Discriminating Form and Meaning in Multilingual Models with Minimal-Pair ABX Tasks

Anonymous ACL submission

Abstract

We introduce a set of training-free ABX-style discrimination tasks to evaluate how multilingual language models represent language identity (form) and semantic content (meaning). Inspired from speech processing, these zero-shot tasks measure whether minimal differences in representation can be reliably detected. This offers a flexible and interpretable alternative to probing. Applied to XLM-R (Conneau et al., 2020a) across pretraining checkpoints and layers, we find that language discrimination declines over training and becomes concentrated in lower layers, while meaning discrimination strengthens over time and stabilizes in deeper layers. We then explore probing tasks, showing some alignment between our metrics and linguistic learning performance. Our results position ABX tasks as a lightweight framework for analyzing the structure of multilingual representations.

1 Introduction

Multilingual Transformer models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020a) have become essential tools for cross-lingual NLP. Trained on large concatenated corpora spanning dozens of languages, these models learn representations that support transfer across languages even in the absence of explicit cross-lingual supervision (Wu and Dredze, 2019; Conneau et al., 2020b; Xue et al., 2021; Philippy et al., 2023, etc.). Despite their success, it remains unclear how these models internally organize linguistic form and shared meaning. Prior work suggests that multilingual models encode both language-specific information (e.g., surface forms, word order) and language-agnostic features (e.g., semantic content), but the nature and interaction of these representations is not fully understood. These encoding choices shape generalization and transfer behavior, including both positive effects (e.g., shared structure benefiting low-resource languages) and negative ones, such as the *curse of multilinguality* (Conneau et al., 2020a), where performance degrades due to interference across languages.

Understanding how form and meaning are represented, and how this balance evolves during pretraining, is essential to explain and improve cross-lingual transfer. If a model strongly encodes language identity, it may better avoid interference between closely related languages[1]. Conversely, if it aligns meanings across languages, it may support more effective semantic generalization. To explore this balance, we ask: *How are languages represented at the form level? How well do models encode shared meanings? And how do these properties evolve across training?*

Previous work has often relied on probing tasks to investigate such questions. While useful, probing requires training classifiers on top of frozen representations, and results are sensitive to probe design and task setup (Belinkov, 2022; Hewitt and Liang, 2019; Voita and Titov, 2020). This makes it difficult to isolate what is truly encoded by the model versus what is learnable with supervision.

We propose a zero-shot alternative: ABX-style discrimination tasks that directly measure model representation structure without additional training. Originating in speech processing (Schatz et al., 2013), ABX tasks evaluate whether a model reliably discriminate minimal contrasts: given a triplet (A, B, X), is X closer to A or B? By designing minimal pairs that differ only in language or in meaning, we isolate and quantify how well models distinguishes these dimensions. Because they are contrastive, zero-shot, and training-free, these metrics can be applied across languages, checkpoints, and architectures with minimal adaptation.

[1] In fact, it was found that forcing some sort of separation in multilingual models can help somewhat alleviate these negative interferences (Pfeiffer et al., 2022; Blevins et al., 2024; Xu et al., 2024; Huang et al., 2024).

Our contributions are as follows:

1. We propose a training-free, ABX-style framework for analyzing multilingual representations by contrasting minimal pairs. Our tasks are designed to isolate language identity (form) and semantic content (meaning), offering a complementary alternative to traditional probing methods.
2. We apply this framework to XLM-R across 36 languages and 630 language pairs, analyzing all pretraining checkpoints and layers. We show that language and meaning discrimination evolve in parallel but are not mutually exclusive: different layers vary in the degree to which they encode each axis.
3. We relate ABX discrimination scores to downstream performance on POS tagging, NER, and NLI. We find that form-oriented tasks correlate more strongly with language discrimination, while NLI, a semantic task, shows no consistent relationship to either axis, highlighting a disconnect between task performance and intrinsic representational structure.

2 Related Work

Multilingual language models are expected to support cross-lingual generalization by encoding both language-specific form and shared semantic content. However, existing evaluation methods typically focus on one of these dimensions in isolation. This section reviews prior work on analyzing multilingual representations and highlight the need for a unified, training-free framework that jointly evaluates both language identity and meaning in a controlled, contrastive setting.

2.1 Evaluating Form and Content in Multilingual Representations

Multilingual pretrained language models aim to map diverse languages into a shared embedding space. This allows for zero-shot and cross-lingual transfer, but raises the question of how these models balance language-specific and language-agnostic features during training.

Content-focused evaluations typically focus on cross-lingual alignment, using methods such as translation retrieval to measure whether semantically equivalent inputs in different languages are mapped to nearby embeddings (Sundar et al., 2025; Pires et al., 2019; Libovický et al., 2020; Hu et al., 2020). Models like LASER (Artetxe and Schwenk, 2019) and LaBSE (Feng et al., 2022) are explicitly trained to optimize such alignment. More recent work introduces contrastive alignment scores such as DALI (Ravisankar et al., 2025) to better capture meaning equivalence. However, these approaches abstract away from language identity and provide little insight into how models handle surface-form distinctions across languages.

Conversely, form-focused evaluations examine how well a model encodes language identity. Clustering analyses show that multilingual embeddings often group by language or script, particularly in lower layers (Libovický et al., 2020; Choenni and Shutova, 2022). Classifiers trained on frozen representations can often identify input language with high accuracy (Choenni and Shutova, 2022), but this depends on probe training and may not reflect the geometry of the representation space itself.

These two evaluation paradigms have remained largely separate. To our knowledge, no existing method allows for simultaneous, controlled evaluation of both dimensions without relying on task-specific training. As a result, we lack a unified evaluation framework that can directly assess both dimensions under comparable, controlled conditions.

2.2 Training Dynamics and Linguistic Emergence in Multilingual Models

Several studies have examined how multilingual representations evolve during pretraining. Blevins et al. (2022) tracked the emergence of linguistic knowledge in XLM-R (Conneau et al., 2020a), showing that different properties emerge at different layers and stages, and that the best-performing checkpoint varies across languages and tasks. Other studies have shown that multilingual models sometimes internally pivot through high-resource languages like English when processing low-resource inputs (Wendler et al., 2024; Schut et al., 2025), while other research suggests that these models juggle both language-specific and language-neutral features (Tang et al., 2024; Libovický et al., 2020; Tanti et al., 2021) These works highlight the complex interplay between form and content in multilingual models and how this balance shifts over time. However, they again mainly rely on task-specific probes or downstream evaluations, which do not offer a way to disentangle form and content in a direct, unsupervised way.

2.3 Prior Uses of ABX Evaluation

The ABX framework offers a contrastive, classifier-free means of evaluating representational structure in a controlled, unsupervised setting. Originally developed in speech processing and psycholinguistics (Schatz et al., 2013, 2014; Schatz, 2016), ABX tests ask whether a test item X is more similar (in embedding space) to a reference item A or to an alternative B. By controlling the design of A, B, and X, ABX evaluations can isolate specific factors of interest (such as phoneme identity in speech) while holding others constant (e.g., speaker, context) (Versteegh et al., 2015; Dunbar et al., 2017; Hallap et al., 2022; Sicherman and Adi, 2023, etc.). ABX tasks have proven robust to variability from other categorical structures, enabling reliable measurement of the target factor even when other linguistic or speaker-related properties vary (Schatz, 2016).

While ABX has primarily been applied to phoneme discrimination, recent work has begun adapting it to the other tasks, testing models' ability to discriminate between languages (Carbajal et al., 2016; de Seyssel and Dupoux, 2020; de Seyssel, 2023), speakers (Thorburn et al., 2019; de Seyssel et al., 2022) and to evaluate syntactic or semantic distinctions (Algayres et al., 2022, 2023).

Our work builds on this foundation by adapting ABX discrimination to text-based multilingual models. We propose a set of zero-shot tasks that independently measure sensitivity to language identity and semantic content using minimal contrast triplets. To our knowledge, this is the first unified, training-free framework that systematically isolates and evaluates these two core dimensions of multilingual representation.

3 Our ABX Discrimination Framework

Understanding how multilingual models structure linguistic information in their internal representations is key to explaining their interactions between different languages, and to a further extent their generalization behaviors. To directly assess the intrinsic structure of multilingual representations without relying on the pitfalls of extrinsic evaluation, we adapt the ABX discrimination paradigm, originally developed for evaluating speech embeddings, to the text domain.

In the original ABX framework (Schatz et al., 2013, 2014; Schatz, 2016), illustrated in Figure 1, three items (A, B, X) are presented, with A and B belonging to different categories, and X matching the category of either A or B. A model is successful when X is closer (according to a distance metric in embedding space) to the item that shares its category. That is, for each triplet, a correct decision is recorded when $d(X, A) < d(X, B)$, with X and A sharing the same category. The score for a given triplet is computed as :

$$\text{score}(A, B, X) = \mathbb{1}[\, d(X, A) < d(X, B)\,]$$

where $\mathbb{1}$ denotes the indicator function. The overall ABX score is the average success rate across all triplets. Importantly, control variables can be introduced to eliminate bias from confounding factors. In that case, both A and B share the same control variable to ensure that the discrimination is based solely on the variable of interest.

The ABX score reflects the proportion of correct decisions, with higher scores indicating better discrimination. We apply this setup to sentence embeddings extracted from XLM-R at various layers and checkpoints, where each sentence is represented by the mean-pooled embedding of its subword tokens. Cosine similarity is used as the distance metric[2].

