to DP defenses, though such defenses incur significant utility degradation. This sensitivity highlights both the vulnerability and fragility of multilingual embeddings. Future work could focus on enhancing the decoder training, e.g., through multilingual pretraining, or incorporating language-specific priors - which we expect could improve inversion performance in both monolingual and crosslingual scenarios.

Computational Resources

We fine-tune the decoder on a single NVIDIA A40 GPU, with training completing in just three hours. Notably, ALGO operates with minimal GPU resource demands, enabling a true few-shot setup.

Ethics Statement

We comply with the ACL Ethics Policy. The inversion attacks implemented in this paper can be misused and potentially harmful to proprietary embeddings. We discuss and experiment with potential mitigation and defense mechanisms, and we encourage further research in developing effective defenses in this attack space.

Acknowledgements

WY is funded by the EU ChipsJU and the Innovation Fund Denmark through the project CLEVER (no. 101097560); YC and JB are funded by the Carlsberg Foundation, under the Semper Ardens: Accelerate programme (project nr. CF21-0454). We further acknowledge the support of the AAU AI Cloud for providing computing resources.

References

- David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba O. Alabi, Shamsuddeen H. Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, and 26 others. 2022. [MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Verena Blaschke, Masha Fedzechkina, and Maartje ter Hoeve. 2025. Analyzing the effect of linguistic similarity on cross-lingual transfer: Tasks and experimental setups matter. *arXiv preprint arXiv:2501.14491*.
- Luiz Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of the ms marco passage ranking dataset. *arXiv preprint arXiv:2108.13897*.
- Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.
- Stephen Boyd. 2010. [Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers](#). *Foundations and Trends® in Machine Learning*, 3(1):1–122.
- Nicholas Carlini, Ulfar Erlingsson, and Nicolas Papernot. 2019. Prototypical examples in deep learning: Metrics, characteristics, and utility.
- Yiyi Chen, Russa Biswas, Heather Lent, and Johannes Bjerva. 2024a. Against all odds: Overcoming typology, script, and language confusion in multilingual embedding inversion attacks. *arXiv preprint arXiv:2408.11749*.
- Yiyi Chen, Heather Lent, and Johannes Bjerva. 2024b. Text embedding inversion security for multilingual language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7808–7827.
- Yiyi Chen, Qiongxiu Li, Russa Biswas, and Johannes Bjerva. 2025a. [Large language models are easily confused: A quantitative metric, security implications and typological analysis](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3810–3827, Albuquerque, New Mexico. Association for Computational Linguistics.
- Yiyi Chen, Qionгкаi Xu, and Johannes Bjerva. 2025b. Algen: Few-shot inversion attacks on textual embeddings using alignment and generation. *arXiv preprint arXiv:2502.11308*.
- Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? studying cross-lingual data sharing during lm fine-tuning. *arXiv preprint arXiv:2305.13286*.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2022. [Scaling instruction-finetuned language models](#). *Preprint*, arXiv:2210.11416.
- Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [Xnli: Evaluating cross-lingual sentence representations](#). *Preprint*, arXiv:1809.05053.
- Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. [Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.
- Minxin Du, Xiang Yue, Sherman SM Chow, and Huan Sun. 2023. Sanitizing sentence embeddings (and labels) for local differential privacy. In *Proceedings of the ACM Web Conference 2023*, pages 2349–2359.
- Vitaly Feldman. 2020. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959.
- Richard Heusdens and Guoqiang Zhang. 2024a. Distributed nonlinear conic optimisation with partially separable structure. *arXiv preprint arXiv:2405.09490*.
- Richard Heusdens and Guoqiang Zhang. 2024b. Distributed optimisation with linear equality and inequality constraints using pdmm. *IEEE Transactions on Signal and Information Processing over Networks*.
- Yu-Hsiang Huang, Yuche Tsai, Hsiang Hsiao, Hong-Yi Lin, and Shou-De Lin. 2024. [Transferable Embedding Inversion Attack: Uncovering Privacy Risks in Text Embeddings without Model Queries](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4193–4205, Bangkok, Thailand. Association for Computational Linguistics.
- Anne Lauscher, Vinit Ravishankar, Ivan Vulic, and Goran Glavas. 2020. [From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Haoran Li, Mingshi Xu, and Yangqiu Song. 2023. [Sentence embedding leaks more information than you expect: Generative embedding inversion attack to](#)

