# MELD: MEASURING MULTILINGUAL CAPABILITY OF LANGUAGE MODELS THROUGH LATENT DISTRIBUTION GEOMETRY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multilingual large language models exhibit substantial variation in capability across languages, yet evaluating this variation typically requires costly downstream benchmarks. We introduce MELD, an intrinsic metric for estimating multilingual capability directly from the geometry of hidden states. MELD is motivated by growing evidence that multilingual LLMs develop a shared semantic and reasoning space in their intermediate layers, while language-specific structure is more prominent near the input and output boundaries. MELD measures distributional language neutrality by contrasting the separation between reference- and target-language representation distributions with their within-language dispersion. We evaluate MELD across 14 open-weight LLMs spanning multiple model families, parameter scales, depths, and training regimes, and validate it against five multilingual benchmark families covering more than 30 languages. We assess MELD from two complementary perspectives: predicting cross-language performance differences within a model and ranking models by their overall multilingual capability. In the within-model setting, MELD achieves average Pearson and Spearman correlations of about 0.94 and 0.93, respectively, with downstream multilingual performance. We further show that this predictive relationship remains strong when sentence-level correspondence is removed using non-parallel reference data, whereas prior methods degrade substantially. MELD also strongly correlates with multilingual capability across all models, achieving average Pearson and Spearman correlations of about 0.78 and 0.75, respectively, with overall downstream performance. Together, these results demonstrate that MELD provides a robust intrinsic signal of multilingual capability across both languages and models.

## 1 INTRODUCTION

Large language models (LLMs) are increasingly deployed across a wide range of languages, yet their capabilities remain highly uneven across them. Models that perform strongly in English can exhibit substantial degradation in medium- and low-resource languages, even when the underlying task or semantic content is unchanged (Xu et al., 2025; Wang et al., 2025; Lu et al., 2025). This disparity makes multilingual evaluation particularly important, but existing downstream benchmarks remain costly to construct, task-specific, and limited in their coverage of the world's languages (Kargaran et al., 2025). As a result, recent work has explored whether multilingual capability can instead be estimated directly from the internal representations of LLMs (Li et al., 2025c; Kargaran et al., 2025).

Existing intrinsic approaches provide strong evidence that cross-lingual representation geometry is predictive of downstream multilingual performance (Li et al., 2025c; Kargaran et al., 2025). However, these methods primarily measure alignment between representations of *paired translations*. Their scores therefore depend on instance-level bilingual correspondence and conflate two distinct properties: whether the model represents the same sentence similarly across languages, and whether the two languages occupy similarly structured regions of the model's latent space. This dependence also requires access to high-quality parallel data, which is itself scarce for many of the languages for which intrinsic evaluation would be most useful.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

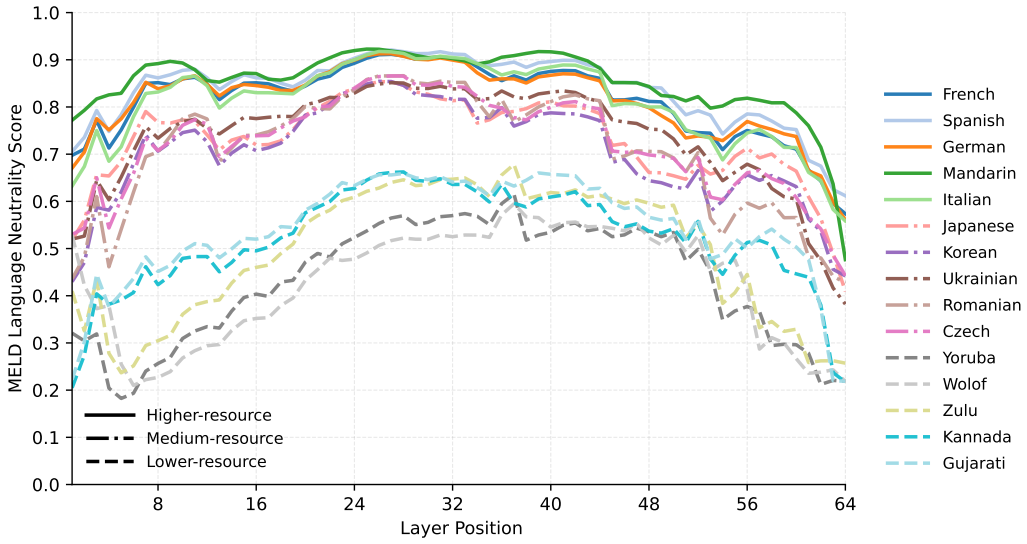


Figure 1: MELD measures multilingual capability from the distributional geometry of intermediate hidden states. At each layer, MELD quantifies the separation between reference- and target-language representation distributions relative to their within-language variability, with higher scores indicating stronger language neutrality. The characteristic rise in neutrality toward intermediate layers motivates MELD's focus on the shared middle-layer regime, where stronger neutrality is predictive of downstream multilingual capability. Shown for Qwen3.5-27B on NTREX-128.

