DLIR: Spherical Adaptation for Cross-Lingual Knowledge Transfer of Sociological Concepts Alignment

Zeqiang Wang[1], Jon Johnson[2], Suparna De[1]*
[1]University of Surrey, Guildford, United Kingdom,
[2]University College London, London, United Kingdom
Correspondence: s.de@surrey.ac.uk

Abstract

Cross-lingual alignment of nuanced sociological concepts is crucial for comparative cross-cultural research, harmonising longitudinal studies, and leveraging knowledge from social science taxonomies (e.g., ELSST). However, aligning these concepts is challenging due to cultural context-dependency, linguistic variation, and data scarcity, particularly for low-resource languages. Existing methods often fail to capture domain-specific subtleties or require extensive parallel data. Grounded in a Vector Decomposition Hypothesis—positing separable domain and language components within embeddings, supported by observed language-pair specific geometric structures—we propose DLIR (Dual-Branch LoRA for Invariant Representation). DLIR employs parallel Low-Rank Adaptation (LoRA) branches: one captures core sociological semantics (trained primarily on English data structured by the ELSST hierarchy), while the other learns language invariance by counteracting specific language perturbations. These perturbations are modeled by Gaussian Mixture Models (GMMs) fitted on minimal parallel concept data using spherical geometry. DLIR significantly outperforms strong baselines on cross-lingual sociological concept retrieval across 10 languages. Demonstrating powerful knowledge transfer, English-trained DLIR substantially surpasses target-language (French/German) LoRA fine-tuning even in monolingual tasks. DLIR learns disentangled, language-robust representations, advancing resource-efficient multilingual understanding and enabling reliable cross-lingual comparison of sociological constructs.

1 Introduction

Cross-lingual comprehension of specialized domains is a long-standing challenge in natural language processing, particularly for sociological texts, where concepts are culturally bound and sensitive to contextual cues. For example, terms such as "unemployment," "social inequality," and "economic stagnation" can have significantly different cross-cultural interpretations despite similar terminology.

Multilingual pre-trained language models have advanced the state of the art in cross-lingual tasks but display significant limitations in capturing domain-specific nuances. Even when text segments appear lexically similar, they may invoke different theoretical frameworks in varying socio-political contexts. For instance, consider an English–Spanish pair: "*Income inequality reflects structural economic factors*" and "*La desigualdad de ingresos refleja factores económicos estructurales*." Although lexically aligned, one text may emphasize market-driven disparities, while the other refers to class-related structures regulated by the state. This divergence highlights the inadequacy of conventional similarity metrics for specialized applications.

The situation worsens in low-resource languages that lack large-scale, domain-specific corpora, making fine-tuning or extensive parallel training impractical. Basic alignment methods overlook crucial conceptual distinctions, particularly within complex sociological frameworks where hierarchical and interlinked concepts create a specialized knowledge graph (Li et al., 2025).

To address these limitations, we propose an approach grounded in a **vector decomposition hypothesis**: embeddings of domain-specific texts in high-resource languages can be decomposed into a *domain knowledge vector* and a *language-specific feature vector*. We suggest that a representation $\mathbf{A}$ includes a core domain component (sociological semantics) and a language-specific component. By learning perturbation vectors on a unit hypersphere from a small set of parallel concept pairs, we transform high-resource embeddings into approxi-

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 2061–2075
November 4-9, 2025

mations of low-resource counterparts, preserving domain-level semantics.

We implement this hypothesis using a **dual-branch low-rank adaptation (LoRA)** design, with one branch capturing domain knowledge and the other introducing language perturbations. Spherical geometric operations yield perturbations that realign high-resource embeddings with low-resource counterparts.

Our contributions are: (i) positing and providing empirical evidence for Vector Decomposition Hypothesis, demonstrating effective disentanglement of domain and language features with limited parallel data; (ii) introducing **Spherical Noise Injection**, utilizing logarithmic and exponential mappings on the unit hypersphere with Gaussian mixture modeling for cross-lingual alignment; (iii) developing a **Dual-Branch LoRA** mechanism to effectively separate domain knowledge from language-specific perturbations; and (iv) showing through rigorous experimentation on sociological corpora that our method surpasses standard baselines in cross-lingual concept alignment.

As illustrated in Figure 1, the proposed DLIR framework offers a resource-efficient and linguistically grounded solution to align sociological concepts between languages, particularly benefiting low-resource and specialized domains [1] .

2 Related Work

Multilingual Text Embeddings. Early research on multilingual embeddings often focuses on removing language-specific artifacts to obtain unified representations. For instance, Tiyajamorn et al. (2021) extract language-agnostic embeddings by filtering out language-specific signals, while Feng et al. (2022) combine translation and masked language modeling objectives to reduce parallel data requirements. More recent models adopt large-scale contrastive pre-training to enhance cross-lingual alignment: Multilingual E5 (Wang et al., 2024) leverages a billion multilingual text pairs, NV-Embed (Lee et al., 2025) removes causal masks to improve encoder-based embeddings, and M3-Embedding (Chen et al., 2024) extends retrieval functionality across more than one hundred languages. Although these efforts produce state-of-the-art performance in general multilingual tasks, fine-grained alignment of specialized concepts (e.g., sociological or biomedical terminologies) remains a persistent challenge.

Progress in LoRA. To address computational bottlenecks in adapting large language models, a variety of parameter-efficient fine-tuning (PEFT) methods have emerged, with LoRA serving as a prominent example. Building on its low-rank decomposition approach, recent work has explored new dimensions of LoRA. Zhang et al. (2023) dynamically allocate rank across layers, LoRETTA (Yang et al., 2024) employs tensor-train decomposition, and DoRA (Liu et al., 2024) distinguishes weight magnitude and direction for better adaptability. Sparse adjustments have also been introduced, as seen in SoRA (Ding et al., 2023), while VeRA (Kopiczko et al., 2024) further reduces trainable parameters by sharing low-rank matrices across layers. Dettmers et al. (2023) complement LoRA with quantization to achieve full-scale fine-tuning on single-GPU setups. Despite these innovations, most LoRA-based methods focus on generic downstream tasks without explicitly separating domain-specific concepts from language-inherent features.

Cross-Lingual Knowledge Transfer. Techniques for cross-lingual alignment range from classical geometric mappings (Jawanpuria et al., 2019) to more recent strategies that adjust pretrained models using parallel corpora or dictionaries (Kulshreshtha et al., 2020). Some approaches incorporate alignment signals in model pre-training (Li et al., 2024), while others merge adapters for distinct languages and tasks (Zhao et al., 2024). Prompt-tuning has also been applied to multilingual dialogues (Tu et al., 2024). Although these methods demonstrate promising zero-shot capabilities, they often rely on large bilingual resources or fail to isolate the conceptual domain from underlying linguistic variations. In contrast, our proposed framework unifies *dual-branch LoRA* with *spherical noise modeling* to achieve domain-sensitive concept alignment using only minimal parallel data. This design explicitly disentangles domain knowledge from language perturbations, addressing key limitations in existing cross-lingual solutions.