- recover the whole sentence. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 14022–14040, Toronto, Canada. Association for Computational Linguistics.
- Qiongxiu Li, Richard Heusdens, and Mads Græsbøll Christensen. 2020. Privacy-preserving distributed optimization via subspace perturbation: A general framework. *IEEE Transactions on Signal Processing*, 68:5983–5996.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.
- John X Morris, Volodymyr Kuleshov, Vitaly Shmatikov, and Alexander M Rush. 2023. Text embeddings reveal (almost) as much as text. *arXiv preprint arXiv:2310.06816*.
- Yuyuan Ouyang, Yunmei Chen, Guanghui Lan, and Eduardo Pasiliao Jr. 2015. An accelerated linearized alternating direction method of multipliers. *SIAM Journal on Imaging Sciences*, 8(1):644–681.
- Jie Peng, Weiyu Li, and Qing Ling. 2021. Byzantine-robust decentralized stochastic optimization over static and time-varying networks. *Signal Processing*, 183:108020.
- Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. *arXiv preprint arXiv:2305.16768*.
- Michael Rabbat and Robert Nowak. 2004. Distributed optimization in sensor networks. In *Proceedings of the 3rd international symposium on Information processing in sensor networks*, pages 20–27.
- Congzheng Song and Ananth Raghunathan. 2020. [Information leakage in embedding models](#). In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, CCS ’20*, page 377–390, New York, NY, USA. Association for Computing Machinery.
- Huahua Wang and Arindam Banerjee. 2014. Bregman alternating direction method of multipliers. *Advances in neural information processing systems*, 27.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2022. [CLDF dataset derived from Wichmann et al.’s "ASJP Database" v20 from 2022](#).
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.
- Wenrui Yu, Qiongxiu Li, Milan Lopuhaä-Zwakenberg, Mads Græsbøll Christensen, and Richard Heusdens. 2024. Provable privacy advantages of decentralized federated learning via distributed optimization. *IEEE Transactions on Information Forensics and Security*.
- Guoqiang Zhang and Richard Heusdens. 2017. Distributed optimization using the primal-dual method of multipliers. *IEEE Transactions on Signal and Information Processing over Networks*, 4(1):173–187.

A Derivation of Normal Equation

The optimal alignment matrix \mathbf{W} is obtained by minimizing a cost function J that quantifies the discrepancy between the attack embedding matrix \mathbf{E}_A and the transformed victim embeddings $\mathbf{E}_{V \rightarrow A} = \mathbf{E}_V \mathbf{W}$:

$$\begin{aligned} J(\mathbf{W}) &= \frac{1}{2}(\mathbf{E}_A - \mathbf{E}_V \mathbf{W})^T (\mathbf{E}_A - \mathbf{E}_V \mathbf{W}) \\ &= \frac{1}{2}(\mathbf{E}_A^T \mathbf{E}_A - \mathbf{E}_A^T \mathbf{E}_V \mathbf{W} - (\mathbf{E}_V \mathbf{W})^T \mathbf{E}_A \\ &\quad + (\mathbf{E}_V \mathbf{W})^T \mathbf{E}_V \mathbf{W}) \\ &= \frac{1}{2}(\mathbf{E}_A^T \mathbf{E}_A - \mathbf{E}_A^T \mathbf{E}_V \mathbf{W} - \mathbf{W}^T \mathbf{E}_V^T \mathbf{E}_A \\ &\quad + \mathbf{W}^T \mathbf{E}_V^T \mathbf{E}_V \mathbf{W}). \end{aligned} \quad (3)$$