At the same time, mechanistic studies point to a broader structural regularity in multilingual LLMs: language-specific information is particularly prominent near the input and output boundaries, while intermediate layers tend to exhibit more shared semantic and reasoning structure (Tang et al., 2024; Wendler et al., 2024; Bandarkar et al., 2025; Harrasse et al., 2025). Functional interventions further suggest that the degree to which languages access this shared computation is related to multilingual performance (Zhao et al., 2026; Bandarkar et al., 2026). This suggests that multilingual capability may be reflected not only in the alignment of individual translation pairs, but in the geometry of language-conditioned representation distributions themselves.

Motivated by this observation, we ask a different question: *to what extent do representations from different languages become distributionally neutral in the region where shared multilingual computation is strongest?* Importantly, we do not assume that successful multilingual models eliminate language-specific structure. Prior work shows that shared and language-sensitive information can coexist within the same representation space (Chang et al., 2022; Hämmerl et al., 2024). Instead, we hypothesize that *stronger multilingual capability is associated with smaller language-conditioned separation relative to the natural variability of each language's representations* in intermediate layers. Our main contributions are:

- We introduce **MELD** (**M**ultilingual **E**valuation of **L**atent **D**istributions), an intrinsic metric that operationalizes this hypothesis by measuring the separation between language-conditioned representation distributions relative to their within-language variability in intermediate layers.
- A key property of MELD is that it is *set-based rather than pair-based*: it depends only on statistics of the two representation sets and does not require sentence-level correspondence or ordering. Consequently, MELD can be computed from independently sampled monolingual corpora rather than aligned translations.
- We evaluate MELD across 14 open-weight LLMs spanning multiple model families, parameter scales, network depths, and training regimes. Using five multilingual benchmark families, we assess whether MELD predicts *cross-language capability differences within a model* and *overall multilingual capability across models*. Across both settings, MELD

shows a strong and consistent relationship with downstream multilingual performance. We also release our source code to facilitate reproducibility and future research[1].

## 2 Background and Related Work

**Intrinsic Multilingual Evaluation.** Multilingual capability is typically assessed using downstream benchmarks that directly measure task performance across languages. But benchmark-based evaluation remains task-dependent and can be costly to construct and run at scale (Wang et al., 2026). This has motivated intrinsic approaches that estimate multilingual capability from model representations rather than task-specific outputs.

Language Ranker uses the similarity between English and non-English hidden representations as a proxy for language-specific capability, showing that stronger representational similarity to English is associated with better downstream performance (Li et al., 2025c). MEXA develops this idea further by measuring cross-lingual alignment through bidirectional retrieval of parallel sentence representations and reports strong agreement with multilingual benchmark performance (Kargaran et al., 2025). Related probing work similarly studies cross-language hidden-state similarity as an indicator of multilingual competence (Li et al., 2025a). These methods establish that multilingual capability is reflected in representation geometry, but they characterize it primarily through instance-level alignment between semantically corresponding sentences.

**Shared and Language-Specific Representations.** A substantial literature studies how multilingual models balance language-independent and language-specific structure. Early geometric analyses show that different languages can occupy similar mean-centered subspaces while remaining separated along language-sensitive directions (Chang et al., 2022). Language-agnostic structure can also be strengthened by removing low-rank language-specific components (Xie et al., 2022), while sparse subnetworks with substantial cross-language overlap suggest that multilingual models reuse common computational machinery across languages (Foroutan et al., 2022). More broadly, surveys of cross-lingual alignment emphasize that useful multilingual representations need not be fully invariant, but may instead combine shared semantic structure with language-dependent information (Hämmerl et al., 2024).

Mechanistic analyses further show that this balance varies systematically with network depth. Language-specific neurons are concentrated disproportionately in lower and upper layers (Tang et al., 2024), whereas intermediate representations often exhibit more abstract or cross-lingually shared structure (Wendler et al., 2024; Zhao et al., 2024; Zhong et al., 2024). Cross-layer analyses identify a similar progression from language-conditioned input representations toward shared intermediate features and then language-specific decoding (Harrasse et al., 2025; Tamo et al., 2026). Related evidence appears across different levels of model organization: representations of task difficulty transfer most strongly across languages at intermediate depth (Civelli et al., 2026), Mixture-of-Experts (MoE) models exhibit their strongest cross-lingual routing alignment in middle layers (Bandarkar et al., 2026), and layer-swapping experiments show that outer layers can supply language-specific functionality while intermediate layers retain transferable task capabilities (Bandarkar et al., 2025). Neuron-level studies similarly describe multilingual inference as a transition between language-specific and shared representational regimes (Tezuka & Inoue, 2025; Zhang et al., 2026), while more recent work argues that increasingly capable models develop functionally important neurons shared across languages (Chen et al., 2026). Together, these findings motivate treating intermediate-layer language neutrality as a graded property rather than assuming the existence of a perfectly language-agnostic space.