We propose two ABX variants for studying multilingual language models: *language discrimination* (LD) and *meaning discrimination* (MD). Both tasks leverage paired multilingual data and are constructed to isolate either language identity or semantic content while controlling for the other. These tasks enable zero-shot, training-free evaluation of key representational properties in multilingual models.

We present both tasks below; see Appendix B for further illustrations and examples.

Language Discrimination In the LD task, the objective is to assess whether the model can distinguish between embedding representations from different languages while controlling for meaning. In other words, the focus is on determining whether the form of the language is encoded in the representations sufficiently to discriminate between

[2]We choose cosine as it is the standard metric in many embedding-based evaluations, particularly in multilingual sentence retrieval and alignment tasks (e.g., Ravisankar et al., 2025; Sundar et al., 2025; Mohammadshahi et al., 2019). Cosine is well-suited to measuring relative orientation in high-dimensional spaces and is less sensitive to differences in embedding magnitude, which makes it particularly effective for comparing representations across languages and layers.

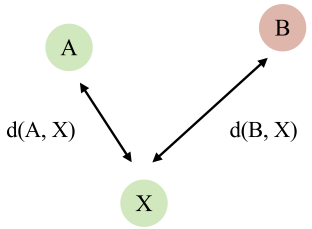


Figure 1: Illustration of the ABX discrimination task. A and X share the target variable, whereas B differs. Control variables may be included, with A and B sharing the same control variable.

languages. Triplets are constructed as follows: X comes from language L_1 and carries meaning M_1; A, the target, is also from language L_1 but conveys a different meaning M_2; and B, the distractor, is from another language L_2 but shares the same meaning M_2 as A, hence controlling for meaning (see Appendix B for an illustration). The task is considered successful if the model leads to the distance between A and X to be smaller than that between B and X.

Meaning Discrimination In the MD task, we test whether the model captures differences in meaning while holding language constant. The goal is to evaluate whether semantic content is encoded in the representations independently of surface form. Triplets are constructed such that A and X share the same meaning M_1 but come from different languages (L_1 and L_2), while B is in the same language as A (L_2) but conveys a different meaning M_2. The model is considered successful if it places X closer to A than to B, indicating that it encodes semantic similarity across languages, beyond surface-level language identity.

While LD primarily probes the presence of language-specific information, MD offers a more direct lens on semantic similarity. High MD scores, especially across languages, suggest that the model encodes meaning in a way that is at least partially language-agnostic. As such, MD may serve as a proxy for cross-lingual semantic alignment within the representation space. In fact, we show in Section 4.1 that a standard cross-lingual retrieval task, commonly used to assess such alignment, correlates highly with our MD task, supporting the idea that MD captures cross-lingual alignment[3]. We do not perform a similar analysis for LD, as no existing metric captures the specific abilities assessed by our language ABX task.

4 Discrimination dynamics in a multilingual models

4.1 General Experimental setup

Model To study pretraining dynamics in a multilingual setting, we use the base version of XLM-R (Conneau et al., 2020a) (L = 12, H = 768, A = 12, 270M parameters), a widely used multilingual masked language model. Specifically, we rely on the checkpoints released by Blevins et al. (2022), who retrained XLM-R from scratch in order to examine the evolution of language representations during pretraining[4]. All evaluations and analyses in this work are based on the representations from these checkpoints.

ABX Languages and Dataset We construct ABX triplets and perform evaluations using the WMT24++ dataset (Deutsch et al., 2025), a multilingual corpus of 55 languages with sentence-level alignments across all language pairs. From this corpus, we select 36 languages spanning a broad range of families, scripts, and typological features (see Appendix A for the complete list). This selection yields 630 unordered language pairs. Triplets are sampled randomly from aligned sentence pairs, ensuring that each triplet satisfies the relevant ABX condition (form or meaning), and that the sample size is sufficient to ensure broad and unbiased coverage. For each evaluation mode and language pair, we generate approximately 100,000 triplets. Unless stated otherwise, we report discrimination scores averaged across all layers. In most analyses, we present scores separately by checkpoint to track how discrimination abilities evolve during training. In addition to language-pair scores, we compute a global LD or MD score for each language, defined as the average across all pairings with the other 35 languages. These global metrics offer a higher-level view of how well a language is discriminated or semantically aligned within the multilingual space.

Validation of ABX Metrics To validate our metrics, we perform two control analyses. First, we

[3]The high correlation does not imply they are identical. Our ABX MD task targets the same underlying ability under more controlled conditions. Instead of ranking many candidates, MD ABX uses contrastive triplets that isolate semantic differences while tightly controlling for language .

[4]Details of the pretraining scheme can be found in Blevins et al. (2022)

confirm that both ABX scores return values at or near chance (0.5) under a "baseline" setup, where the variable of interest (language for LD, meaning for MD) is held constant across all three elements of the triplet. This serves as a sanity check to rule out bias in the construction of the triplets or evaluation procedure. Second, we compare MD scores at the final checkpoint with performance on a standard cross-lingual retrieval task. Following the setup of Sundar et al. (2025), we compute, for each language pair (L_1, L_2), the top-1 accuracy of retrieving the most semantically similar sentence in language L_2 given a sentence in language L_1. The retrieval pool consists of L_2 candidates from the WMT24 dataset (and vice versa). While retrieval tasks typically rely on mean-pooled representations from the final layer, our ABX evaluations average scores across all layers. Despite this difference, we find a strong correlation between the two metrics (Pearson $r = 0.77$)[5]. This supports the validity of ABX as a proxy for semantic alignment. Importantly, ABX goes further by explicitly controlling for surface form (in the case of MD), enabling a more fine-grained assessment of the model's semantic representations.

4.2 Experiments

We begin by analyzing how the model's ability to discriminate between language identity (form) and semantic meaning (content) evolves during training. Figures 2, 3, and 4 present complementary views of these dynamics across checkpoints and layers.

Checkpoint-level evolution. Figure 2 shows the evolution of average LD and MD scores across checkpoints, aggregated over all language pairs. First of all, we can see that all scores are consistently above the 0.5 baseline, meaning that the model, at all checkpoints, can discriminate between languages and meanings (cross-lingually) to some extent. LD score declines rapidly during early training steps and gradually recovers in later stages, while MD score steadily improves. This suggests that as training progresses, the model increasingly prioritizes semantic abstraction over explicit language-specific cues.

We also observe a negative correlation between the two measures when considering all language pairs across checkpoints (Spearman's $\rho = -0.74$, $p < 0.001$; Pearson's $r = -0.68$, $p < 0.001$), computed over individual (language pair × checkpoint) points. We also ensure that this correlation is not merely driven by training dynamics by examining the final checkpoint (step 150,000) in isolation. The relationship remains strong (Spearman's $\rho = -0.83$, $p < 0.001$; Pearson's $r = -0.72$, $p < 0.001$), confirming that language pairs which are more separable by form tend to exhibit lower meaning preservation, even in the fully trained model.

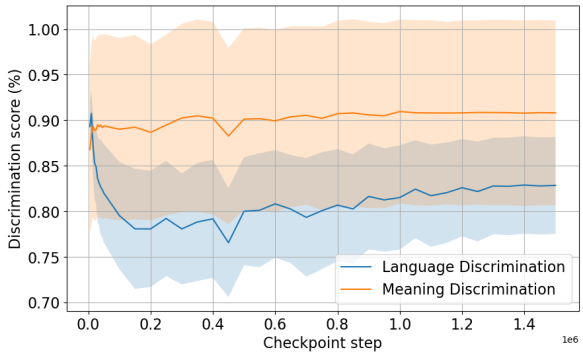


Figure 2: Language and meaning ABX discrimination scores across checkpoints (averaged over layers and all language pairs). Baseline score is 0.5.

Layer-level patterns. To better understand how these abilities are distributed within the model, Figure 3 plots discrimination scores across layers for the final checkpoint. LD is strongest in the lower layers and gradually decreases with depth, reaching a plateau in the upper layers before rising again in the final layer. In contrast, MD starts lower but quickly rises and remains high in the upper layers, but decreases slightly in the last layer. This pattern suggests that earlier layers focus more on identifying the language of the input, while later layers capture its meaning more effectively.

We also find a significant negative correlation between language and meaning discrimination across layers (Spearman's $\rho = -0.66$, $p < 0.001$; Pearson's $r = -0.53$, $p < 0.001$). This indicates that, as representations evolve through the network, increases in meaning discrimination are generally accompanied by decreases in language separability. However, the correlation coefficients are weaker than those observed across language pairs, suggesting that this trade-off is not strictly enforced at the layer level. Instead, the model exhibits a more flexible allocation of representational capacity across form and meaning over its depth.

[5]When both use last-layer embeddings, $r = 0.73$; when comparing last-layer retrieval to all-layer ABX, Pearson drops to $r = 0.53$, but remains significant ($p < 0.001$).