By calculating the derivatives of $J(\mathbf{W})$, we have

$$\begin{aligned} \nabla_{\mathbf{W}} J(\mathbf{W}) &= \frac{1}{2} \nabla_{\mathbf{W}} (\mathbf{E}_A^T \mathbf{E}_A - \mathbf{E}_A^T \mathbf{E}_V \mathbf{W} \\ &\quad - \mathbf{W}^T \mathbf{E}_V^T \mathbf{E}_A + \mathbf{W}^T \mathbf{E}_V^T \mathbf{E}_V \mathbf{W}) \\ &= 2\mathbf{E}_V^T \mathbf{E}_V \mathbf{W} - 2\mathbf{E}_V^T \mathbf{E}_A. \end{aligned} \quad (4)$$

The optimized \mathbf{W} is achieved when the derivative is equal to 0,

$$\mathbf{E}_V^T \mathbf{E}_V \mathbf{W} = \mathbf{E}_V^T \mathbf{E}_A. \quad (5)$$

Then, the matrix \mathbf{W} that minimizes $J(\mathbf{W})$ is

$$\mathbf{W} = (\mathbf{E}_V^T \mathbf{E}_V)^{-1} \mathbf{E}_V^T \mathbf{E}_A. \quad (6)$$

B Topology Construction

To illustrate this approach, consider the syntactic distance matrix obtained from Lang2vec for English (eng), French (fra), and Italian (ita):

$$\mathbf{D} = \begin{bmatrix} 0 & 0.46 & 0.51 \\ 0.46 & 0 & 0.55 \\ 0.51 & 0.55 & 0 \end{bmatrix}$$

where each entry D_{ij} represents the syntactic dissimilarity between language pairs. By applying different threshold values r , we construct distinct topological configurations of language relationships. Fig. 8 demonstrates how the network connectivity varies with increasing r values, revealing:

- At $r = 0.45$: No edges form
- At $r = 0.47$: eng-fra connection emerges
- At $r = 0.52$: eng-ita connection appears while fra-ita remains disconnected
- At $r = 0.56$: Complete graph forms

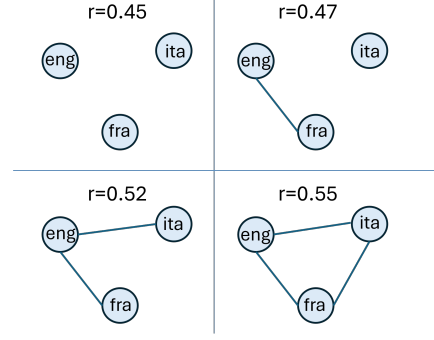


Figure 8: Linguistic topological graph of English, French and Italian with different threshold r . The higher the threshold, the denser the connectivity.

C Other Experimental Results

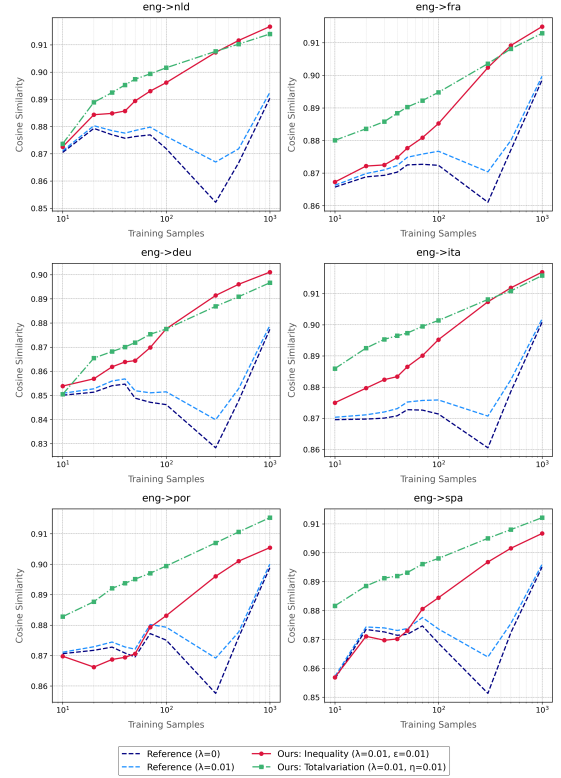


Figure 9: Cross-lingual Inversion Performances with Lang2vec Graph in Cosine Similarities across Training Samples.