**Multilingual Alignment and Capability.** A complementary line of work asks whether cross-lingual representational alignment is functionally related to multilingual capability. Projecting non-English representations toward English-aligned spaces can improve factual knowledge transfer (Xu et al., 2023), and multilingual alignment training can improve performance even in languages not explicitly included during alignment (Zhang et al., 2024). Deep supervision of intermediate representations likewise improves non-English performance, suggesting that the quality of internal cross-lingual transformations matters for downstream capability (Huo et al., 2025). At inference time, removing

[1] https://anonymous.4open.science/r/MELD-ICLR2027

language-specific components from hidden states can improve multilingual reasoning across a range of models and languages (Zhao et al., 2026).

Mechanistic work on factual reasoning provides a related interpretation of multilingual failures. Models may possess the relevant factual knowledge in shared internal mechanisms but fail to access, route, or express it correctly in lower-resource languages (Wang et al., 2025; Lu et al., 2025; Sun et al., 2026). At the same time, successful multilingual computation does not imply that language-specific structure should disappear entirely. Sparse-autoencoder analyses identify causally relevant language-specific features (Deng et al., 2025), while circuit-level analyses find shared computation supplemented by language-specific components when linguistic structure requires them (Zhang et al., 2025). Other work identifies latent romanization as an intermediate bridge for some non-Roman scripts (Saji et al., 2025), and multilingual reasoning can benefit from dynamically mixing languages rather than enforcing a single reasoning language (Li et al., 2025b; Weihua et al., 2026). Moreover, directly manipulating isolated language-specific neurons does not consistently improve cross-lingual transfer (Mondal et al., 2025), suggesting that multilingual capability is distributed across interacting representations and computational mechanisms.

# 3 METHODOLOGY

## 3.1 MOTIVATION

MELD is based on two observations from prior work: multilingual representations contain both shared and language-specific structure (Chang et al., 2022; Hämmerl et al., 2024), and this shared structure tends to be most prominent at intermediate model depths (Tang et al., 2024; Harrasse et al., 2025). Hence, we treat language neutrality as a graded property of the representation space rather than assuming that intermediate representations are completely language-agnostic. This motivates the central hypothesis underlying MELD:

$$\boxed{\text{stronger middle-layer language neutrality} \implies \text{stronger multilingual capability}}$$

To empirically assess this hypothesis, MELD measures how strongly two language-conditioned representation distributions are separated relative to their within-language variability. We focus this measurement on intermediate layers, where shared multilingual computation is expected to be strongest. The following sections formalize this distributional notion of language neutrality and define the resulting MELD score.

## 3.2 MELD LANGUAGE NEUTRALITY SCORE

We consider a reference-language corpus $\mathcal{D}^p = \{s_i^p\}_{i=1}^N$ and a target-language corpus $\mathcal{D}^q = \{s_i^q\}_{i=1}^N$, each containing $N$ samples. Balanced sampling is used in our main experiments for symmetry and to match the available parallel corpora, although MELD also extends to $N_p \neq N_q$ (see Appendix A.1). Here, the shared index $i$ is not assumed to indicate sentence-level correspondence. At layer $\ell$, let

$$H_\ell^a = \{\mathbf{h}_{i,\ell}^a\}_{i=1}^N, \qquad \boldsymbol{\mu}_\ell^a = \frac{1}{N}\sum_{i=1}^N \mathbf{h}_{i,\ell}^a, \qquad a \in \{p, q\}, \tag{1}$$

where $\mathbf{h}_{i,\ell}^a \in \mathbb{R}^d$ is a sentence-level hidden representation. MELD measures between-language separation relative to a shared within-language covariance geometry. We define the pooled covariance, $\Sigma_\ell$, as the average of the sample covariances $\widehat{\Sigma}_\ell^a$ for the two languages $a \in \{p, q\}$,

$$\widehat{\Sigma}_\ell^a = \frac{1}{N-1}\sum_{i=1}^N (\mathbf{h}_{i,\ell}^a - \boldsymbol{\mu}_\ell^a)(\mathbf{h}_{i,\ell}^a - \boldsymbol{\mu}_\ell^a)^\top \quad \Rightarrow \quad \Sigma_\ell = \frac{1}{2}\left(\widehat{\Sigma}_\ell^p + \widehat{\Sigma}_\ell^q\right). \tag{2}$$

Using this pooled covariance, we compute the regularized Mahalanobis distance between two representations $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ as

$$d_{\Sigma_\ell}^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^\top (\Sigma_\ell + \epsilon I)^{-1}(\mathbf{x} - \mathbf{y}), \tag{3}$$

where $\epsilon > 0$ ensures numerical stability. Under this geometry, the **between-language separation**, $D_{\text{inter}}$, and the **within-language dispersion**, $D_{\text{intra}}$, are defined as

$$D_{\text{inter}}^{(\ell)} = d_{\Sigma_\ell}^2\left(\boldsymbol{\mu}_\ell^p, \boldsymbol{\mu}_\ell^q\right), \qquad D_{\text{intra}}^{(\ell)} = \frac{1}{2}\sum_{a\in\{p,q\}} \frac{1}{N}\sum_{i=1}^{N} d_{\Sigma_\ell}^2\left(\mathbf{h}_{i,\ell}^a, \boldsymbol{\mu}_\ell^a\right). \tag{4}$$