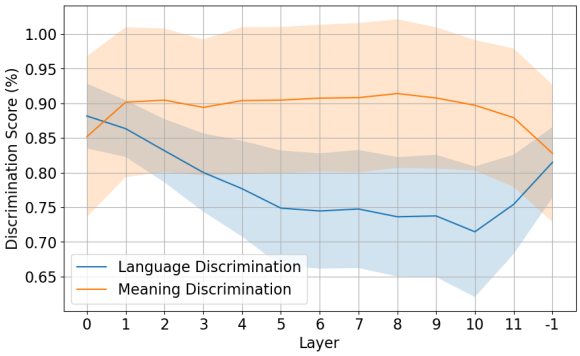


Figure 3: Language and meaning ABX discrimination scores across layers (averaged over all language pairs) for the last checkpoint (step 150 000)

Joint checkpoint and layer dynamics. Figure 4 presents ABX discrimination scores as a function of both checkpoint and layer. LD (left panel) is initially high across most layers but gradually becomes concentrated in the lower layers and the output layer as training advances. In contrast, MD (right panel) improves steadily across all layers, especially in the deeper ones.

Taken together, these patterns suggest that the model initially relies heavily on language-specific features but gradually shifts toward encoding more abstract, language-invariant semantic structures. Importantly, the two forms of discrimination are not strictly opposed at the layer level. While a trade-off exists, its moderate strength suggests that the model can support both language sensitivity and semantic alignment to some degree simultaneously.

We provide an additional analysis in Appendix D, showing how both discrimination scores vary across individual languages and training checkpoints.

4.3 Discussion

These findings support the view that pretraining leads to a progressive decoupling of surface form and semantic content. Early in training, language identity is clearly encoded across the model. As training proceeds, this information becomes increasingly concentrated in the lower layers, while deeper layers develop language-invariant semantic representations. This aligns with prior work suggesting that lower layers encode form-related properties, while higher layers abstract away toward more conceptual information (Pires et al., 2019; Tenney et al., 2019). Notably, at convergence, several middle layers appear to support both types of discrimination to a moderate degree, suggesting a partial overlap between structural and semantic signals rather than strict exclusivity.

5 Correlation of ABX discrimination metrics with linguistic learning

We then examine whether the model's discrimination patterns relate to linguistic task performance, focusing on monolingual probing and cross-lingual transfer.

5.1 Experimental Setup

Following Blevins et al. (2022), we evaluate both monolingual probing and cross-lingual transfer to test how our ABX discrimination metrics relate to linguistic generalization. We use part-of-speech tagging (POS), named entity recognition (NER), and natural language inference (NLI) as representative tasks. POS and NLI were used in the original analysis; we additionally include NER, which offers a complementary view of lexical-level information and the form–content divide. These tasks span different linguistic levels, from surface form to sentence-level semantics.

All probes are trained independently per language with early stopping on validation loss. We run 6 iterations per setup with different random seeds and report average results. Unless otherwise specified, all experiments use the final XLM-R checkpoint (step 150,000).

Part-of-Speech Tagging (POS) We use Universal Dependencies (UD) (Nivre et al., 2020). Monolingual performance is evaluated on all 36 languages from our ABX setup, using standard UD splits. For cross-lingual transfer, we follow Blevins et al. (2022) and use the Parallel UD (PUD) subset at test time, covering 18 languages (Appendix A).

Named Entity Recognition (NER) We use WikiAnn (Rahimi et al., 2019), providing NER labels in 36 languages. Monolingual evaluation mirrors the POS setup. NER was not included in prior analyses and serves as a new probe of lexical-level representations. For cross-lingual transfer, we use the same 18-language subset used in POS.

Natural Language Inference (NLI) For NLI, we use XNLI (Conneau et al., 2018), a multilingual extension of standard NLI benchmarks. We evaluate both monolingual and cross-lingual performance on the 13 XNLI languages that overlap with our 36-language set.

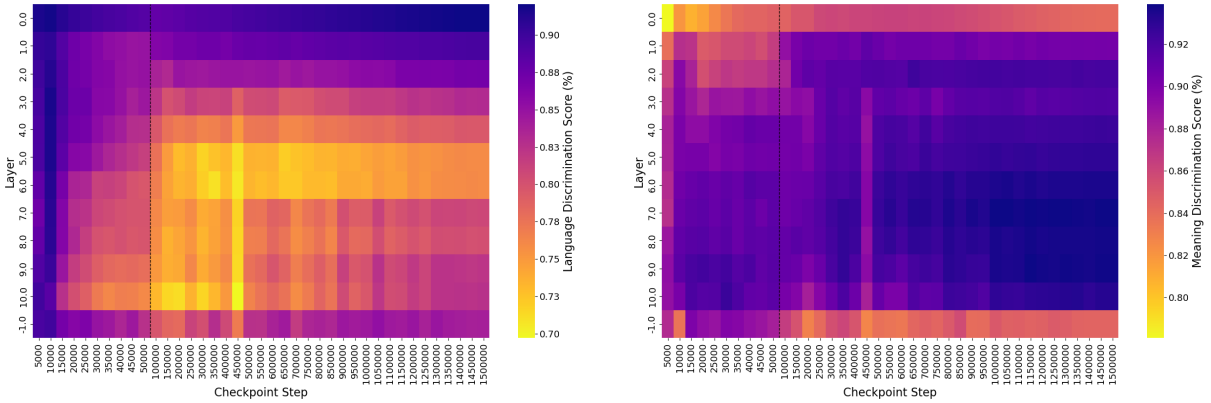


Figure 4: Evolution of ABX discrimination scores across model checkpoints and layers. Dark regions indicate higher discrimination scores. (Left: LD; right: MD)

5.2 Discrimination Scores and Monolingual Linguistic Probing

Following prior work (Blevins et al., 2022), we find the best checkpoint for probe accuracy varies across tasks and languages (see Appendix E)[6].

To assess whether language (LD) or meaning discrimination (MD) predict probing performance, we regress POS, NER, and NLI accuracy against each language's global LD and MD scores, using both the final checkpoint (Last) and the mean across checkpoints (Avg.). For each setting, we fit a multiple linear regression of the form:

$$\text{Accuracy} \sim \beta_0 + \beta_1 \cdot \text{LD} + \beta_2 \cdot \text{MD} + \epsilon$$

Table 1 summarizes the results[7]. For POS, language discrimination is a robust negative predictor of accuracy, while meaning discrimination shows no significant effect. This suggests that languages which are more easily distinguishable from others (i.e., with higher LD scores) tend to perform worse on syntactic probing tasks, consistent with the idea that strong language-specific encoding may hinder generalization of structural information across languages (see Appendix F for visualization).

In contrast, neither LD nor MD significantly predicts performance on NER or XNLI. These tasks may depend less consistently on cross-lingual structural overlap than POS, which could explain the absence of LD as a predictor. While one might expect MD to be predictive, especially for NLI which is explicitly semantic in nature, both tasks may rely on aspects of meaning not well captured by our ABX-based definition of semantic alignment

Setting	Task	Ckpt	R^2	LD Coef (p)	MD Coef (p)
Prob.	POS	Avg.	0.395	$\mathbf{-2.34}$ **(p < .01)**	-0.26 (n.s.)
Prob.	POS	Last	0.37	$\mathbf{-1.78}$ **(p < .01)**	-0.36 (n.s.)
Prob.	NER	Avg.	0.085	-0.578 (n.s.)	-0.074 (n.s.)
Prob.	NER	Last	0.087	-0.584 (n.s.)	-0.147 (n.s.)
Prob.	NLI	Avg	0.275	-0.07 (n.s.)	-0.12 (n.s.)
Prob.	NLI	Last	0.224	-0.08 (n.s.)	-0.14 (n.s.)
CL	POS	Last	0.324	$\mathbf{-1.66}$ **(p < .001)**	-0.07 (n.s.)
CL	NER	Last	0.146	$\mathbf{-0.51}$ **(p < .001)**	$+0.06$ (n.s.)
CL	NLI	Last	0.3009	-0.1 (n.s.)	-0.015 (n.s.)

Table 1: Summary of linear regression results predicting POS and NER accuracy from language discrimination (LD) and meaning discrimination (MD) scores. Each row corresponds to a probing (Prob.) or cross-lingual (CL) evaluation setting.

We also explore whether ABX scores can guide language-specific checkpoint selection, under the hypothesis that lower language discrimination might signal better generalization. We find that LD-based ABX selection improves performance for POS (see Appendix G for details).

5.3 Discrimination Scores and Cross-Lingual Transfer

We evaluate cross-lingual transfer on POS, NER, and NLI at the final checkpoint. As originally found by Blevins et al. (2022), transfer accuracy varies widely across source–target pairs (see Appendix H for detailed heatmaps). To test whether ABX discrimination explains this variation, we fit linear regression models predicting transfer accuracy from LD and MD scores between language pairs (see Table 1). We find that LD is a significant negative predictor for both POS and NER. Neither LD nor MD is predictive of NLI performance. This

[6]We exclude checkpoint 450,000 from all analyses due to a training instability, probably due to gradient clipping, that affects both probing and discrimination metrics (see Figure 2).

[7]We also ensure that these effects are not driven by training data size. Language-wise probing accuracy shows no significant correlation with pretraining data quantities (taken from Conneau et al. (2020a)).

supports the view that strong language-specific encoding can hinder generalization across languages (see Appendix I for visualization). We also test whether ABX language discrimination can guide source language selection for transfer. While it does not consistently identify the single best source, it often selects competitive candidates and outperforms random baselines (see Appendix J).