Model	Huggingface	Architecture	#Languages	Dimension
FLAN-T5 (Chung et al., 2022)	google/flan-t5-small	Encoder-Decoder	60	512
E5-SMALL-V2 (Wang et al., 2022)	intfloat/e5-small-v2	Encoder	1	384
MT5 (Xue et al., 2020)	google/mt5-base	Encoder-Decoder	102	768
TEXT-EMBEDDING-ADA-002	OpenAI API	Encoder	100+	1536

Table 3: Details of LLMs and Embeddings.

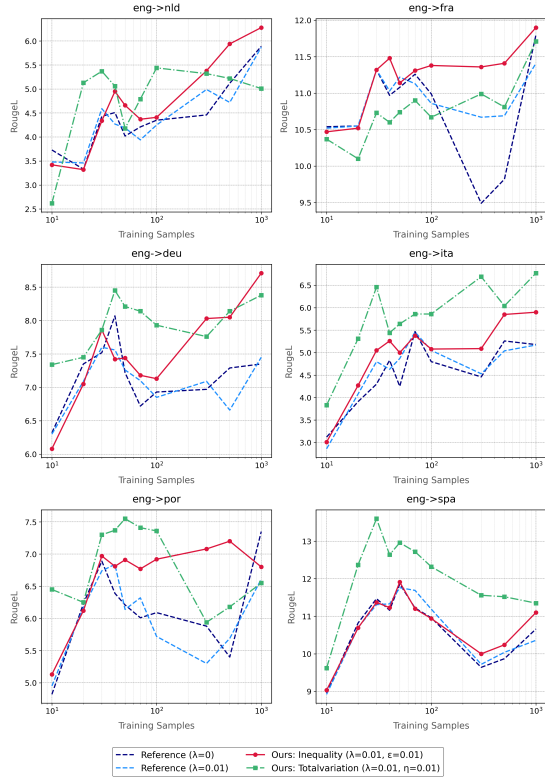


Figure 10: Cross-lingual Inversion Performances with ADA-2 Victim Model in Rouge-L Scores across Training Samples.

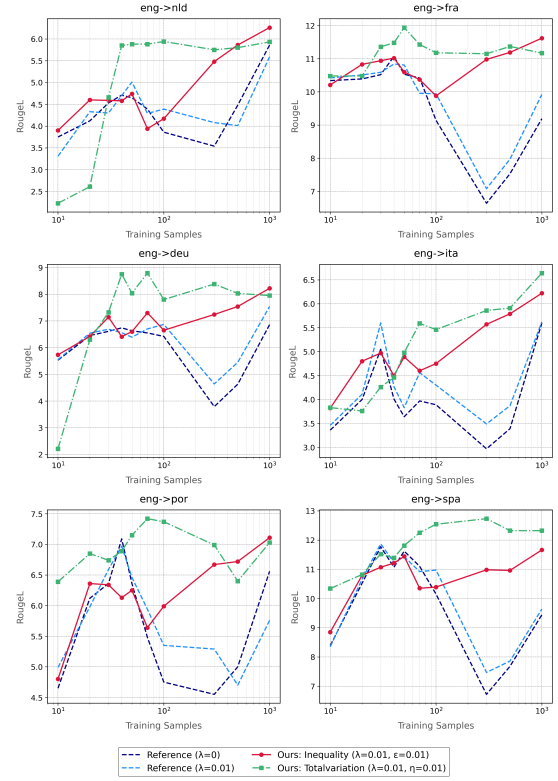


Figure 11: Cross-lingual Inversion Performances with E5 Victim Model in Rouge-L Scores across Training Samples.