3 Methodology

This section details **DLIR (Dual-Branch LoRA for Invariant Representation)**, targeting effective

[1]To facilitate reproducibility, the code and data for this research will be made available upon acceptance.

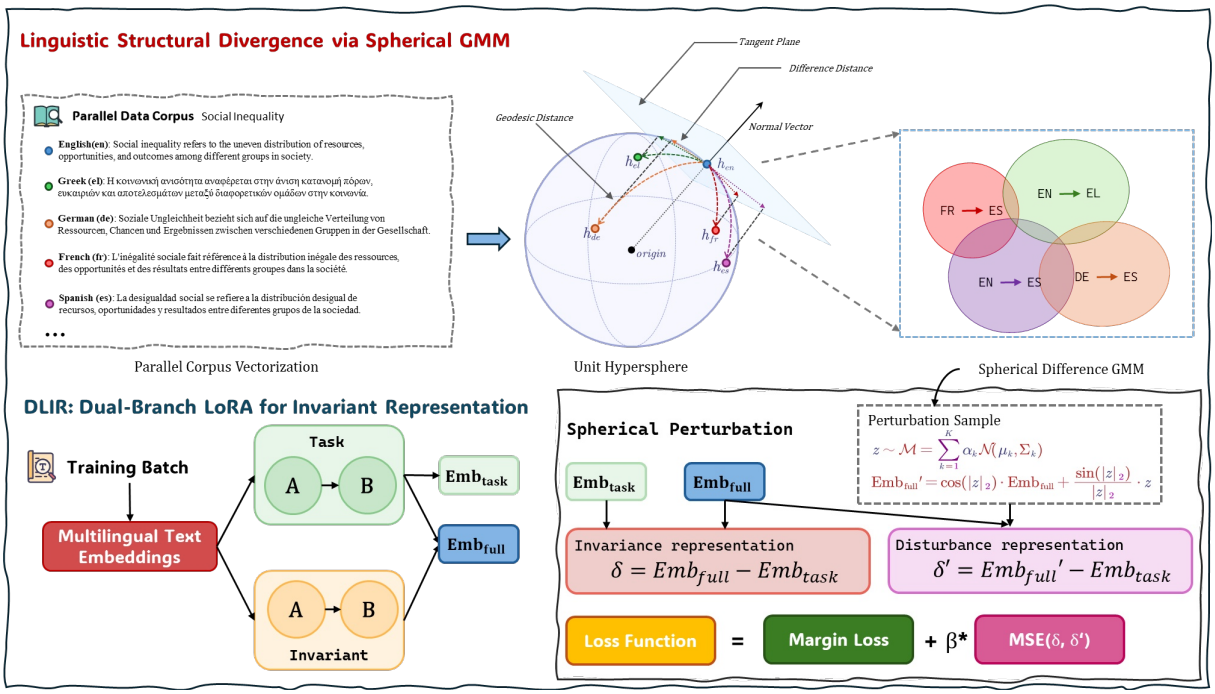

Figure 1: Overview of the DLIR framework. Top: linguistic divergence is modeled by projecting multilingual embeddings onto a hypersphere and learning geodesic deltas via Gaussian Mixture Models (GMMs). Bottom: a dual-branch LoRA encoder with GMM-based perturbations enforces semantic consistency under language variation.

cross-lingual knowledge transfer in specialized domains like sociology, where concept alignment is challenged by nuance and cultural context. DLIR achieves robust transfer by explicitly disentangling domain semantics from language features. This approach is grounded in the **Vector Decomposition Hypothesis**—that embeddings contain separable domain and language components, empirically supported via Gaussian Mixture Model (GMM) analysis of parallel concept differences on a hypersphere. DLIR implements this using a **Dual-Branch LoRA architecture**: one branch captures domain structure (primarily from English), while the other, guided by GMMs, counteracts language variations. This yields language-invariant domain representations trained via a composite objective combining domain ranking and spherical invariance learning.

3.1 Task Formulation and Evaluation Scope

Evaluation is conducted via a cross-lingual concept retrieval task within sociology. Given a source text c_ℓ in a target language ℓ (e.g., Spanish, French, Greek) describing a sociological concept, the goal is to retrieve the corresponding English text t^* from a candidate set $\{t_1, t_2, \ldots, t_n\}$; note that the set contains one correct match (t^*) and $n-1$ distractors (related but distinct concepts). Both c_ℓ and the English texts t_i are detailed descriptions that go beyond simple keyword matching.

The task is cast as semantic matching on the unit hypersphere $\mathcal{S}^{d-1}$, where d is the dimensionality of the embedding space. An embedding function

$$E : \text{Text} \to \mathcal{S}^{d-1}$$

maps texts into a shared space, and prediction is performed using cosine similarity:

$$t^* = \arg\max_{t_i} \text{SIM}(E(c_\ell), E(t_i)).$$

While primary evaluation is target-language-to-English retrieval, the ultimate goal is to build a language-invariant space facilitating bidirectional transfer. The training leverages the European Language Social Science Thesaurus (ELSST) hierarchy[2] to structure the semantic space. ELSST is a broad-based, multilingual thesaurus for the social sciences, covering core disciplines such as politics, sociology, and economics, and is designed to facilitate access to data resources across Europe.

3.2 Vector Decomposition Hypothesis

The core hypothesis asserts that the embedding $\mathbf{h}$ of a domain-specific text can be decomposed into

[2] https://elsst.cessda.eu/index.html

two primary components: $\mathbf{v}_{\text{domain}}$, representing the language-agnostic semantic core (e.g., the sociological theory), and $\mathbf{v}_{\text{language}}$, capturing language-specific features (such as lexical or syntactic patterns). For two texts $\mathcal{A}$ (English) and $\mathcal{B}$ (French) conveying the same concept, their embeddings $\mathbf{a} = E(\mathcal{A})$ and $\mathbf{b} = E(\mathcal{B})$ should share the same $\mathbf{v}_{\text{domain}}$ while differing in $\mathbf{v}_{\text{language}}$. Our objective is to learn transformations (parameterized by language-pair specific offsets) so that

$$\mathbf{b} \approx \text{Transform}\big(\mathbf{a};\, \mathbf{z}_{\text{EN}\to\text{FR}}\big)$$
$$\mathbf{a} \approx \text{Transform}\big(\mathbf{b};\, \mathbf{z}_{\text{FR}\to\text{EN}}\big)$$

3.3 Empirical Support via GMM on Spherical Differences

To validate the hypothesis, we analyze the differences between embeddings of parallel concepts. Since text embeddings are L2-normalized, they reside on the surface of a unit hypersphere—a curved manifold. Standard Euclidean subtraction inadequately measures differences in this space. Therefore, we employ principles from Riemannian geometry to capture these differences accurately. For a parallel pair (A_i, B_i) with embeddings (a_i, b_i) on the hypersphere, we use the *Logarithmic Map* (log) to project a_i onto the tangent space at b_i. This yields a tangent vector u_i (representing the true geometric difference in a locally flat space), which is computed as:

$$\mathbf{u}_i = \frac{\theta_i}{\sin\theta_i}\Big(\mathbf{a}_i - \cos\theta_i\, \mathbf{b}_i\Big), \quad \theta_i = \arccos\big(\mathbf{a}_i^\top \mathbf{b}_i\big).$$

The collection $\{\mathbf{u}_i\}$ is then modeled with a GMM:

$$p(\mathbf{u} \mid \Theta) = \sum_{k=1}^{K} \alpha_k \mathcal{N}\Big(\mathbf{u} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\Big),$$

where the parameters $\Theta = \{(\alpha_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ are estimated via Expectation-Maximization (EM) algorithm (Dempster et al., 1977). The resulting clusters (as visualized in Figure 2, which shows distinct groupings of difference vectors based on the language pair) reveal systematic, language-dependent shifts, which both validate our hypothesis and provide a generative model of language perturbations (denoted $\mathcal{M}$) for subsequent training.