5.4 Discussion

A key finding is that language discrimination negatively correlates with POS performance in both monolingual and cross-lingual settings. For NER, LD is predictive only in the cross-lingual setting, suggesting that language-specific encoding affects transfer between languages but has less impact on within-language structure. Importantly, ABX discrimination scores' interpretation differs slightly between settings. In monolingual probing, LD/MD scores are averaged across all language pairs per language, while cross-lingual transfer uses pair-specific scores for each source–target combination. This finer granularity may help capture transfer-specific effects, explaining why LD predicts cross-lingual NER but not monolingual performance: interference may depend more on the relationship between particular languages than on a language's overall discriminability. Overall, these results suggest that when a language is highly discriminable from others, its representations may become more isolated, reducing structural sharing and hindering transfer. In the case of POS, this is especially apparent in monolingual probing, where high LD may reflect a failure to encode shared syntactic patterns.

By contrast, MD does not significantly predict downstream accuracy in any task. While one might expect MD to relate to semantically oriented tasks like NLI, success there may depend on higher-level reasoning unaccounted for by our contrastive ABX metric. Prior work has also highlighted problematic annotation artifacts, not to mention hypothesis-only biases in the original SNLI dataset from which XNLI was developed, that limit its use for measuring semantic generalization (Poliak et al., 2018; Gururangan et al., 2018).

6 Conclusion and Future Work

This work introduces ABX-style discrimination metrics for testing how multilingual encoder models discriminate language identity (form) and semantic content (meaning). Adapting ABX to text-based multilingual models, we provide a lightweight, interpretable tool for analysing representational structure.

Applied to XLM-R (Conneau et al., 2020a), our analysis reveals consistent trends across training and depth: language discrimination decreases and concentrates in lower layers, while meaning discrimination increases and stabilizes in deeper ones. This suggests a shift from form-sensitive to meaning-oriented representations during training, without implying a strict trade-off. We also examine how these metrics relate to downstream performance. Higher language discrimination correlates with lower accuracy on form-sensitive tasks such as POS and NER, while meaning discrimination shows no consistent link, pointing to a possible disconnect between representational alignment and task requirements. These findings position ABX discrimination as a useful metric for analyzing how multilingual models separate linguistic form from content. They offer a new lens on the evolving structure of multilingual representation spaces and the balance between language-specific and language-invariant information. These metrics could also support practical use cases, such as adaptive checkpoint selection or lightweight diagnostics in multilingual pipelines.

Future work can build on this in several directions. First, discrimination patterns could be related to typological linguistic features, and it is worth investigating how this typological differences can influences form and content discrimination scores, as previous work has found positive transfer when pairing typologically similar languages (Wu and Dredze, 2020). Second, while we evaluated tasks spanning syntax and semantics (POS, NER, NLI), deeper semantic tasks could better test the role of meaning discrimination. Third, because ABX metrics are architecture-agnostic, they can be applied to decoder-only LLMs, enabling cross-architecture comparisons. Finally, while ABX does not directly measure representational separation, high discrimination may suggest that a language occupies a distinct subspace. This raises a broader question: how much language sensitivity (i.e., the ability to discriminate languages) can a model have without harming cross-lingual transfer, and how can models balance this trade-off between promoting representational sharing and avoiding interference?

Limitations

While our approach provides a detailed analysis of discrimination in multilingual models, it comes with a number of limitations that constrain its generality and suggest directions for future research.

Encoder architecture Our analysis focuses exclusively on encoder-only architectures, specifically XLM-R. This choice is motivated by the fact that encoder models based on masked language modeling provide stable, structured, layer-wise representations, which are well suited to probing and contrastive analysis. While this makes them a natural starting point for validating our ABX discrimination framework, it remains an open question whether similar dynamics hold for decoder-only or encoder–decoder models, which are trained using autoregressive or sequence-level objectives. Extending our framework to such architectures is an important direction for future work.

Discrimination vs Separation Although we distinguish clearly between language and meaning discrimination, we do not explicitly quantify separation in the representation space (e.g., via clustering structure, inter-class variance). Our results suggest that discrimination scores may indirectly reflect separation, but further work is needed to validate this link and to determine whether a model can be discriminative without being structurally partitioned.

Task coverage Our evaluation focuses on POS tagging, NER, and NLI, which primarily target syntactic and sentence-level semantic understanding. While these tasks are widely used and informative, they may not capture deeper semantic, pragmatic, or discourse-level capabilities. As a result, the role of meaning discrimination in supporting more abstract or context-sensitive generalization remains an open question.

Cross-linguistic generality vs. language-specific phenomena Finally, our analysis examines language and meaning discrimination broadly across multiple languages, but does not investigate the intricacies of specific languages or language families. Languages exhibit unique structural properties, morphological complexity, and semantic nuances that may be represented differently in multilingual models. Future work should explore language-specific discrimination patterns, particularly for typologically diverse languages, to better understand how models encode both universal and language-specific linguistic properties. This would provide insights into representational trade-offs that occur when accommodating multiple languages within a shared parameter space.

References

Robin Algayres, Yossi Adi, Tu Nguyen, Jade Copet, Gabriel Synnaeve, Benoît Sagot, and Emmanuel Dupoux. 2023. Generative spoken language model based on continuous word-sized audio tokens. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3008–3028.

Robin Algayres, Tristan Ricoul, Julien Karadayi, Hugo Laurençon, Salah Zaiem, Abdelrahman Mohamed, Benoît Sagot, and Emmanuel Dupoux. 2022. Dpparse: Finding word boundaries from raw speech with an instance lexicon. *Transactions of the Association for Computational Linguistics*, 10:1051–1065.

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

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2022. Analyzing the mono-and cross-lingual pretraining dynamics of multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3575–3590.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837.

Maria Julia Carbajal, Emmanuel Dupoux, and 1 others. 2016. Modeling language discrimination in infants using i-vector representations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 38.

Rochelle Choenni and Ekaterina Shutova. 2022. Investigating language relationships in multilingual sentence encoders through the lens of linguistic typology. *Computational Linguistics*, 48(3):635–672.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised

cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

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

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

Maureen de Seyssel. 2023. *Unsupervised multilingual models of speech representation, an approach inspired by cognitive science*. Ph.D. thesis, Ecole Normale Supérieure (ENS).

Maureen de Seyssel and Emmanuel Dupoux. 2020. Does bilingual input hurt? a simulation of language discrimination and clusteringusing i-vectors. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 42.

Maureen de Seyssel, Guillaume Wisniewski, and Emmanuel Dupoux. 2022. Is the language familiarity effect gradual? a computational modelling approach. In *CogSci 2022-44th Annual Meeting of the Cognitive Science Society*.

Daniel Deutsch, Eleftheria Briakou, Isaac Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, and 1 others. 2025. Wmt24++: Expanding the language coverage of wmt24 to 55 languages & dialects. *arXiv preprint arXiv:2502.12404*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE.

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.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.

Mark Hallap, Emmanuel Dupoux, and Ewan Dunbar. 2022. Evaluating context-invariance in unsupervised speech representations. *arXiv preprint arXiv:2210.15775*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: a massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421.

Yongxin Huang, Kexin Wang, Goran Glavaš, and Iryna Gurevych. 2024. Modular sentence encoders: Separating language specialization from cross-lingual alignment. *arXiv preprint arXiv:2407.14878*.

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

Alireza Mohammadshahi, Rémi Lebret, and Karl Aberer. 2019. Aligning multilingual word embeddings for cross-modal retrieval task. In *Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN)*, pages 11–17.

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

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495.

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. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891.

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

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. 2018. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 180–191.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for ner. In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*, pages 151–164. Association for Computational Linguistics-ACL.

Kartik Ravisankar, Hyojung Han, and Marine Carpuat. 2025. Can you map it to english? the role of cross-lingual alignment in multilingual performance of llms. *arXiv preprint arXiv:2504.09378*.

Thomas Schatz. 2016. *ABX-discriminability measures and applications*. Ph.D. thesis, Université Paris 6 (UPMC).

Thomas Schatz, Vijayaditya Peddinti, Francis Bach, Aren Jansen, Hynek Hermansky, and Emmanuel Dupoux. 2013. Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, pages 1–5.

Thomas Schatz, Vijayaditya Peddinti, Xuan-Nga Cao, Francis R Bach, Hynek Hermansky, and Emmanuel Dupoux. 2014. Evaluating speech features with the minimal-pair abx task (ii): resistance to noise. In *INTERSPEECH*, pages 915–919.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Amitay Sicherman and Yossi Adi. 2023. Analysing discrete self supervised speech representation for spoken language modeling. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. *arXiv preprint arXiv:2502.15639*.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.

Marc Tanti, Lonneke van der Plas, Claudia Borg, and Albert Gatt. 2021. On the language-specificity of multilingual bert and the impact of fine-tuning. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 214–227.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Craig A Thorburn, Naomi H Feldman, and Thomas Schatz. 2019. A quantitative model of the language familiarity effect in infancy. In *Proceedings of the conference on cognitive computational neuroscience*.

Maarten Versteegh, Roland Thiolliere, Thomas Schatz, Xuan-Nga Cao, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2015. The zero resource speech challenge 2015. In *Interspeech*, volume 15, pages 3169–3173.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2024. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. *arXiv preprint arXiv:2410.03115*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

A Languages Used in Evaluations

Table 2 lists all languages selected for our different evaluations, including ABX discrimination tasks and probing tasks. The selection covers a wide range of language families, scripts, and typological characteristics.