The inter-language term quantifies the variance-normalized separation between the two language centroids. The intra-language term provides the corresponding within-language reference scale under the same geometry. The shared covariance geometry downweights centroid differences along dimensions with high within-language variability and emphasizes differences along comparatively stable dimensions. We combine these quantities into the layer-wise MELD language-neutrality score,

$$M_\ell = \frac{D_{\text{intra}}^{(\ell)}}{D_{\text{inter}}^{(\ell)} + D_{\text{intra}}^{(\ell)}}, \qquad 0 \le M_\ell \le 1. \tag{5}$$

Since $M_\ell \in [0, 1]$, it has a direct interpretation. A high value indicates that the variance-normalized separation between the reference- and target-language centroids is small relative to the corresponding within-language scale, which we interpret as stronger distributional language neutrality. Conversely, a low value indicates stronger language-conditioned separation.

### 3.3 Middle-Layer MELD Score

Across model depth, $M_\ell$ typically exhibits an inverted-U profile, as illustrated for Qwen3.5-27B in Figure 1; corresponding profiles across all evaluated models, are provided in Appendix A.4. Language neutrality generally increases toward intermediate layers and decreases toward the output. This is consistent with prior evidence for more shared multilingual representations at intermediate depths and stronger language-specific processing near model boundaries (Tang et al., 2024; Harrasse et al., 2025; Civelli et al., 2026; Zhao et al., 2026).

Therefore, we aggregate MELD language-neutrality scores over a region centered on the middle of the network, where shared multilingual processing is expected to be strongest. Let $\mathcal{M} = \{\ell_1, \ldots, \ell_K\}$ denote the selected layers and $0 = t_1 < \cdots < t_K = 1$ their normalized positions. We define the MELD score as the normalized area under the layer-wise neutrality curve, computed using the trapezoidal rule:

$$\text{MELD}(p, q) = \sum_{k=1}^{K-1} \frac{M_{\ell_k} + M_{\ell_{k+1}}}{2}(t_{k+1} - t_k). \tag{6}$$

Since $M_\ell \in [0, 1]$ and the normalized layer interval has unit length, $0 \le \text{MELD}(p, q) \le 1$. A larger score indicates weaker language-conditioned separation throughout the intermediate processing region, relative to the shared within-language covariance geometry. We use this score as an intrinsic estimate of multilingual capability.

### 3.4 Diagonal Mahalanobis Approximation

The Mahalanobis geometry above relies on the pooled covariance $\Sigma_\ell \in \mathbb{R}^{d\times d}$. However, in high-dimensional hidden spaces, estimating and inverting the full covariance is statistically challenging. Under balanced sampling, the pooled covariance is estimated from at most $2N - 2$ independent centered observations and is therefore rank deficient whenever $2N - 2 < d$. Even when it is full rank, estimating the $O(d^2)$ cross-dimensional covariance terms can have high variance when $d$ is large relative to the sample size. Therefore, we use a diagonal approximation in the default MELD implementation, retaining only the pooled marginal variances. The resulting distance is

$$d_{\widetilde{\Sigma}_\ell}^2(\mathbf{x}, \mathbf{y}) = \sum_{k=1}^{d} \frac{(x_k - y_k)^2}{v_{\ell,k} + \epsilon}, \qquad v_{\ell,k} = \frac{\sum_{a\in\{p,q\}}\sum_{i=1}^{N}(h_{i,\ell,k}^a - \mu_{\ell,k}^a)^2}{2N - 2}. \tag{7}$$

This corresponds to feature-wise variance normalization: differences along high-variance dimensions are down-weighted, while differences along more stable dimensions receive greater weight. The diagonal approximation does not assume that hidden dimensions are independent; rather, it

Table 1: Overview of the evaluated models, including parameter scale, depth, and training regime.

| | **Gemma-3** | **Gemma-3** | **Gemma-4** | **Gemma-4** | **Phi-4** | **Phi-4-mini** | **Phi-4-mini** |
|---|---|---|---|---|---|---|---|
| **#Params** | 27B | 12B | 31B | 12B | 15B | 4B | 4B |
| **#Layers** | 63 | 49 | 61 | 49 | 41 | 33 | 33 |
| **Training** | Pretrained | Pretrained | Pretrained | Pretrained | Reasoning | Reasoning | Instruct |
| | **Llama-3.1** | **Llama-3.1** | **Ministral-3** | **Ministral-3** | **Qwen3.5** | **Qwen3.5** | **Qwen3.5** |
| **#Params** | 70B | 8B | 8B | 3B | 27B | 9B | 4B |
| **#Layers** | 81 | 33 | 35 | 27 | 65 | 33 | 33 |
| **Training** | Pretrained | Pretrained | Instruct | Instruct | Pretrained | Pretrained | Pretrained |

trades cross-dimensional covariance information for more stable estimation of $d$ marginal variances instead of $O(d^2)$ covariance terms. In our ablations, we assess the effect of retaining cross-dimensional covariance by first applying PCA and then computing the full Mahalanobis distance in the reduced space; see Appendix A.2.