We specifically chose GMMs because they provide a generative model necessary for sampling perturbations in our invariance loss, and they can capture multi-modal distributions (i.e., multiple systematic shifts between a language pair), balancing expressive power with the constraints of minimal parallel data.

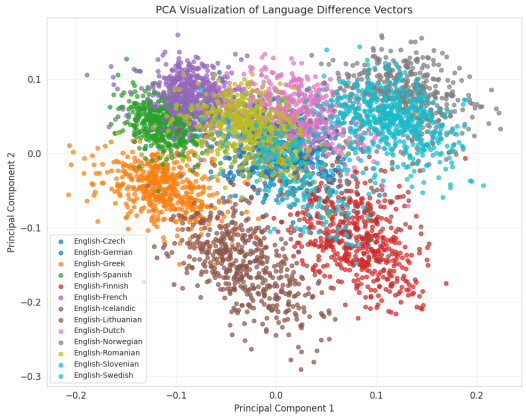

Figure 2: PCA visualization of difference vectors $\mathbf{u}_i = \text{Log}_{\mathbf{b}_i}(\mathbf{a}_i)$ colored by language pair, supporting the existence of systematic, language-dependent offsets.

3.4 DLIR: Dual-Branch LoRA for Invariant Representation

DLIR operationalizes the hypothesis via a dual-branch LoRA design. Standard LoRA adapts a model $M(\cdot; W)$ with low-rank updates $\Delta W = BA^\top$. Here, we introduce two parallel updates:

$$\Delta W_{\text{task}} = B_{\text{task}} A_{\text{task}}^\top, \quad \Delta W_{\text{inv}} = B_{\text{inv}} A_{\text{inv}}^\top.$$

The **Task Branch** (ΔW_{task}) learns the domain-specific structure from the ELSST hierarchy (using predominantly English data), targeting $\mathbf{v}_{\text{domain}}$. The **Invariance Branch** (ΔW_{inv}) learns to neutralize $\mathbf{v}_{\text{language}}$ by counteracting language-specific offsets. The final representation is computed as

$$M_{\text{both}}(x) \;=\; M\Big(x;\, W + \frac{\alpha}{r}\big(\Delta W_{\text{task}} + \Delta W_{\text{inv}}\big)\Big),$$

where r is the LoRA rank and α is the scaling factor used to adjust the magnitude of the adaptation. while a task-only representation is given by

$$M_{\text{task}}(x) = M\Big(x; W + \frac{\alpha}{r}\Delta W_{\text{task}}\Big).$$

During training, language perturbations are simulated by sampling $\tilde{\mathbf{z}} \sim \mathcal{M}$ and applying the spherical exponential map:

$$\text{Exp}_{\mathbf{q}}(\mathbf{z}) = \cos\|\mathbf{z}\|_2\, \mathbf{q} + \frac{\sin\|\mathbf{z}\|_2}{\|\mathbf{z}\|_2}\,\mathbf{z} \quad (\mathbf{z} \neq \mathbf{0}),$$

so that an embedding $\mathbf{h}$ is perturbed as $\mathbf{h}' = \text{Exp}_{\mathbf{h}}(\tilde{\mathbf{z}})$.

3.5 Learning Objectives: Domain Structure and Invariance

DLIR is trained by optimizing a composite loss that combines domain structure learning and invariance:

Hierarchical Ranking Loss ($\mathcal{L}_{\text{task}}$) For an anchor t_{anchor} and texts t_{τ_1}, t_{τ_2} with t_{τ_1} closer than t_{τ_2}, we define the margin-based loss as:

$$\Delta_{\tau_1,\tau_2} = \text{SIM}(\mathbf{h}_{anchor}, \mathbf{h}_{\tau_1}) - \text{SIM}(\mathbf{h}_{anchor}, \mathbf{h}_{\tau_2})$$
$$\mathcal{L}_{\text{task}} = \sum_{(\tau_1,\tau_2)\in\mathcal{P}} \max\{0,\, m - \Delta_{\tau_1,\tau_2}\}$$

where $\mathcal{P}$ represents the set of all ranking pairs derived from the ELSST hierarchy (e.g., comparing a narrower concept t_{τ_1} vs. a broader concept t_{τ_2} relative to the anchor t_{anchor}), and m is a margin hyperparameter. This standard margin-based ranking loss encourages the model to learn representations reflecting the underlying semantic hierarchy.

Spherical Invariance Loss ($\mathcal{L}_{\text{inv}}$) Defining the contribution of the invariance branch as $\boldsymbol{\delta}_j = \mathbf{h}_{\text{both},j} - \mathbf{h}_{\text{task},j}$ and perturbing the full embedding to obtain $\mathbf{h}'_{\text{both},j} = \text{Exp}_{\mathbf{h}_{\text{both},j}}(\tilde{\mathbf{z}}_j)$ (with $\tilde{\mathbf{z}}_j \sim \mathcal{M}$), the loss is given by

$$\mathcal{L}_{\text{inv}} = \frac{1}{|\text{Batch}|} \sum_{j\in\text{Batch}} \mathbb{E}_{\tilde{\mathbf{z}}_j} \left\| \boldsymbol{\delta}_j - \left(\mathbf{h}'_{\text{both},j} - \mathbf{h}_{\text{task},j} \right) \right\|_2^2.$$

The intuition behind minimizing this loss is as follows: $\boldsymbol{\delta}_j$ represents the learned language-specific vector added by the invariance branch. $\mathbf{h}'_{\text{both},j}$ is the embedding perturbed by realistic language noise sampled from $\mathcal{M}$. Minimizing the squared difference between the original offset $\boldsymbol{\delta}_j$ and the offset remaining after perturbation $(\mathbf{h}'_{\text{both},j} - \mathbf{h}_{\text{task},j})$ forces the invariance branch to counteract the effects of language variations modeled by the GMM, thus encouraging the task branch embedding $\mathbf{h}_{\text{task},j}$ to become invariant to these specific linguistic shifts. Minimizing this loss thus encourages the task branch embedding $\mathbf{h}_{\text{task},j}$ to represent the stable, language-invariant semantic core, while the invariance branch learns to dynamically counteract the specific language variations modeled by $\mathcal{M}$.