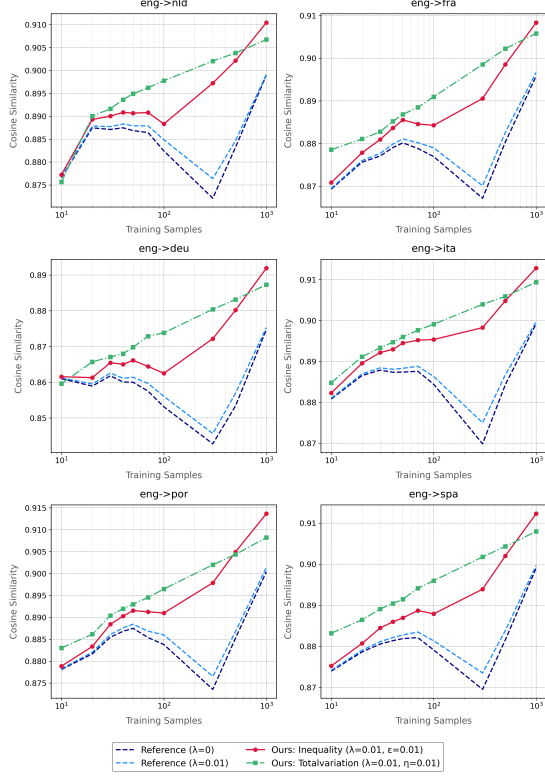


Figure 12: Cross-lingual Inversion Performances with ADA-2 Victim Model in Cosine Similarities across Training Samples.

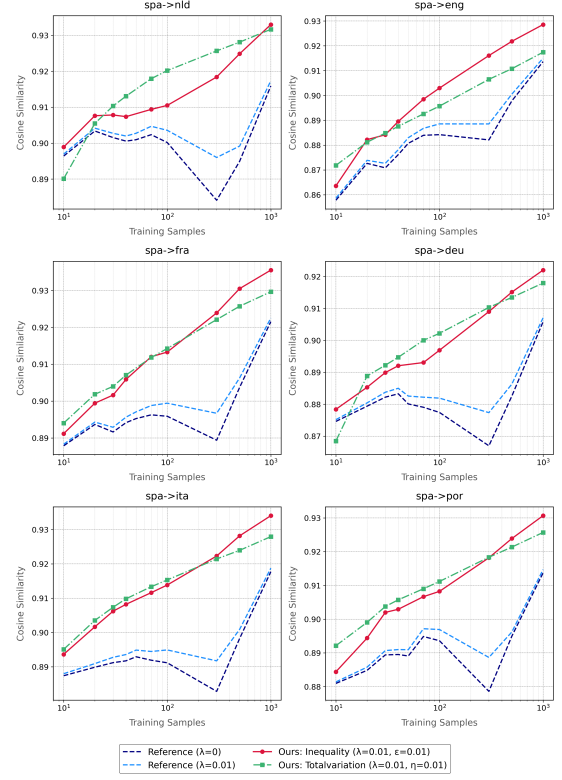


Figure 14: Cross-lingual Inversion Performances with Attack Model trained in Spanish in Cosine Similarities across Training Samples.

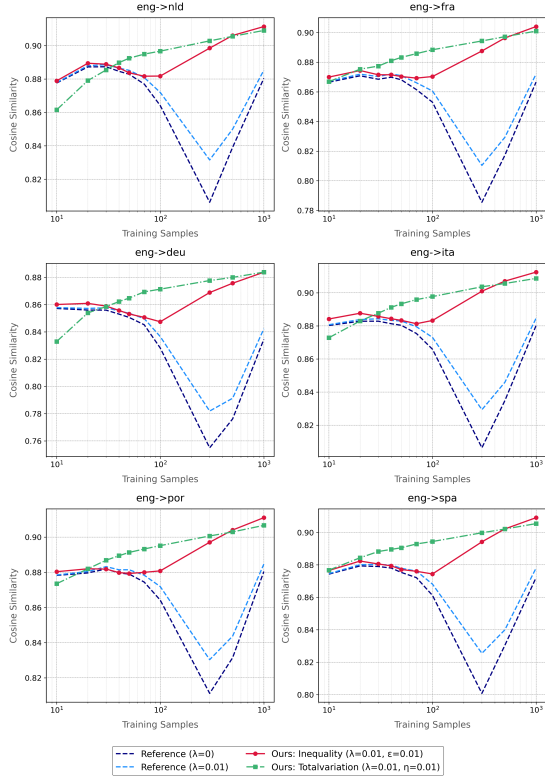


Figure 13: Cross-lingual Inversion Performances with E5 Victim Model in Cosine Similarities across Training Samples.