Code	Language	ABX	POS		NER		NLI	
			mono.	CL	mono.	CL	mono.	CL
ar	Arabic	✓	✓	✓	✓	✓	✓	✓
bg	Bulgarian	✓	✓		✓	✓	✓	✓
ca	Catalan	✓	✓		✓			
cs	Czech	✓	✓	✓	✓	✓		
da	Danish	✓	✓		✓			
de	German	✓	✓	✓	✓	✓	✓	✓
el	Greek	✓	✓		✓	✓	✓	✓
en	English	✓	✓	✓	✓	✓	✓	✓
es	Spanish	✓	✓	✓	✓	✓	✓	✓
et	Estonian	✓	✓		✓			
fa	Persian	✓	✓		✓			
fi	Finnish	✓	✓	✓	✓	✓		
fr	French	✓	✓	✓	✓	✓	✓	✓
he	Hebrew	✓	✓		✓			
hi	Hindi	✓	✓	✓	✓	✓	✓	✓
hr	Croatian	✓	✓		✓			
hu	Hungarian	✓	✓		✓			
is	Icelandic	✓	✓	✓	✓	✓		
it	Italian	✓	✓	✓	✓	✓		
ja	Japanese	✓	✓	✓	✓	✓		
ko	Korean	✓	✓	✓	✓	✓		
lv	Latvian	✓	✓		✓			
nl	Dutch	✓	✓		✓			
pl	Polish	✓	✓	✓	✓	✓		
pt	Portuguese	✓	✓	✓	✓	✓		
ro	Romanian	✓	✓		✓			
ru	Russian	✓	✓	✓	✓	✓	✓	✓
sk	Slovak	✓	✓		✓			
sl	Slovenian	✓	✓		✓			
sr	Serbian	✓	✓		✓			
sv	Swedish	✓	✓	✓	✓	✓		
tr	Turkish	✓	✓	✓	✓	✓	✓	✓
uk	Ukrainian	✓	✓		✓			
ur	Urdu	✓	✓		✓	✓	✓	✓
vi	Vietnamese	✓	✓		✓		✓	✓
zh	Chinese	✓	✓	✓	✓	✓	✓	✓

Table 2: Languages and related ISO codes used in discrimination evaluations (ABX), and probing tasks (mono for monolingual probing and CL for cross-lingual). A checkmark indicates the language is used in that task/subset.

B Illustration of Language and Meaning ABX Setups

Figure 5 illustrates our adaptation of the ABX discrimination paradigm for evaluating multilingual text representations. The figure depicts our two complementary evaluation setups: Language Discrimination (LD) and Meaning Discrimination (MD).

In both setups, we follow a consistent structure where A and X share the variable of interest (the property we want the model to discriminate), while B and X share a control variable (the property we want to control for). Success is measured by whether the model places X closer to A than to B in the embedding space.

For the Language Discrimination task (left panel), the variable of interest is language identity, while meaning serves as the control variable. Specifically, A and X share the same language (L_1) but express different meanings, while A and B share the same meaning but are expressed in different languages. When $d(X, A) < d(X, B)$, the model successfully discriminates based on language identity despite semantic differences.

An example is given below:

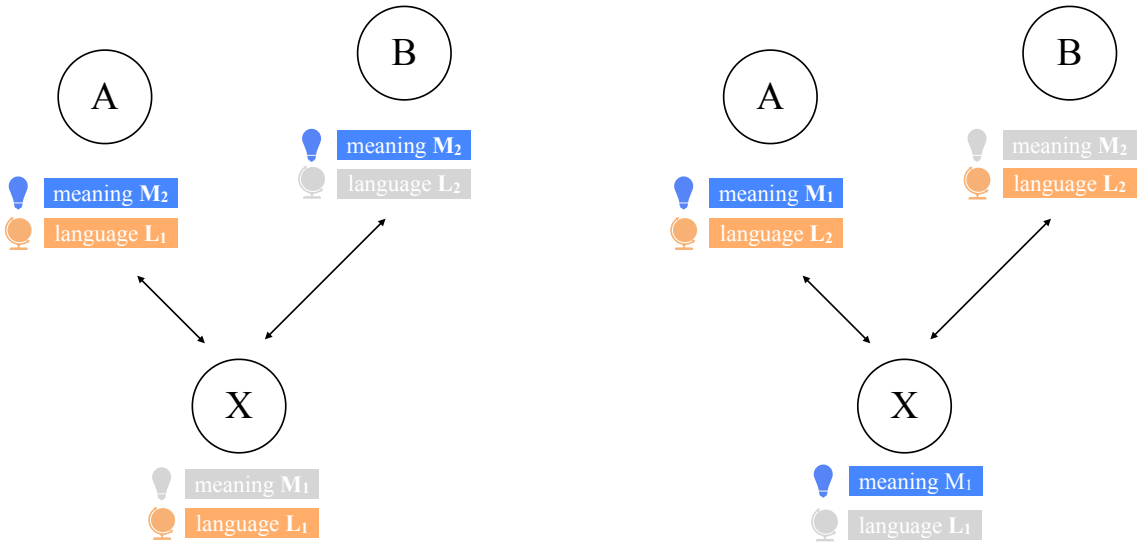
- X: "The weather is nice today." (English (L_1), meaning M_1)
- A: "I need to buy groceries." (English (L_1), meaning M_2)
- B: "Je dois acheter des provisions." (French (L_2), meaning M_2: "I need to buy groceries")

For the Meaning Discrimination task (right panel), the variable of interest is semantic content, while language identity serves as the control variable. Here, A and X share the same meaning (M_1) but are expressed in different languages, while A and B share the same language (L_2) but express different meanings. When $d(X, A) < d(X, B)$, the model successfully discriminates based on semantic similarity across languages despite surface form differences.

Here is an example for the MD task:

- X: "The weather is nice today." (English (L_1), meaning M_1)
- A: "La météo est bonne aujourd'hui" (French (L_2), meaning M_1: "The weather is nice today")
- B: "Je dois acheter des provisions." (French (L_2), meaning M_2: "I need to buy groceries")

This systematic approach allows us to isolate specific properties in multilingual representations by controlling for potential confounding factors. The ABX score for each task reflects the proportion of triplets where the model correctly places items sharing the variable of interest closer together than those sharing only the control variable, providing a direct measure of how the model structures linguistic information along these dimensions.



(a) Language Discrimination

(b) Meaning Discrimination

Figure 5: Illustration of the Language Discrimination (left) and Meaning Discrimination (right) ABX tasks.

C Correlation Analysis Between Language and Meaning Discrimination in XLM-R

This appendix provides additional analysis on the relationship between language discrimination (LD) and meaning discrimination (MD) in our model.

We observe a strong overall negative correlation between LD and MD across all language pairs and checkpoints (Spearman's $\rho = -0.74$, $p < 0.001$; Pearson's $r = -0.68$, $p < 0.001$), computed at the (language pair × checkpoint) level. This suggests that, throughout training, language pairs that are more separable in form tend to be less effective in preserving semantic structure.

To verify that this effect is not simply an artifact of training progression, we examine the relationship at the final checkpoint (step 150,000) alone. The inverse correlation persists with even greater magnitude (Spearman's $\rho = -0.83$, $p < 0.001$; Pearson's $r = -0.72$, $p < 0.001$), confirming that the tradeoff between language and meaning discrimination remains pronounced even in the fully trained model. Figure 6 visualizes this relationship: the scatterplot reveals a clear monotonic trend, with almost no high–high co-occurrence (i.e., no language pairs simultaneously scoring high on both MD and LD), which supports the interpretation of a representational tradeoff.

We further analyze the dynamics of this relationship across training by computing correlation coefficients at each checkpoint (Figure 7). Spearman's correlation remains consistently strong and statistically significant across all training stages, suggesting a stable monotonic inverse relationship. Pearson's correlation, while also consistently negative, varies in magnitude but remains significant as training progresses, indicating that the relationship is not only ordinal but approximately linear in later stages.

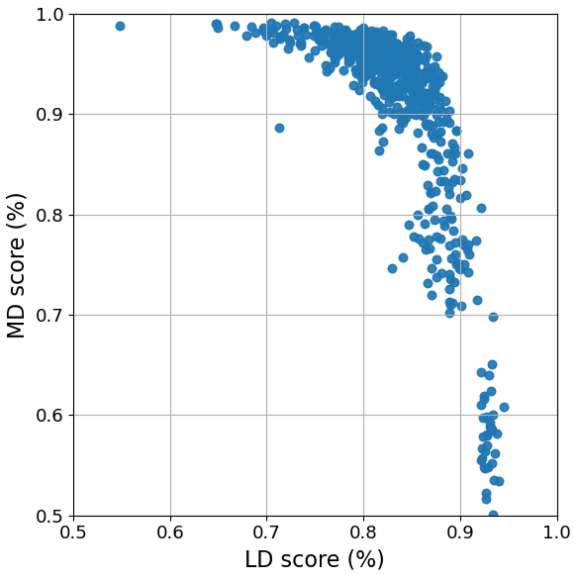


Figure 6: Scatterplot showing the relationship between language discrimination (x-axis) and meaning discrimination (y-axis) scores for all language pairs at checkpoint 150,000 (last). Each point represents a language pair.