Overall Training Objective The final loss is a linear combination:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \beta\, \mathcal{L}_{\text{inv}},$$

with β balancing domain fidelity and invariance.

Training starts with offline fitting of the GMM $\mathcal{M}$ using limited parallel data. The dual LoRA branches ($A_{\text{task}}, B_{\text{task}}, A_{\text{inv}}, B_{\text{inv}}$) are then iteratively updated (keeping the base model fixed) on English hierarchical text data to minimize $\mathcal{L}_{\text{total}}$.

This framework thus achieves effective cross-lingual transfer: the task branch captures domain structure while the invariance branch, guided by GMM-modeled perturbations, factors out language-specific features. The resulting embedding space $E_{\text{both}}(\cdot)$ clusters texts describing the same sociological concept regardless of language, facilitating robust cross-lingual alignment with high parameter efficiency and minimal reliance on parallel data.

4 Experiments

In this section, we empirically evaluate the effectiveness of our proposed DLIR. We detail the dataset derived from sociological resources, the baseline models and ablation studies conducted for comparison, the evaluation tasks and metrics employed, and the specific implementation settings.

4.1 Experimental Setup

Datasets and Tasks. Our experiments leverage the European Language Social Science Thesaurus (ELSST) hierarchy, complemented by synthetic data for realistic training and evaluation scenarios. Minimal parallel data (concept labels/definitions) from ELSST was used *only* for offline GMM fitting ($\mathcal{M}$) for all 13 language pairs listed in Appendix Table 7. The primary cross-lingual concept retrieval evaluation (results in Tables 2, 3, 4) was performed on the 10 target languages with sufficient test data (more than 50 samples, excluding Czech, Icelandic, and Slovenian).

- **Core Training Data (English Hierarchical Ranking).** We created a core English training set using ELSST concepts. To ensure the model learns robust semantic representations beyond simple keyword matching, for each concept, diverse descriptive texts were generated via DeepSeek-V3. The generation process employed varied prompts designed to elicit nuanced expressions of sociological concepts. We utilized a template-based strategy varying context (e.g., academic, applied), perspective (e.g., researcher, policymaker), and style to maximize diversity (See Appendix A.3 for specific prompt templates and generation details). These prompts incorporated

hypothetical scenarios (e.g., "Describe a situation where social stratification is clearly evident in a modern city"), different personas (e.g., "Explain youth unemployment from the perspective of a recent graduate"), and diverse linguistic styles, all aimed at encouraging implicit referencing of the core concept without its direct naming. This approach aimed to create training data that reflects the complexity and contextual dependency of sociological terms. This synthetic dataset, structured by ELSST's hierarchy (e.g., sampling broader/narrower relations for ranking pairs (τ_1, τ_2)), was exclusively used to optimize the hierarchical ranking loss $\mathcal{L}_{\text{task}}$ (§3.5) for all model variants (DLIR, LoRA, Gaussian).

- **Synthetic Data Quality Assurance.** To validate the quality of the generated texts, we conducted a human evaluation with domain experts. A random sample of 40 concept-text pairs (focusing on the crucial 'self' relation) was evaluated for Accuracy (1-5 scale, measuring alignment with the ELSST definition) and Fluency (1-3 scale). The evaluation yielded a high average Accuracy score of 3.95 (out of 5) and a Fluency score of 2.71 (out of 3), confirming the data's suitability for training nuanced conceptual representations. (See Appendix A.3 for detailed results and the evaluation protocol).

- **GMM Training Data (Parallel Concept Pairs).** To model language-specific perturbations (§3.3) for the invariance loss $\mathcal{L}_{\text{inv}}$ (§3.5), minimal parallel data was extracted from ELSST: pairs of English concept labels/definitions $(\mathcal{A}_i)$ and their counterparts $(\mathcal{B}_i)$ in 10 target languages (incl. Spanish, French, Romanian). The number of pairs per language ranged from **73 to 843** (details in Appendix Table 7). This data was used *only* for the offline fitting of language-pair specific Gaussian Mixture Models (GMMs, $\mathcal{M}$) and was not involved in parameter gradient updates during main model training.

- **Evaluation Task 1: Cross-Lingual Concept Retrieval.** Our primary benchmark (results in Table 2) evaluates retrieving the correct English synthetic text t^* for a given target-language synthetic query c_ℓ from a candidate set $\{t_1, \ldots, t_n\}$. Distractors t_i $(i \neq *)$ represent other ELSST concepts. The candidate set size n varied per query, reflecting a dynamic negative sampling strategy designed to enhance evaluation realism: distractors prioritized semantically related concepts (hard negatives) based on ELSST structure density (e.g., siblings, parent/child concepts relative to t^*), supplemented by a smaller number of randomly selected, unrelated concepts. Consequently, the number of candidates n per query ranged from 2 to 7 (mean=3.08, mode=2). Evaluation metrics (R@1, MRR, NDCG) were computed based on the rank within each query's specific candidate list, accommodating the variable list size.

- **Evaluation Task 2: Monolingual Concept Ranking.** We assessed representation quality and transferability via monolingual concept ranking (Table 1) in English, French, and German. The English data is the core training set (§ 4.1); French/German data are its translations via the ByteDance service. Crucially, this translated data was used only for training the monolingual LoRA-FR/DE baselines and for this specific evaluation task; the DLIR model itself never encounters this translated data during training. Given an anchor text, models rank other texts describing related concepts (per ELSST hierarchy: self, narrower, related, broader, see §3.5) within the same language. The number of texts to rank per anchor (i.e., list size) varied slightly, ranging from 9 to 12 (mode=11), primarily due to the differing density of concepts within specific relationship categories (e.g., 'related' or 'broader' neighbors) in the ELSST hierarchy relative to the anchor concept. This task crucially compares English-trained DLIR against LoRA models fine-tuned directly on the target language (LoRA-FR/DE), highlighting knowledge transfer capabilities.

Evaluation Metrics. We evaluate performance using standard retrieval and ranking metrics: Recall@1 (R@1), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain (NDCG).

Base Model. We employ `multilingual e5 small` as the base model for all fine-tuning experiments.