# 4 Experimental Setup

## 4.1 Models

We evaluate MELD on 14 open-weight LLMs spanning six model families, parameter scales from 3B to 70B, depths from 27 to 81 layers, and multiple training regimes, including pretrained, instruction-tuned, and reasoning-oriented models. Table 1 summarizes the evaluated models. This diversity allows us to assess whether MELD remains predictive across substantial differences in model family, scale, depth, and post-training strategy.

## 4.2 Representation Corpora

We use English as the reference language in all main experiments and compare it against each target language available in the corresponding representation corpus and downstream benchmark. We use FLORES+ (Burchell et al., 2024) and NTREX-128 (Federmann et al., 2022) as multilingual parallel corpora for evaluating MELD under sentence-level semantic correspondence. Specifically, we use the FLORES+ devtest split (1,012 sentences per language) and all 1,997 NTREX-128 sentences per language. Together, these corpora provide broad typological and geographic coverage for evaluating MELD. To assess MELD without sentence-level correspondence, we additionally construct an independent English corpus from five GLUE tasks (Wang et al., 2018). This corpus replaces the aligned English side of FLORES+ and NTREX-128 in the non-parallel setting, ensuring that no sentence-level semantic alignment exists between the reference and target-language corpora. Additional corpus details, sampling procedures, and language coverage are provided in Appendix A.3.

## 4.3 Multilingual Benchmarks

We evaluate multilingual capability using five standard benchmarks: m-MMLU, m-ARC, and m-HellaSwag from the Okapi corpus (Lai et al., 2023), OpenAI MMMLU, and BELEBELE (Bandarkar et al., 2024). The Okapi benchmarks are machine-translated versions of MMLU (Hendrycks et al., 2020), ARC (Clark et al., 2018), and HellaSwag (Zellers et al., 2019) covering 31 languages. OpenAI MMMLU is a professionally translated version of MMLU covering 14 languages, while BELEBELE is a multilingual reading-comprehension benchmark spanning 122 languages. All downstream benchmarks are evaluated using the EleutherAI lm-evaluation-harness (Gao et al., 2024). We use five-shot evaluation for instruction-tuned and reasoning-oriented models and zero-shot evaluation for all other models.

### 4.4 BASELINES

We compare MELD against two prior intrinsic multilingual evaluation methods: Language Ranker (LR) (Li et al., 2025c) and MEXA (Kargaran et al., 2025). LR measures cross-lingual representation similarity between English and target-language sentence pairs using cosine similarity, while MEXA formulates cross-lingual alignment as a bidirectional retrieval problem over parallel translations. For LR, we follow its evaluation procedure, evaluating every fifth layer across the full depth of the network. For MEXA, we use the authors' publicly available source code and adopt its weighted-average token pooling strategy, which yields the best overall performance in their study. Both baselines rely on sentence-level bilingual correspondence, but operationalize it differently.

### 4.5 EVALUATION METRICS

We evaluate whether intrinsic scores predict multilingual capability from two complementary perspectives. First, for *within-model language ranking*, we associate each evaluated target language with its downstream accuracy and corresponding intrinsic score for a given model and benchmark. We then compute Pearson $r$ and Spearman $\rho$ correlations across languages for each model–benchmark pair, and macro-average the resulting correlations across models for each benchmark. Pearson correlation captures agreement in the magnitude of cross-language performance variation, whereas Spearman correlation measures agreement in the relative ranking of languages. Second, for *overall model ranking*, we aggregate downstream performance across benchmarks within each model–language pair and then average across languages to obtain a single multilingual capability score per model. Intrinsic model scores are obtained by averaging each intrinsic metric across the same language set. We then compute Pearson $r$ and Spearman $\rho$ across models to measure how well each intrinsic metric predicts overall multilingual model capability.

### 4.6 HIDDEN-STATE EXTRACTION AND MELD CONFIGURATION

For all intrinsic methods, we feed each corpus sentence directly to the model, without adding prompts or task-specific templates. We obtain sentence-level representations using mean pooling over token-level hidden states for MELD, weighted mean pooling for MEXA, and the last-token representation for LR. The MELD language neutrality score is computed independently at each candidate layer according to Section 3. The final MELD score is obtained by aggregating over approximately the central 10% of layers in each model. Defining this region proportionally, rather than using fixed layer indices, allows MELD to adapt naturally to architectures of varying depths while reducing representation-extraction cost. Unlike MEXA, which requires hidden states across the full model depth, and LR, which samples regularly spaced layers throughout the network, MELD operates on only a small, proportionally defined subset of middle layers. We use the central 10% by default because it provides a favorable trade-off between computational cost and predictive correlation with downstream multilingual performance. In Section A.2.2 of the Appendix, we examine the sensitivity of MELD to this choice by progressively increasing the proportion of middle layers included in the aggregation.