Lang	ϵ_{dp}	Rouge-L↓		COS↓	
		LapMech	PurMech	LapMech	PurMech
eng->eng	1	13.16	0.0751	13.35	0.0199
	4	12.95	0.0257	12.61	0.0510
	8	14.01	0.0845	13.88	0.1320
	12	13.52	0.1720	13.86	0.1162
eng->fra	1	1.60	-0.0168	1.90	-0.0189
	4	1.77	-0.0161	2.10	0.1081
	8	2.02	0.1040	2.10	0.1428
	12	1.92	0.1271	2.46	0.1853
eng->deu	1	0.86	0.0080	0.62	-0.0240
	4	0.99	0.0259	0.62	-0.0216
	8	0.77	0.0960	0.64	0.0881
	12	0.70	0.1815	1.22	0.1944
eng->sap	1	1.58	0.0431	1.78	0.0729
	4	1.35	0.0318	1.45	0.0360
	8	1.87	0.2408	1.94	0.1119
	12	1.65	0.1875	2.29	0.1846

Table 4: Cross-lingual Inversion Performance with $|D_V|=100$ on Classification Tasks on SNLI dataset with Local DP (Total Variation). From a defender’s perspective, ↓ means lower are better.

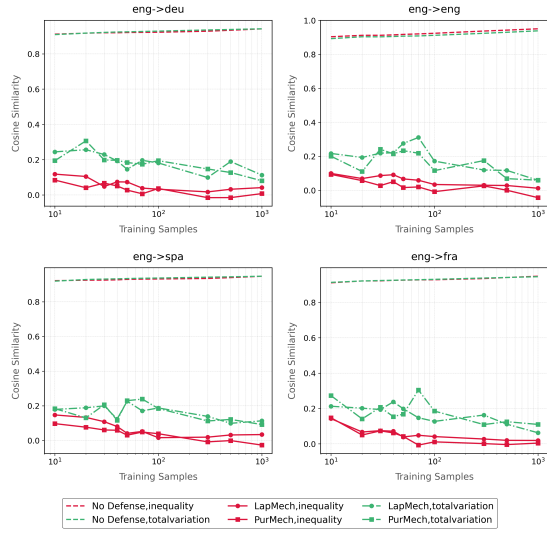


Figure 15: Cross-lingual Inversion Performance on Classification Tasks on SNLI dataset with Local DP ($\epsilon_{dp} = 12$) in Cosine Similarities.

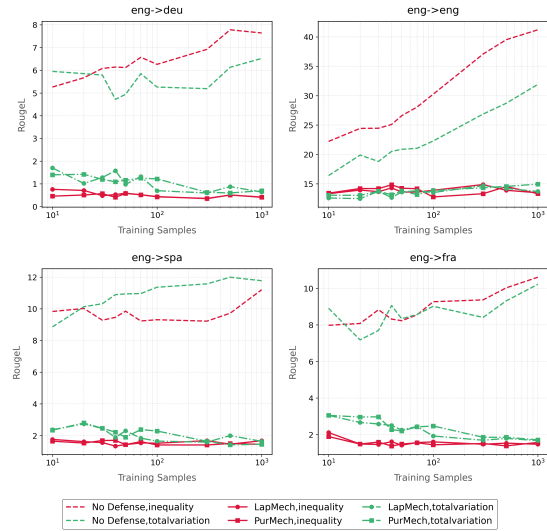


Figure 16: Cross-lingual Inversion Performance on Classification Tasks on SNLI dataset with Local DP ($\epsilon_{dp} = 12$) in Rouge-L Scores.