Compared Methods. We compare our proposed method, DLIR, against several baselines and state-

of-the-art models:

- **E5s (Base Model):** The zero-shot performance of `multilingual e5 small` without any fine-tuning.
- **LoRA:** Standard single-branch LoRA fine-tuning, optimized *only* on the **English synthetic descriptive text data** using the task loss $\mathcal{L}_{\text{task}}$. This baseline assesses the effectiveness of basic domain adaptation.
- **LoRA-FR / LoRA-DE (for Table 1 only):** Standard single-branch LoRA models fine-tuned monolingually, *only* on the **translated French or German synthetic descriptive text data**, respectively, using $\mathcal{L}_{\text{task}}$. These serve as strong baselines to evaluate the cross-lingual transfer capabilities of DLIR against models trained directly on the target language data.
- **Gaussian:** This baseline utilizes the same **dual-branch LoRA architecture** as DLIR and is trained with the identical composite loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \beta\mathcal{L}_{\text{inv}}$. The crucial difference lies in the invariance training: the perturbation vectors $\tilde{\mathbf{z}}$ for the spherical noise injection step $\mathbf{h}' = \text{Exp}_{\mathbf{h}}(\tilde{\mathbf{z}})$ are sampled from a standard **isotropic Gaussian distribution** ($\mathcal{N}(0, \sigma^2 I)$ with $\sigma = 0.01$), subsequently projected onto the tangent space at $\mathbf{h}$), *instead of* the learned GMM $\mathcal{M}$. This baseline isolates the effect of using language-pair-specific GMM modeling versus employing generic random noise for promoting invariance. The task loss $\mathcal{L}_{\text{task}}$ is trained on the English synthetic data.
- **DLIR (Proposed Method):** Our complete proposed method. It employs the **dual-branch LoRA architecture**, trains the task branch using $\mathcal{L}_{\text{task}}$ on the **English synthetic descriptive text data**, and simultaneously trains the invariance branch using $\mathcal{L}_{\text{inv}}$. The perturbation vectors $\tilde{\mathbf{z}}$ for the spherical noise injection are sampled from the GMM $\mathcal{M}$ learned offline from limited parallel ELSST concept pairs.
- **SOTA Models:** We compare against publicly available state-of-the-art multilingual embedding models in a zero-shot setting:
  - **KaLM:** KaLM embedding multilingual mini v1 is a multilingual embedding model developed by HITsz-TMG, leveraging high-quality training data and advanced training techniques to achieve superior performance across multiple languages (Hu et al., 2025).
  - **OpenAI:** text embedding ada 002 is OpenAI's state-of-the-art embedding model designed to capture deep semantic meanings, widely used for tasks such as semantic search and text similarity `https://platform.openai.com/docs/guides/embeddings`.
  - **E5l:** multilingual e5 large is a large-scale multilingual embedding model from the E5 family, supporting 100 languages and excelling in tasks like passage retrieval and semantic similarity (Wang et al., 2024).

4.2 Results and Analysis

This section presents the quantitative results of our experiments and provides analysis comparing our proposed method against baselines and ablations.

Overall Performance in Cross-Lingual Retrieval. Table 2 summarizes the overall average performance across 10 languages on the cross-lingual concept retrieval task. Our proposed method, **DLIR**, achieves the best performance among all compared models (R@1 0.821, MRR 0.902, NDCG 0.928), surpassing strong zero-shot baselines like E5l (R@1 0.794), KaLM (R@1 0.713), and Ada-002 (R@1 0.695). While not a direct apples-to-apples comparison (as DLIR utilizes minimal parallel data for GMM fitting), this comparison highlights the practical efficiency of our approach. DLIR attains this SOTA performance by adapting the significantly smaller `E5-small` model (110M parameters) with minimal LoRA updates (approx. 0.3M parameters, <0.3% of base model), demonstrating high parameter efficiency compared to the larger E5l (560M).

Ablation Study: Deconstructing DLIR's Gains. The ablation study (Table 2 bottom, Appendix Table 3) shows clear gains from each component. Basic LoRA improves substantially over E5s (+0.135 R@1), confirming the benefits of domain adaptation. Adding dual-branch and generic Gaussian noise (`Gaussian` baseline) yields further gains (+0.023 R @ 1), suggesting that encouraging invariance is beneficial. This improvement may also

Model	English			French			German		
	R@1	MRR	NDCG	R@1	MRR	NDCG	R@1	MRR	NDCG
E5 small	0.407	0.587	0.686	0.533	0.676	0.753	0.359	0.531	0.640
LoRA	0.680	0.788	0.839	0.630	0.747	0.808	0.532	0.660	0.740
%↑ vs E5s	*(+67.0)*	*(+34.2)*	*(+22.3)*	*(+18.2)*	*(+10.5)*	*(+7.3)*	*(+48.0)*	*(+24.3)*	*(+15.6)*
DLIR	**0.884**	**0.922**	**0.941**	**0.806**	**0.865**	**0.897**	**0.680**	**0.778**	**0.831**
%↑ vs E5s	*(+117.2)*	*(+57.1)*	*(+37.2)*	*(+51.1)*	*(+27.9)*	*(+19.1)*	*(+89.2)*	*(+46.6)*	*(+29.8)*

Table 1: Performance on **monolingual concept ranking** tasks in English, French, and German. DLIR (trained only on English data) is compared against the E5s baseline and standard LoRA fine-tuning. High performance of DLIR in French and German demonstrates effective **cross-lingual knowledge transfer**. Metrics: R@1, MRR, and NDCG. Percentage gains are relative to E5 baseline.

Model	R@1	MRR	NDCG
Comparison with State-of-the-Art Models			
KaLM	0.713	0.840	0.881
OpenAI	0.695	0.831	0.875
E5l	<u>0.794</u>	<u>0.891</u>	<u>0.919</u>
DLIR	**0.821**	**0.902**	**0.928**
Ablation Study for DLIR			
E5s	0.647	0.803	0.853
LoRA	0.782	0.879	0.910
Gaussian	0.805	0.893	0.921

Table 2: Overall average performance on the cross-lingual concept retrieval task. Best scores within each section are in **bold**, second best are <u>underlined</u>. Detailed per-language results are in Appendix Tables 3 and 4.

indicate that the invariance objective inherently enhances model robustness by introducing noise during training, a phenomenon worthy of future investigation. Crucially, replacing Gaussian noise with GMM-sampled perturbations (full **DLIR**) gives the largest final boost (+0.016 R@1). This strongly supports our hypothesis: explicitly modeling systematic, geometric language differences via GMMs enables more effective disentanglement than unstructured noise, validating the Vector Decomposition Hypothesis and our GMM-guided spherical invariance mechanism.

Monolingual Performance and Knowledge Transfer. Table 1 presents additional evidence for the effectiveness of DLIR in the transfer of knowledge between languages. This evaluation assesses performance on monolingual concept ranking tasks in English, French, and German. The key comparison is between DLIR (trained *only* on English synthetic data) and the LoRA-FR/LoRA-DE baselines (trained *directly* on the corresponding translated French/German synthetic data). Remarkably, DLIR significantly outperforms the monolingual LoRA baselines on their respective languages. For instance, on French, DLIR achieves an R@1 of 0.806, far exceeding LoRA-FR's 0.630 (a relative improvement of 27.9%). Similarly, on German, DLIR achieves an R@1 of 0.680 compared to LoRA-DE's 0.532 (a relative improvement of 27.8%). DLIR also shows massive gains over the zero-shot baseline E5s across all three languages (e.g., a 117.2% R@1 gain in English, 51.2% in French and 89.4% in German). These results demonstrate that knowledge about the conceptual structure of the sociological domain, primarily learned from English data, is effectively transferred to French and German through the invariant representation learned. The disentanglement mechanism enables DLIR to apply this domain knowledge more successfully in new languages than standard fine-tuning directly performed on translated data, highlighting DLIR's ability to learn truly transferable, language-robust representations. Collectively, these results strongly back the Vector Decomposition Hypothesis and demonstrate the practical efficacy of our GMM-guided spherical adaptation for robust cross-lingual knowledge transfer in specialized domains.