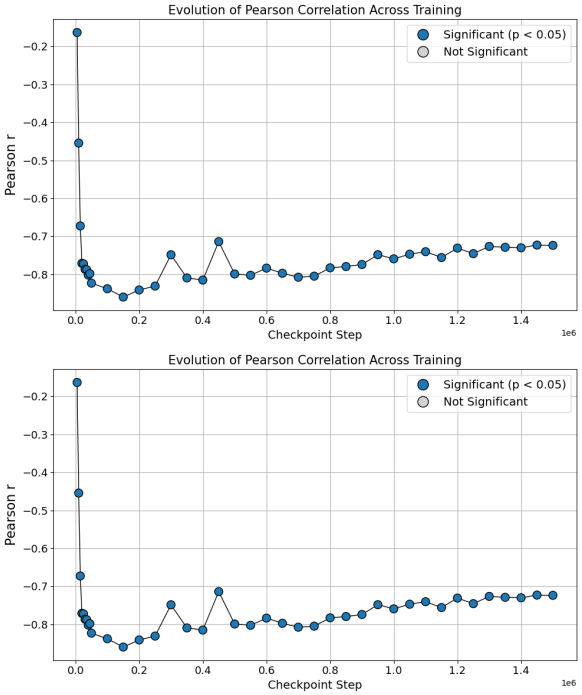


Figure 7: Evolution of Spearman (top) and Pearson (bottom) correlation coefficients between language and meaning discrimination scores across training checkpoints. Statistically significant correlations (p < 0.05) are highlighted.

D Discrimination scores across languages and checkpoints

To examine how discrimination evolves at the individual language level, we present heatmaps of language and meaning discrimination scores across checkpoints (Figure 8). Scores are normalised per language to highlight relative changes over time. For language discrimination (left), we observe a sharp decline during early training steps for most languages, followed by a partial recovery. However, the timing and extent of this rebound varies across languages, suggesting that some retain language-specific features more robustly. In contrast, meaning discrimination (right) increases steadily for all languages, but again at different rates, with certain languages benefiting earlier from semantically structured representations. These differences may reflect both linguistic factors and data resource disparities. Additional views of final-layer behaviour are included in Figure 9.

E Checkpoint-wise Probe Accuracy

Figure 10 shows per-language probe accuracy across checkpoints for POS, NER and NLI highlighting the variability in when each language reaches its peak performance.

F Additional Probing Analyses

Figure 11 shows the negative relationship between ABX language discrimination and POS accuracy across languages. Higher language discrimination scores are associated with lower probing performance, consistent with the idea that strong language-specific encoding may limit generalization.

G ABX-Guided Checkpoint Selection

Given that language discrimination is a strong global predictor of probing accuracy for POS, we ask whether ABX scores can serve as lightweight, unsupervised heuristics for language-specific checkpoint selection. Specifically, we evaluate whether selecting, for each language, the checkpoint with minimal LD brings the model closer to its optimal performance, compared to using the final checkpoint uniformly.

We compare probing accuracy at the ABX-selected checkpoint to that at the final training step, measuring their respective distances from each language's best-performing checkpoint.

ABX-guided selection yields a closer match to the best checkpoint in 29 out of 36 languages, with a mean improvement of 0.034 ± 0.048, and a Wilcoxon signed-rank test confirming significance over choosing the final checkpoint ($p < 0.001$). This suggests that LD dynamics during training can inform language-specific model selection, particularly when the final checkpoint is suboptimal.

These patterns are visualised in Figure 13, which presents per-language deltas. For each language, we compute:

$$\Delta = \text{Final} - \text{ABX}$$

where positive values indicate that ABX selection yields a checkpoint closer to the best-performing one. Bars are sorted by the absolute delta, highlighting languages with the largest impact.

H Cross-Lingual Transfer Accuracy Matrices

Figure 14 shows the full cross-lingual probing results for POS, NER, and XNLI at the final checkpoint. Each heatmap shows transfer accuracy from a source language (row) to a target language (column). The highest-performing source language for each target is highlighted in yellow.

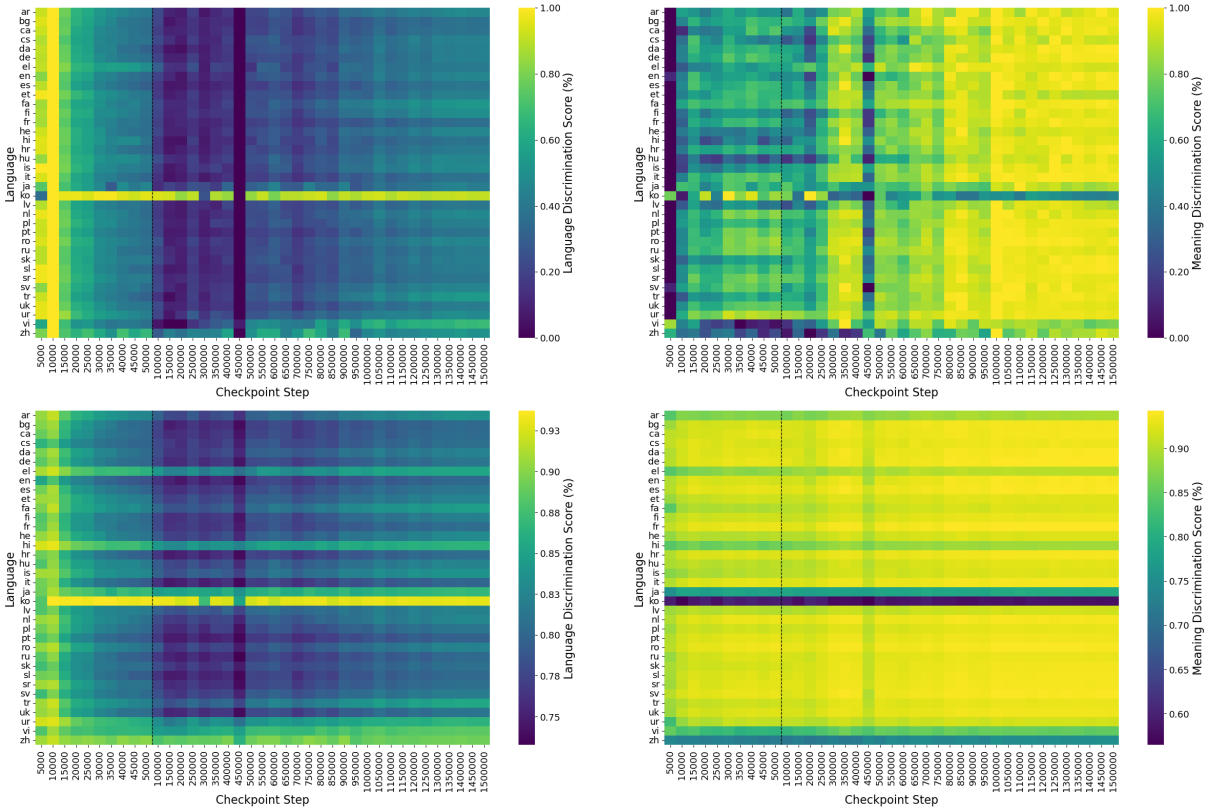


Figure 8: Heatmaps showing the evolution of language discrimination (left) and meaning discrimination (right) *(averaged across layers)* across checkpoints. Scores are normalized per language to highlight the differences across checkpoints (top row). We also provide the non-normalized scores (bottom row). Each heatmap's row represents a language, and each column a checkpoint step. Bright regions indicate high relative discrimination at that training stage for the given language.

I Visualization of Language Discrimination Effects on Cross-Lingual Performance

Figure 15 illustrates the relationship between language discrimination scores and cross-lingual transfer accuracy for all source-target language pairs in our experiments. For both POS tagging and NER tasks, we observe a strong negative correlation: language pairs with higher discrimination scores (indicating more distinct linguistic forms) consistently show lower transfer performance. This visualization reinforces our regression findings that language discrimination acts as a significant negative predictor of cross-lingual transfer success.

The scatter plots reveal that when models encode languages in ways that make their forms highly distinguishable from each other, their ability to transfer knowledge between those languages for form-focused task as POS and NER diminishes. Conversely, when language forms are less discriminable (more shared or mixed representations), cross-lingual transfer improves.

J ABX-Guided Source Language Selection

Inspired by our earlier use of ABX scores to guide checkpoint selection (Section G), we investigate whether ABX language discrimination can also inform source language selection in cross-lingual transfer. Specifically, for each target language, we test whether the source language with the lowest ABX language discrimination score yields the highest transfer performance.

Exact Match and Top-k Accuracy We first compare, for each target language, the true best source (i.e., the one yielding the highest transfer accuracy) with the ABX-selected source (i.e., the one with minimal ABX LD). Exact matches occur in 2/18 (POS) and 7/18 (NER) cases. When considering the top-3 sources, ABX guidance succeeds in 6/18 (POS) and 12/18 (NER) cases, suggesting it often identifies competitive transfer candidates.