## 5 RESULTS AND FINDINGS

We evaluate MELD in three complementary settings. First, we assess whether it captures cross-language variation in downstream performance within individual models. We then test whether this relationship persists when sentence-level correspondence is removed, before examining whether MELD also reflects differences in overall multilingual capability across models.

### 5.1 WITHIN-MODEL LANGUAGE RANKING

Table 2 summarizes the within-model language-ranking results using the evaluation protocol defined in Section 4.5. As indicated in this table, MELD consistently shows strong agreement with downstream performance across benchmarks and representation corpora and is generally competitive with or stronger than prior methods. The advantage is most consistent for Pearson correlation, suggesting that MELD captures not only the relative ordering of languages but also the magnitude of cross-language capability differences. Notably, these results are obtained in the parallel-data

Table 2: Within-model correlations between intrinsic scores and downstream multilingual performance using parallel corpora, macro-averaged across models. Best results are shown in bold. * denotes significance under a paired bootstrap (95% CIs; 10,000 replicates).

| | | m-MMLU | | m-ARC | | m-HellaSwag | | MMMLU | | Belebele | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Corpus** | **Method** | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ |
| FLORES | LR | 0.888 | 0.905 | 0.866 | 0.892 | 0.882 | 0.920 | 0.889 | 0.898 | 0.678 | 0.825 |
| | MEXA | 0.908 | 0.888 | 0.931 | 0.915 | 0.933 | 0.913 | 0.958 | 0.911 | 0.878 | 0.876 |
| | MELD (Ours) | **0.940** | **0.934** | **0.956** | **0.941** | **0.961** | **0.946** | **0.961** | **0.942** | **0.887** | **0.886** |
| NTREX | LR | 0.899 | 0.913 | 0.874 | 0.889 | 0.899 | 0.933 | 0.872 | 0.855 | 0.687 | 0.797 |
| | MEXA | 0.933 | **0.943** | 0.955 | **0.946** | 0.960 | **0.955** | 0.918 | 0.923 | 0.818 | 0.881 |
| | MELD (Ours) | **0.948** | 0.940 | **0.961** | 0.944 | **0.965** | 0.947 | **0.955** | **0.942** | **0.873** | **0.886** |
| Average | LR | 0.894 | 0.909 | 0.870 | 0.890 | 0.891 | 0.927 | 0.880 | 0.877 | 0.682 | 0.811 |
| | MEXA | 0.921 | 0.915 | 0.943 | 0.930 | 0.947 | 0.934 | 0.938 | 0.917 | 0.848 | 0.879 |
| | MELD (Ours) | **0.944*** | **0.937*** | **0.958*** | **0.942*** | **0.963*** | **0.946*** | **0.958*** | **0.942*** | **0.880*** | **0.886** |

setting for which pair-based methods such as MEXA or LR are explicitly designed. MELD therefore achieves comparable or stronger predictive performance without incorporating sentence-level correspondence into its scoring function.

### 5.2 Robustness to Non-Parallel Data

Table 3 evaluates whether MELD remains predictive when sentence-level correspondence between the reference- and target-language corpora is removed. In this setting, we construct an independent English reference corpus from GLUE, as described in Section 4.2. We randomly partition this corpus into five approximately equal subsets and, in each run, replace the English side of the parallel FLORES+ or NTREX-128 corpus with one of these non-parallel reference sets. We repeat the evaluation over five runs, recomputing all intrinsic scores and measuring their correlations with downstream multilingual performance in each case. Table 3 reports the mean Pearson and Spearman correlations across the five runs together with their standard deviations.

MELD retains strong agreement with downstream multilingual performance across both representation corpora and all benchmark families, with very little variation across reference subsets. MEXA degrades substantially once aligned translation pairs are removed, consistent with its retrieval-based formulation, which depends directly on sentence-level correspondence. LR also degrades, but markedly less than MEXA. This behavior is consistent with the observations in Kargaran et al. (2025), that absolute cosine similarities for parallel and non-parallel sentence pairs are themselves highly correlated, indicating that cosine similarity can retain substantial language-level structure even when sentence-level correspondence is absent. MELD remains substantially more stable than both baselines, supporting the view that its predictive signal is primarily associated with distribution-level properties of multilingual representations rather than with pairwise bilingual alignment.

### 5.3 Overall Multilingual Model Ranking

Table 4 considers a complementary question: *whether the intrinsic metrics can distinguish multilingual capability across models* rather than only across languages within a fixed model. We compute correlations between model-level downstream performance and intrinsic scores, following the evaluation protocol in Section 4.5. MELD shows the strongest agreement with aggregate downstream performance across both representation corpora. This indicates that MELD remains informative across models differing in family, scale, depth, and training regime, and extends the within-model findings by showing that its signal also reflects broader differences in multilingual capability across models.