4.3 Qualitative Analysis

To illustrate how DLIR captures domain-specific nuances more effectively than baselines, we present case studies from the cross-lingual retrieval task.

Case 1: Disambiguating Related Concepts (German to English). Consider a German query de-

scribing *structural unemployment* (*strukturelle Arbeitslosigkeit*)—unemployment resulting from industrial reorganization, rather than fluctuations in demand. The E5l baseline (strongest zero-shot model) incorrectly prioritized an English text discussing *cyclical unemployment*, likely misled by high lexical overlap in general economic terms (e.g., "recession," "layoffs"). In contrast, DLIR correctly retrieved the corresponding structural unemployment text. This suggests that the invariance training successfully suppressed language-specific lexical signals, allowing the core conceptual distinction (structural vs. cyclical causes) learned by the task branch to dominate.

Failure Analysis. DLIR still faces challenges when concepts exhibit significant cultural divergence not fully captured by the GMM. For example, aligning the French concept of *Laïcité* (a specific form of state secularism) with English concepts of *Secularism* remains difficult. DLIR sometimes fails to prioritize the nuance of strict state–religion separation inherent in the French term, indicating that capturing highly specific cultural–political constructs may require modeling more complex perturbations than our current GMM approach allows.

5 Conclusion

This paper introduces DLIR, a novel framework for aligning nuanced sociological concepts cross-lingually, directly addressing the challenge of semantic heterogeneity prevalent in longitudinal sociological studies. Such studies, including national surveys from diverse institutions and temporal periods, often see identical concepts manifesting through varied expressions, creating significant barriers to robust longitudinal and cross-sectional analyses. Our work is particularly pertinent given that longitudinal population and cross-sectional studies form the backbone of empirical research in the social, economic, and behavioural sciences, as well as in epidemiology and health research, providing the basis for evidence-based policy advice.

The architecture of DLIR, grounded in the Vector Decomposition Hypothesis, effectively disentangles domain-specific semantics from language-specific features. By modeling language perturbations with Gaussian Mixture Models (GMMs) on spherical geometry using minimal parallel data, it learns language-invariant representations. This is crucial for overcoming the scarcity of manually tagged multilingual concepts, especially in low-resource languages, which currently renders many cross-national study comparisons unviable.

Our experiments demonstrate DLIR's state-of-the-art performance in cross-lingual sociological concept retrieval and its remarkable ability for knowledge transfer to new languages, even outperforming models fine-tuned directly on target language data. This suggests that DLIR can be a significant step change in how natural language processing techniques are applied to comparative social science. By enabling more reliable alignment of sociological constructs across linguistic and cultural divides, DLIR offers a promising, parameter-efficient approach that can serve as an enabler for the automated harmonisation of longitudinal surveys.

While acknowledging limitations such as GMM fitting with sparse data and reliance on synthetic text for core training, DLIR significantly advances cross-lingual understanding in culturally rich fields like sociology. Future work will focus on enhancing robustness and extending the framework to other specialized domains, further paving the way for more reliable and scalable multilingual NLP applications in interdisciplinary research contexts. In essence, DLIR contributes to the recontextualization of natural language processing by providing tools that can directly support and enhance established methodologies within the social sciences, facilitating deeper and more comparable cross-cultural insights.

Limitations

While DLIR demonstrates promising results, several limitations should be acknowledged:

- **GMM Fitting Dependency and Robustness:** The effectiveness of the invariance branch relies on the quality of the GMMs fitted on minimal parallel data (§3.3). While our results show efficacy even with as few as 73 pairs (for en-is, see Table 7), the stability and representativeness of GMMs learned from such sparse data, especially for capturing complex language divergences, might be limited. This could explain some performance variations across languages (Appendix Table 3, 4), particularly for the lowest-resource pairs, and warrants further investigation into more robust density estimation techniques for extremely low-data scenarios. The selection of GMM

components (K) also introduces a hyperparameter.

- **Reliance on Synthetic Data:** Both the core training ($\mathcal{L}_{\text{task}}$) and the primary evaluation tasks heavily rely on synthetic descriptive texts generated by DeepSeek-V3 (§4.1). While designed for diversity and implicit referencing, this data may not fully capture the nuances, complexities, or potential biases present in authentic sociological discourse. Performance on naturally occurring sociological texts might differ.

- **Translation Quality for Baselines:** The strong LoRA-FR/DE baselines in Task 2 were trained on machine-translated data (ByteDance service). While standard practice, the quality of these translations could potentially influence the upper bound of their performance compared to DLIR trained solely on original English data.

- **Scope of Evaluation and Analysis:** Our evaluation covers 10 European languages, primarily dictated by the coverage of ELSST. It does not include typologically distant languages (e.g., East Asian languages). Furthermore, it focuses on Target-to-English transfer rather than direct transfer between non-English pairs. The analysis of cross-language performance variation (§4.2) is preliminary and doesn't deeply correlate results with factors like linguistic typology or concept ambiguity. A comprehensive qualitative error analysis across all languages was beyond the scope of this work but would be valuable.

References

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Ning Ding, Xingtai Lv, Qiaosen Wang, Yulin Chen, Bowen Zhou, Zhiyuan Liu, and Maosong Sun. 2023. Sparse low-rank adaptation of pre-trained language models. *Preprint*, arXiv:2311.11696.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Xinshuo Hu, Zifei Shan, Xinping Zhao, Zetian Sun, Zhenyu Liu, Dongfang Li, Shaolin Ye, Xinyuan Wei, Qian Chen, Baotian Hu, Haofen Wang, Jun Yu, and Min Zhang. 2025. Kalm-embedding: Superior training data brings a stronger embedding model. *Preprint*, arXiv:2501.01028.

Pratik Jawanpuria, Arjun Balgovind, Anoop Kunchukuttan, and Bamdev Mishra. 2019. Learning multilingual word embeddings in latent metric space: A geometric approach. *Transactions of the Association for Computational Linguistics*, 7:107–120.

Dawid J. Kopiczko, Tijmen Blankevoort, and Yuki M. Asano. 2024. Vera: Vector-based random matrix adaptation. *Preprint*, arXiv:2310.11454.

Saurabh Kulshreshtha, Jose Luis Redondo Garcia, and Ching-Yun Chang. 2020. Cross-lingual alignment methods for multilingual BERT: A comparative study. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 933–942, Online. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. Nv-embed: Improved techniques for training llms as generalist embedding models. *Preprint*, arXiv:2405.17428.

Jiahuan Li, Shujian Huang, Aarron Ching, Xinyu Dai, and Jiajun Chen. 2024. PreAlign: Boosting cross-lingual transfer by early establishment of multilingual alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10246–10257, Miami, Florida, USA. Association for Computational Linguistics.