Comparison to Random Selection Next, we evaluate how ABX-guided selection compares to

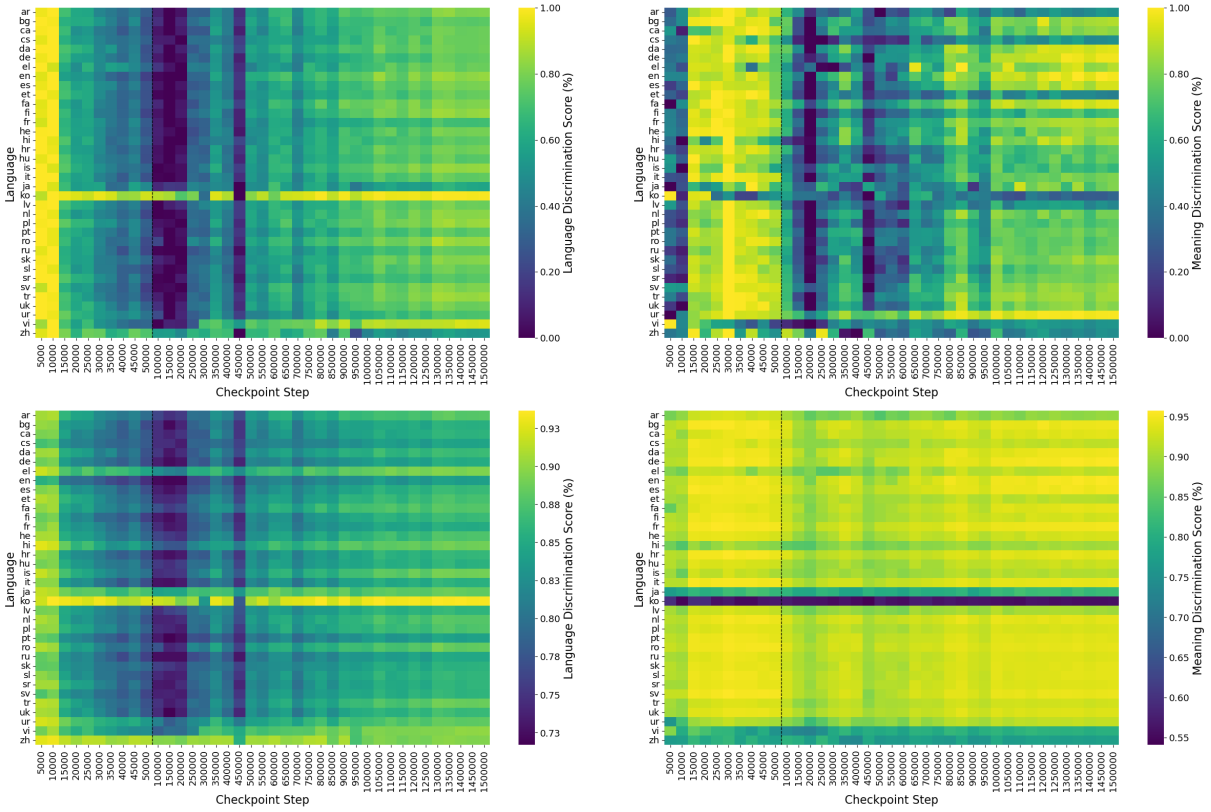


Figure 9: Heatmaps showing the evolution of language discrimination (left) and meaning discrimination (right) on *the last layer*, across checkpoints. Scores are normalized per language to highlight the differences across checkpoints (top row). We also provide the non-normalized scores (bottom row). Each heatmap's row represents a language, and each column a checkpoint step. Bright regions indicate high relative discrimination at that training stage for the given language.

a naive random baseline. For each target, we compare the transfer accuracy of the ABX-selected source to that of 100 randomly sampled sources, and compute the proportion of wins. The ABX-guided source outperforms a random one in 73.0% ± 27.5% of trials for POS, and 84.8% ± 23.4% for NER.

Figure 16 shows the full distribution of these per-target win rates. Most values exceed 70–80%, and very few fall below the 50% chance level, indicating that ABX LD offers a consistent and effective heuristic for source selection.

Conclusion While ABX-guided source selection does not always identify the single best transfer source, it reliably outperforms random baselines. Compared to typological or lexical similarity heuristics (which are often noisy or task-specific) ABX LD offers a simple, data-driven alternative for identifying effective source languages in cross-lingual transfer.

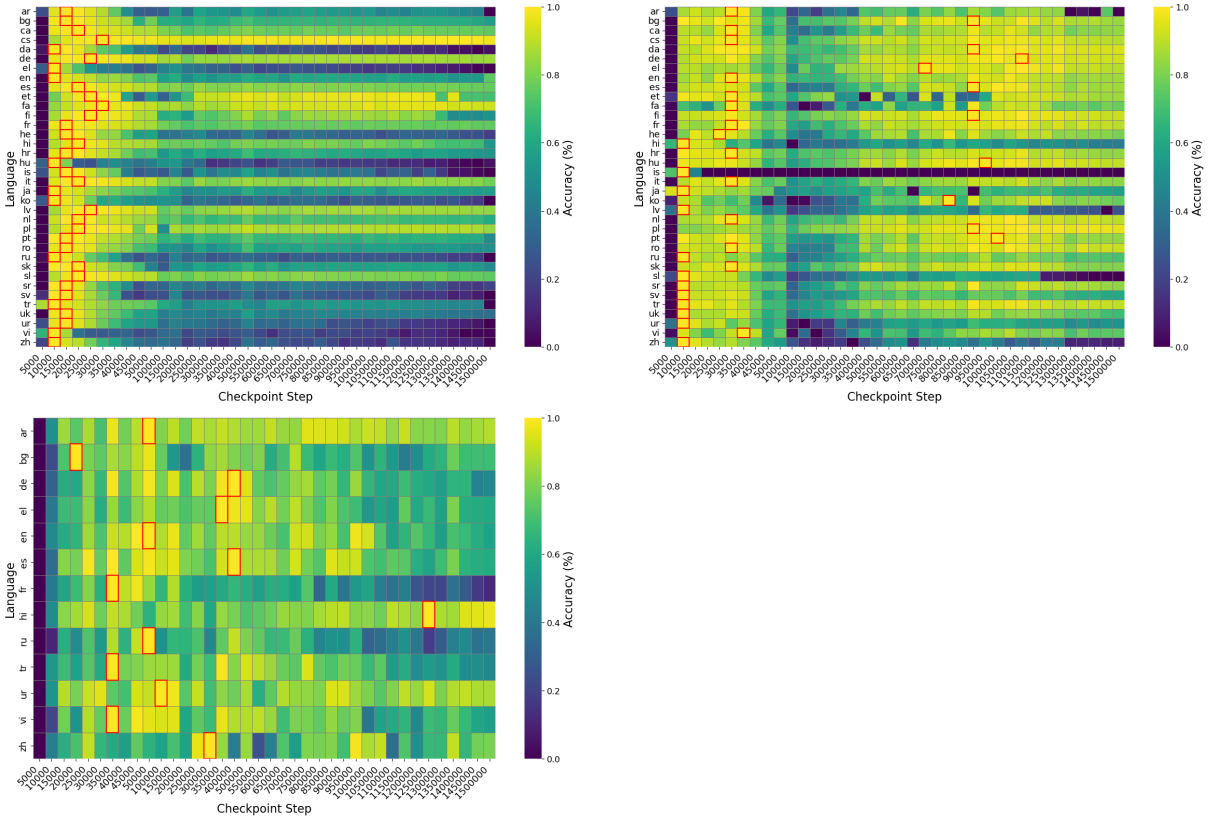


Figure 10: Checkpoint-wise probe accuracy across languages for POS (top left), NER (right), and NLI (bottom left) normalized per language. Each row corresponds to a language, and red boxes mark the checkpoint at which that language reaches peak accuracy for the probing task. Lighter regions mean higher accuracy scores.

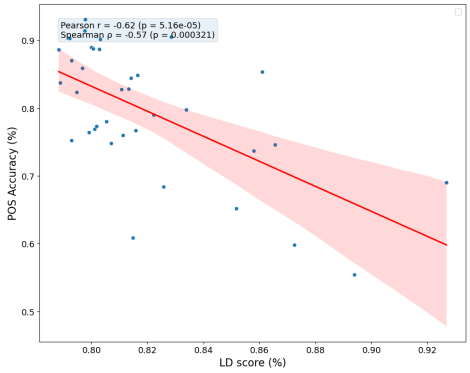


Figure 11: Relationship between ABX-based Language Discrimination scores and downstream probing POS accuracy, averaged across checkpoints. Each point represents a single evaluation language. The x-axis shows how well the model distinguishes that language from others (higher = more discriminable), while the y-axis shows its average performance on the downstream task. Red lines indicate linear regression fits with shaded 95% confidence intervals.