Table 3: Within-model correlations between intrinsic scores and downstream multilingual performance using non-parallel corpora. Results are averaged over five disjoint non-parallel reference subsets and reported as mean with standard deviation. Best results are shown in bold.

| | | **Pearson $r$** | | | | |
|---|---|---|---|---|---|---|
| **Corpus** | **Method** | m-MMLU | m-ARC | m-HellaSwag | MMMLU | Belebele |
| $\text{FLORES}_{+\text{GLUE}}$ | LR | 0.717 (0.005) | 0.685 (0.005) | 0.701 (0.005) | 0.790 (0.003) | 0.450 (0.002) |
| | MEXA | 0.107 (0.102) | 0.113 (0.100) | 0.117 (0.108) | 0.150 (0.098) | 0.130 (0.050) |
| | MELD (Ours) | **0.937** (0.000) | **0.951** (0.001) | **0.957** (0.000) | **0.962** (0.000) | **0.890** (0.000) |
| $\text{NTREX}_{+\text{GLUE}}$ | LR | 0.682 (0.003) | 0.641 (0.003) | 0.659 (0.004) | 0.837 (0.001) | 0.485 (0.001) |
| | MEXA | 0.177 (0.062) | 0.200 (0.059) | 0.182 (0.061) | 0.244 (0.084) | 0.187 (0.074) |
| | MELD (Ours) | **0.942** (0.000) | **0.953** (0.000) | **0.958** (0.000) | **0.959** (0.000) | **0.872** (0.000) |
| Average | LR | 0.700 (0.019) | 0.663 (0.024) | 0.680 (0.023) | 0.814 (0.025) | 0.468 (0.019) |
| | MEXA | 0.142 (0.087) | 0.156 (0.090) | 0.149 (0.089) | 0.197 (0.100) | 0.159 (0.067) |
| | MELD (Ours) | **0.940** (0.003) | **0.952** (0.001) | **0.957** (0.001) | **0.960** (0.002) | **0.881** (0.010) |

| | | **Spearman $\rho$** | | | | |
|---|---|---|---|---|---|---|
| **Corpus** | **Method** | m-MMLU | m-ARC | m-HellaSwag | MMMLU | Belebele |
| $\text{FLORES}_{+\text{GLUE}}$ | LR | 0.700 (0.006) | 0.708 (0.006) | 0.717 (0.006) | 0.775 (0.008) | 0.623 (0.004) |
| | MEXA | 0.112 (0.124) | 0.117 (0.115) | 0.115 (0.130) | 0.129 (0.131) | 0.133 (0.092) |
| | MELD (Ours) | **0.932** (0.001) | **0.938** (0.000) | **0.944** (0.001) | **0.938** (0.000) | **0.884** (0.000) |
| $\text{NTREX}_{+\text{GLUE}}$ | LR | 0.643 (0.004) | 0.642 (0.004) | 0.656 (0.004) | 0.677 (0.006) | 0.559 (0.002) |
| | MEXA | 0.219 (0.083) | 0.234 (0.075) | 0.225 (0.074) | 0.273 (0.103) | 0.234 (0.074) |
| | MELD (Ours) | **0.933** (0.000) | **0.940** (0.001) | **0.943** (0.000) | **0.942** (0.001) | **0.876** (0.000) |
| Average | LR | 0.672 (0.030) | 0.675 (0.035) | 0.687 (0.033) | 0.726 (0.052) | 0.591 (0.034) |
| | MEXA | 0.165 (0.114) | 0.175 (0.110) | 0.170 (0.115) | 0.201 (0.135) | 0.183 (0.095) |
| | MELD (Ours) | **0.933** (0.001) | **0.939** (0.001) | **0.944** (0.001) | **0.940** (0.003) | **0.880** (0.004) |

Table 4: Overall multilingual model-ranking correlations. For each model, downstream accuracy is averaged across benchmarks within each language and then across languages, while intrinsic scores are averaged over the same language set. Best results are shown in bold.

| **Corpus** | **Method** | **Pearson $r$** | **Spearman $\rho$** |
|---|---|---|---|
| FLORES | LR | 0.419 | 0.415 |
| | MEXA | 0.694 | 0.622 |
| | MELD (Ours) | **0.800** | **0.763** |
| NTREX | LR | 0.399 | 0.389 |
| | MEXA | 0.685 | 0.626 |
| | MELD (Ours) | **0.751** | **0.741** |
| Average | LR | 0.409 | 0.402 |
| | MEXA | 0.690 | 0.624 |
| | MELD (Ours) | **0.776** | **0.752** |

## 6 Conclusion

We introduced MELD, an intrinsic metric that estimates multilingual capability from the distributional geometry of intermediate hidden states. Across 14 open-weight LLMs and five multilingual benchmark families, MELD strongly correlates with downstream performance both across languages within models and across models overall. Its predictive signal remains robust when sentence-level correspondence is removed, distinguishing it from pair-based alignment methods. These results suggest that middle-layer distributional language neutrality provides a useful and scalable signal of multilingual capability.