Wing Yan Li, Zeqiang Wang, Jon Johnson, and Suparna De. 2025. Are information retrieval approaches good at harmonising longitudinal survey questions in social science? *Preprint*, arXiv:2504.20679.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. Dora: Weight-decomposed low-rank adaptation. *Preprint*, arXiv:2402.09353.

Nattapong Tiyajamorn, Tomoyuki Kajiwara, Yuki Arase, and Makoto Onizuka. 2021. Language-agnostic representation from multilingual sentence encoders for cross-lingual similarity estimation. In

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7764–7774, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lifu Tu, Jin Qu, Semih Yavuz, Shafiq Joty, Wenhao Liu, Caiming Xiong, and Yingbo Zhou. 2024. Efficiently aligned cross-lingual transfer learning for conversational tasks using prompt-tuning. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1278–1294, St. Julian's, Malta. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *Preprint*, arXiv:2402.05672.

Yifan Yang, Jiajun Zhou, Ngai Wong, and Zheng Zhang. 2024. LoRETTA: Low-rank economic tensor-train adaptation for ultra-low-parameter fine-tuning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3161–3176, Mexico City, Mexico. Association for Computational Linguistics.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *Preprint*, arXiv:2303.10512.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *Preprint*, arXiv:2402.18913.

A Appendix

A.1 Performance

Table 3: Ablation study results on the cross-lingual concept retrieval task. The table compares the baseline E5-small (E5s), standard LoRA fine-tuning (LoRA), dual-branch LoRA with generic Gaussian noise (Gaussian), and our proposed method (DLIR). Languages with fewer than 50 concept test samples (en-is: 9, en-cs: 16, en-sl: 45) have been excluded. Best results within each row are in **bold**, second best are underlined. Group and overall averages are recalculated based on the 10 languages shown.

	Languages	E5s			LoRA			Gaussian			DLIR		
		R@1	MRR	NDCG	R@1	MRR	NDCG	R@1	MRR	NDCG	R@1	MRR	NDCG
Romance	Spanish	0.779	0.885	0.915	0.900	0.946	0.960	**0.914**	**0.954**	**0.966**	0.900	0.948	0.962
	French	0.752	0.869	0.903	0.904	0.946	0.960	0.888	0.939	0.955	**0.912**	**0.952**	**0.964**
	Romanian	0.743	0.861	0.897	**0.886**	**0.940**	**0.956**	0.871	0.925	0.944	**0.886**	0.937	0.953
Germanic	Dutch	0.594	0.768	0.827	**0.820**	**0.904**	**0.928**	**0.820**	0.905	0.930	0.797	0.895	0.923
	German	0.592	0.775	0.833	0.782	0.882	0.912	**0.838**	**0.916**	**0.938**	0.810	0.900	0.926
	Norwegian	0.573	0.751	0.814	0.750	0.862	0.897	0.823	0.903	0.928	**0.831**	**0.906**	**0.930**
	Swedish	0.627	0.798	0.850	0.761	0.868	0.902	0.799	0.888	0.917	**0.806**	**0.897**	**0.924**
Other	Greek	0.717	0.836	0.878	0.717	0.846	0.886	0.772	0.876	0.908	**0.804**	**0.894**	**0.921**
	Finnish	0.566	0.747	0.812	0.614	0.775	0.832	0.627	0.791	0.844	**0.711**	**0.830**	**0.873**
	Lithuanian	0.529	0.736	0.804	0.686	0.824	0.869	0.700	0.839	0.880	**0.750**	**0.863**	**0.898**
Romance Languages		0.758	0.872	0.905	0.897	0.944	0.959	0.891	0.939	0.955	**0.899**	**0.946**	**0.960**
Germanic Languages		0.597	0.773	0.831	0.778	0.879	0.910	**0.820**	**0.903**	**0.928**	0.811	0.900	0.926
Other Languages		0.604	0.773	0.831	0.672	0.815	0.862	0.700	0.835	0.877	**0.755**	**0.862**	**0.897**
Overall Average		0.647	0.803	0.853	0.782	0.879	0.910	0.805	0.893	0.921	**0.821**	**0.902**	**0.928**

Table 4: Comparison of DLIR with state-of-the-art multilingual embedding models. The table includes results for 10 languages, excluding those with fewer than 50 concept test samples (en-is: 9, en-cs: 16, en-sl: 45). Best results within each row are in **bold**, second best are underlined. Group and overall averages are recalculated based on the 10 languages shown.

	Languages	DLIR			KaLM			OpenAI			E5l		
		R@1	MRR	NDCG	R@1	MRR	NDCG	R@1	MRR	NDCG	R@1	MRR	NDCG
Romance	Spanish	**0.900**	**0.948**	**0.962**	0.864	0.931	0.949	0.836	0.914	0.936	0.857	0.926	0.945
	French	**0.912**	**0.952**	**0.964**	0.856	0.923	0.943	0.824	0.905	0.930	0.832	0.907	0.931
	Romanian	**0.886**	**0.937**	**0.953**	0.886	0.940	0.956	0.771	0.879	0.910	0.829	0.912	0.935
Germanic	Dutch	**0.797**	**0.895**	**0.923**	0.767	0.872	0.905	0.759	0.872	0.905	0.782	0.882	0.913
	German	**0.810**	**0.900**	**0.926**	0.775	0.879	0.910	0.761	0.873	0.906	0.782	0.883	0.913
	Norwegian	**0.831**	**0.906**	**0.930**	0.718	0.841	0.882	0.685	0.827	0.871	0.750	0.870	0.904
	Swedish	**0.806**	**0.897**	**0.924**	0.769	0.869	0.902	0.709	0.838	0.880	0.791	0.892	0.920
Other	Greek	**0.804**	**0.894**	**0.921**	0.337	0.624	0.720	0.413	0.668	0.753	0.750	0.871	0.905
	Finnish	0.711	0.830	0.873	0.590	0.770	0.829	0.590	0.767	0.827	**0.783**	**0.882**	**0.912**
	Lithuanian	0.750	0.863	0.898	0.621	0.785	0.840	0.600	0.769	0.828	**0.786**	**0.880**	**0.911**
Romance Languages		**0.899**	**0.946**	**0.960**	0.850	0.921	0.941	0.810	0.899	0.925	0.839	0.915	0.937
Germanic Languages		**0.811**	**0.900**	**0.926**	0.757	0.865	0.900	0.729	0.853	0.891	0.776	0.882	0.913
Other Languages		0.755	0.862	0.897	0.516	0.726	0.796	0.534	0.735	0.803	**0.773**	**0.878**	**0.909**
Overall Average		**0.821**	**0.902**	**0.928**	0.713	0.840	0.881	0.695	0.831	0.875	0.794	0.891	0.919

A.2 Dataset Details

Table 5: Relation Type and Length Statistics in the Training and Development Sets