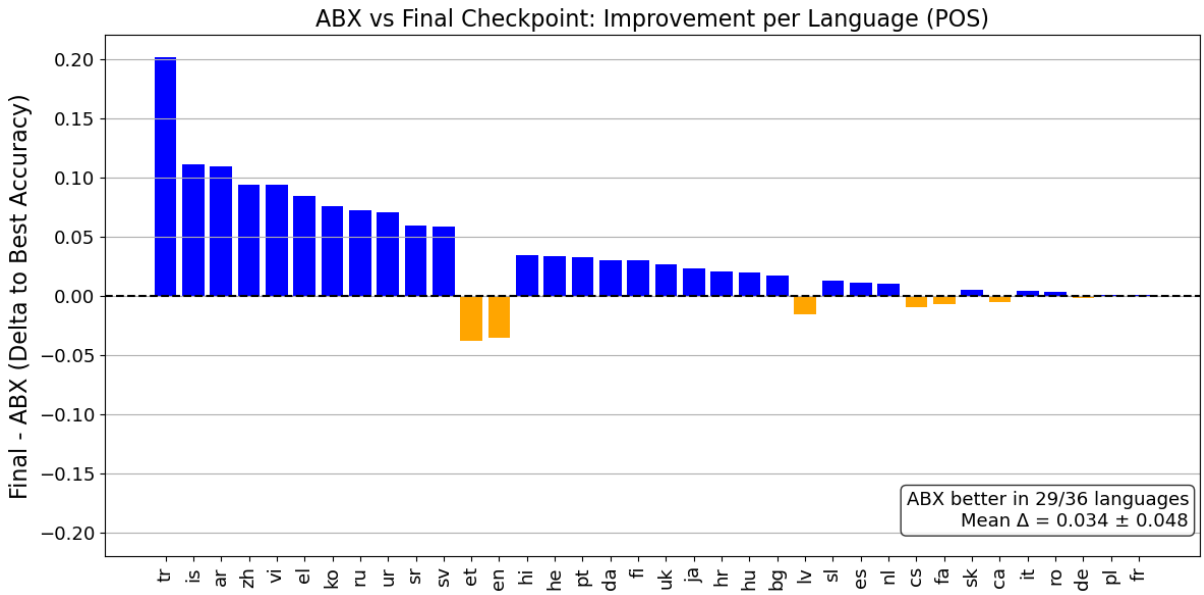


Figure 12: POS

Figure 13: Difference in performance gap to the best checkpoint for each language, comparing ABX-selected (lowest LD ABX) vs. final checkpoints for the POS task. Bars show the difference in delta (Final - ABX); positive values indicate that the ABX-selected checkpoint is closer to the best-performing one (i.e., smaller gap to optimal accuracy).

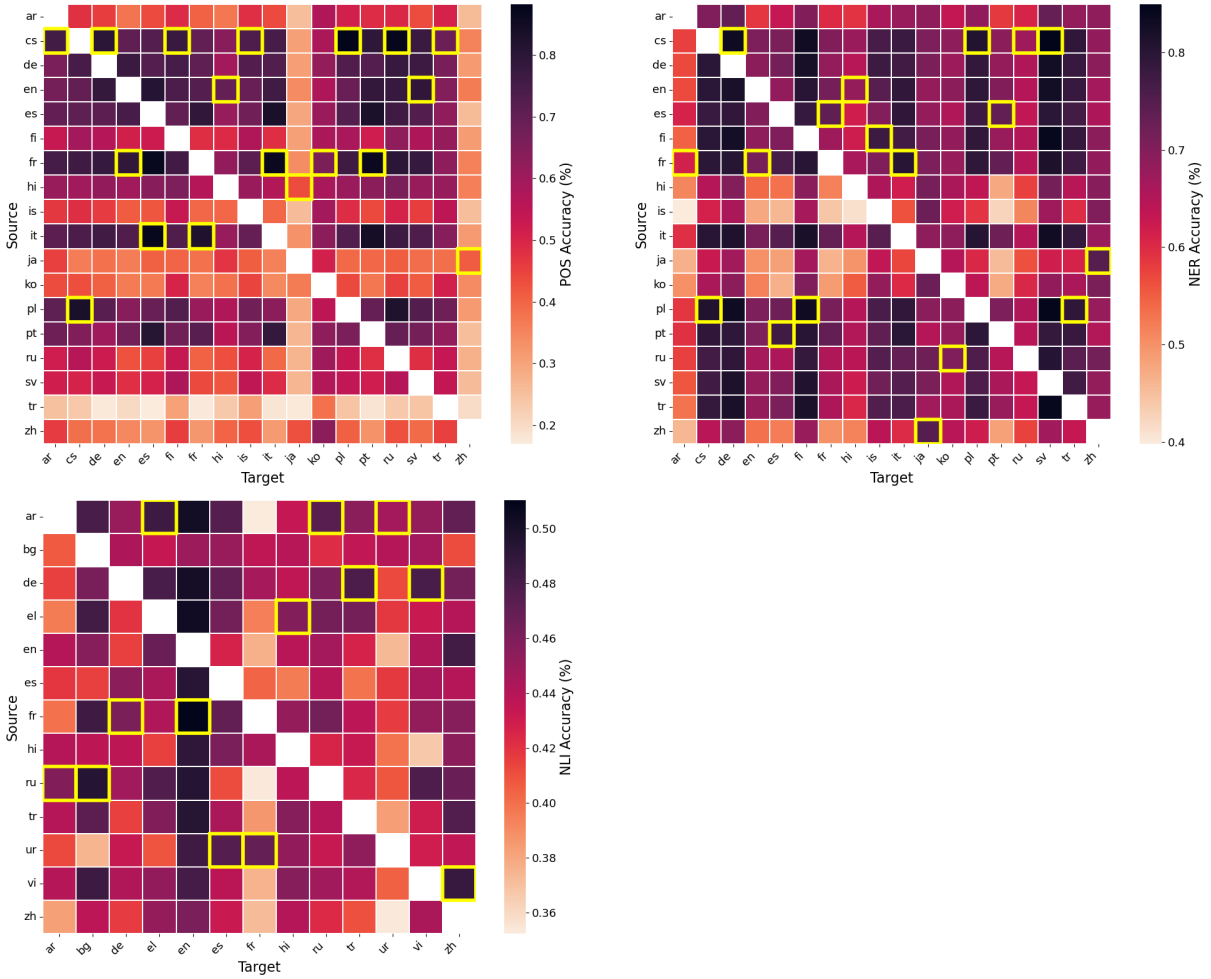


Figure 14: Cross-lingual probing accuracy at the final checkpoint for POS (top left), NER (top right), and XNLI (bottom). Each cell shows accuracy of a probe trained on the source language (row) and evaluated on the target (column). Best source for each target is highlighted in yellow.

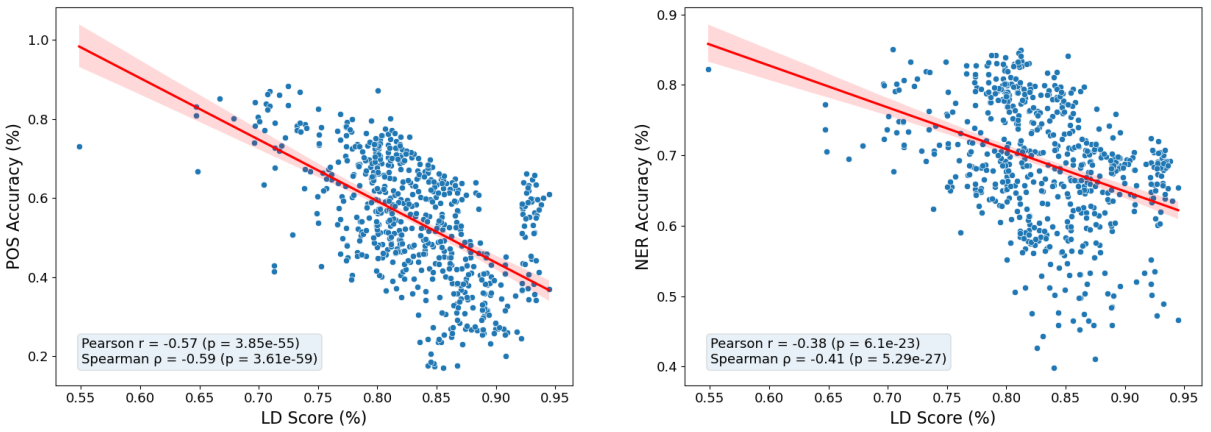


Figure 15: Relationship between language discrimination scores and cross-lingual transfer accuracy for POS tagging (left) and NER (right) across all source-target language pairs. Each point represents a language pair, with the x-axis showing the language discrimination score and the y-axis showing transfer accuracy. The downward trend demonstrates that higher language discrimination (more distinct language forms) is associated with lower cross-lingual transfer performance for POS and NER.

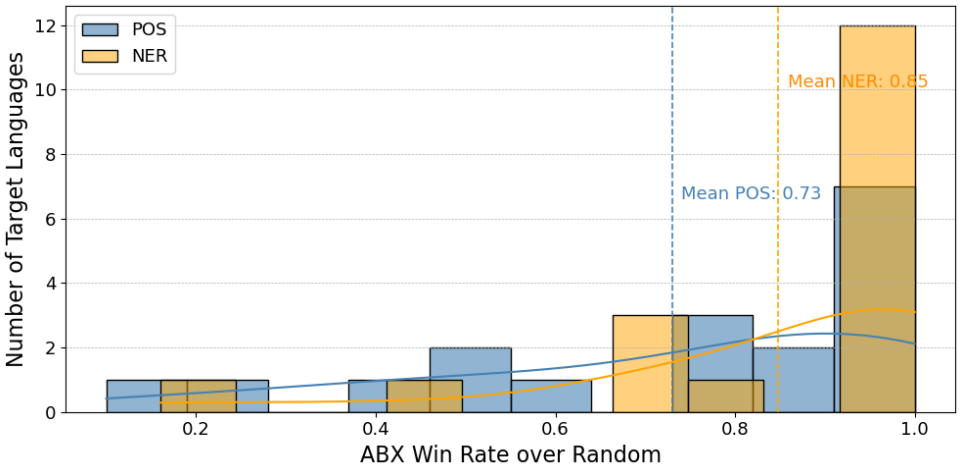


Figure 16: Distribution of ABX win rates across target languages for POS (blue) and NER (orange). Dashed lines indicate average win rate per task. A value above 0.5 reflects better-than-random performance.