Relation Type	#Train	Avg. Words (Train)	#Dev	Avg. Words (Dev)
self	6,236	338.21	1,556	336.12
related	5,264	333.92	1,340	334.29
broader	5,154	336.79	1,332	341.28
narrower	2,046	332.46	440	331.72

Table 6: Word Count Statistics for the Multilingual Test Set

Text Field	#Samples	Avg.	Median	Min	Max
Source Concept	1,253	1.94	2.16	1.51	2.69
Source Definition	1,253	15.58	15.15	10.19	18.61
Correct Candidate	1,253	333.25	332.69	328.31	344.33
Incorrect Candidates	2,601	350.90	350.81	341.44	354.46

Table 7: Parallel Corpus Statistics Across Language Pairs

Lang. Pair	#Sent.	Tokens (EN)	Tokens (XX)	Vocab (EN)	Vocab (XX)
en–cs	118	2,181	1,951	783	977
en–de	828	15,147	15,616	3,006	4,080
en–el	582	10,294	11,271	2,294	3,173
en–es	830	15,381	17,637	3,041	3,329
en–fi	503	9,609	6,418	2,249	3,036
en–fr	768	14,066	17,150	2,856	3,413
en–is	73	1,277	1,169	501	520
en–lt	843	15,453	13,415	3,046	4,658
en–nl	807	14,922	15,232	2,997	3,341
en–no	803	14,731	13,284	2,950	3,394
en–ro	511	9,548	11,087	2,265	3,193
en–sl	193	3,367	2,986	1,087	1,374
en–sv	812	14,831	13,965	2,982	3,558

A.3 Synthetic Data Generation and Human Validation

To train and evaluate DLIR on nuanced sociological concepts, we required a dataset of descriptive texts that accurately reflect the ELSST hierarchy while exhibiting linguistic diversity. This appendix details the pipeline used to generate this synthetic dataset using DeepSeek-V3 and the human validation study conducted to ensure its quality.

A.3.1 A.3.1 Synthetic Data Generation Pipeline

Objective. The primary goal of the generation pipeline was to create varied and context-rich descriptions for each concept. Crucially, the pipeline was designed to generate **implicit references**—where the text describes the concept without explicitly naming it. This forces the model to learn deep semantic representations rather than relying on superficial keyword matching.

Generation Strategy. We employed a highly structured, template-based prompting strategy defined in a YAML configuration. To ensure diversity, we systematically varied the prompts across several dimensions:

- **Roles and Perspectives:** We defined 16 distinct roles (see Table 8), categorized into Professional Sociology Roles and Social Practice Roles. Each role was associated with specific focus areas (e.g., "power relations" for a critical theorist) and linguistic styles (e.g., "technical", "practical").

- **Templates:** A combination of role-specific templates (70% probability) and common templates (30% probability) were used to structure the prompt, guiding the LLM on how to integrate the role, focus, and style.
- **Instruction and Implicitness:** A core instruction block was prepended to every prompt, explicitly forbidding the LLM from mentioning the concept name or using obvious referring terms (e.g., "this concept").

Table 8: Roles utilized in the prompt generation strategy to ensure diverse perspectives.

Professional Sociology Roles	**Social Practice Roles**
Theoretical Sociologist	Social Worker
Empirical Social Researcher	Social Studies Educator
Applied Sociologist	Community Organizer
Cultural Sociologist	Social Affairs Journalist
Social Policy Specialist	Social Impact Consultant
Critical Social Theorist	Sociology Student
	Public Administrator
	Healthcare Professional
	NGO Practitioner
	Engaged Citizen

Pipeline Implementation. For each concept and its ELSST definition, the pipeline randomly selected a unique role and a template. The system prompt established the LLM's persona, and the user prompt combined the core instruction with the specific analysis task. We generated up to 24 unique descriptions per concept. An example of a constructed prompt is shown in Figure 3.

System Prompt:
You are a critical social theorist. Your task is to analyze a sociological concept without explicitly naming it.

User Prompt (Example):
Based on the concept "Social Stratification" and its definition "The classification of persons into groups based on shared socio-economic conditions; a relational set of inequalities...", generate an analysis text.

Important requirements:
1. DO NOT mention the concept name "Social Stratification" directly in your response.
2. DO NOT use obvious referring terms like "this concept", "this theory", etc.
3. Write your description so readers can understand what concept you're discussing without naming it.
4. Use plain text format, not Markdown.
5. Ensure you analyze from the perspective of your assigned role.

Task: As a critical social theorist, examine relationships to power structures and social inequality. Analyze power relations, social inequalities, systemic critique, transformative potential using critical and transformative analysis, citing critical analyses and social critiques to deepen understanding.

Figure 3: Example of a constructed prompt for the concept "Social Stratification" under the role of "Critical Social Theorist". The prompt combines role assignment, strict constraints on implicit expression, and specific analytical focus.

A.3.2 Human Validation Study

To ensure the reliability of the synthetic dataset, we conducted a rigorous human evaluation study with domain experts.

Objective and Scope. The evaluation focused specifically on the quality of the "self" relation examples. These examples are critical as they represent different expressions of the core concept definition and form the foundation of the model's semantic understanding.

Methodology. We randomly sampled **40** concept-text pairs from the generated dataset. These pairs were evaluated by **[2]** independent annotators with expertise in sociology. The evaluators were provided with the concept name and its official ELSST definition (serving as the ground truth). They assessed each generated text based on two criteria: Accuracy and Fluency.

Evaluation Criteria. The criteria were defined as follows:

- **Accuracy (1-5 Scale):** Measures how accurately and unambiguously the generated text represents the core definition of the concept.
  - **5 (Highly Accurate):** Perfect, clear application or explanation.
  - **4 (Basically Accurate):** Consistent core idea, minor extraneous info or suboptimal phrasing.
  - **3 (Partially Accurate):** Related, but misses the core point or shows concept confusion.
  - **2 (Misleading):** Appears related but misrepresents the definition.
  - **1 (Completely Wrong):** Irrelevant or factually incorrect.
- **Fluency (1-3 Scale):** Measures the linguistic quality of the text, independent of accuracy.
  - **3 (Excellent):** Natural, grammatically correct.
  - **2 (Acceptable):** Generally fluent, minor errors or signs of machine generation.
  - **1 (Poor):** Incoherent, difficult to understand.

Results. The results of the human validation study are summarized in Table 9. The dataset achieved a high average accuracy score of 3.950 (out of 5) and an average fluency score of 2.712 (out of 3). The inter-annotator agreement (IAA), measured using Krippendorff's Alpha (Ordinal), was 0.85 for Accuracy and 0.72 for Fluency, indicating substantial to high agreement among the experts. These results confirm that the synthetic data generation pipeline successfully produced high-quality, nuanced, and accurate representations of the sociological concepts, suitable for training our DLIR framework.

Table 9: Summary of Human Validation Results for Synthetic Data Quality (N=40).

Metric	Mean	Median	Std. Dev.	Scale
Accuracy	3.950	4.0	0.967	1-5
Fluency	2.712	3.0	0.482	1-3
IAA: 0.85 (Accuracy), 0.72 (Fluency)				