## AI USE STATEMENT

In this work, we used generative AI tools to assist with the formatting and presentation of mathematical equations, provide feedback on research methodology and experimental design, interpret experimental results, and format tables. We did not use generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, implement methods, or perform qualitative or thematic data analysis. The remaining required disclosure categories are not applicable to this work.

We reviewed all AI-assisted work. In particular, numerical values presented in the tables were checked against the final outputs of our code to verify their correctness and consistency. We also independently reviewed the mathematical arguments, methodological descriptions, experimental interpretations, and all AI-assisted text included in the manuscript. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

## REFERENCES

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, 2024.

Lucas Bandarkar, Benjamin Muller, Pritish Yuvraj, Rui Hou, Nayan Singhal, Hongjiang Lv, and Bing Liu. Layer swapping for zero-shot cross-lingual transfer in large language models. In *International Conference on Learning Representations*, volume 2025, pp. 51315–51338, 2025.

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Violet Peng. Multilingual routing in mixture-of-experts. In *International Conference on Learning Representations*, volume 2026, pp. 16305–16327, 2026.

Laurie Burchell, Jean Maillard, Antonios Anastasopoulos, Christian Federmann, Philipp Koehn, and Skyler Wang. Findings of the wmt 2024 shared task of the open language data initiative. In *Proceedings of the Ninth Conference on Machine Translation*, pp. 110–117, 2024.

Tyler A Chang, Zhuowen Tu, and Benjamin K Bergen. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 119–136, 2022.

Yuxin Chen, Yiran Zhao, Yang Zhang, An Zhang, Kenji Kawaguchi, Shafiq Joty, Junnan Li, Tat-Seng Chua, Michael Shieh, and Wenxuan Zhang. The emergence of abstract thought in large language models beyond any language. *Advances in Neural Information Processing Systems*, 38: 9902–9933, 2026.

Stefano Civelli, Pietro Bernardelle, Nicolò Brunello, and Gianluca Demartini. A shared geometry of difficulty in multilingual language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 796–807, 2026.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. Unveiling language-specific features in large language models via sparse autoencoders. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4563–4608, 2025.

William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the International Workshop on Paraphrasing*, 2005.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In Kabir Ahuja, Antonios Anastasopoulos, Barun Patra, Graham Neubig, Monojit Choudhury, Sandipan Dandapat, Sunayana Sitaram, and Vishrav Chaudhary (eds.), *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pp. 21–24, Online, November 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sumeval-1.4.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7560–7575, 2022.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024. URL https://zenodo.org/records/12608602.

Katharina Hämmerl, Jindřich Libovickỳ, and Alexander Fraser. Understanding cross-lingual alignment—a survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10922–10943, 2024.

Abir Harrasse, Florent Draye, Punya Syon Pandey, Zhijing Jin, and Bernhard Schölkopf. Tracing multilingual representations in llms with cross-layer transcoders. *arXiv preprint arXiv:2511.10840*, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Wenshuai Huo, Xiaocheng Feng, Yichong Huang, Chengpeng Fu, Baohang Li, Yangfan Ye, Zhirui Zhang, Dandan Tu, Duyu Tang, Yunfei Lu, et al. Enhancing non-english capabilities of english-centric large language models through deep supervision fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24185–24193, 2025.

Amir Hossein Kargaran, Ali Modarressi, Nafiseh Nikeghbal, Jana Diesner, François Yvon, and Hinrich Schütze. Mexa: Multilingual evaluation of english-centric llms via cross-lingual alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 27001–27023, 2025.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 318–327, 2023.

Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. Exploring multilingual probing in large language models: A cross-language analysis. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pp. 61–70, 2025a.

Yihao Li, Jiayi Xin, Miranda Muqing Miao, Qi Long, and Lyle Ungar. The impact of language mixing on bilingual llm reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 32519–32536, 2025b.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 28186–28194, 2025c.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. Paths not taken: Understanding and mending the multilingual factual recall pipeline. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15077–15107, 2025.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pp. 46–62, 2025.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pp. 2383–2392, 2016.

Alan Saji, Jaavid Aktar Husain, Thanmay Jayakumar, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. Romanlens: The role of latent romanization in multilinguality in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26410–26429, 2025.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. Natural language understanding with the quora question pairs dataset. *arXiv preprint arXiv:1907.01041*, 2019.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*, pp. 1631–1642, 2013.

Chenghao Sun, Zhen Huang, Yonggang Zhang, Xinmei Tian, Xu Shen, and Jieping Ye. Bridging the language gap: Uncovering and aligning shared circuits for multi-hop reasoning in multilingual llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33091–33099, 2026.

J Ben Tamo, Daniel Carlander-Reuterfelt, Jonathan Rubin, Dezhi Hong, Mingxian Wang, and Oleg Poliannikov. Linguamap: Which layers of llms speak your language and how to tune them? *arXiv preprint arXiv:2601.20009*, 2026.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 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)*, pp. 5701–5715, 2024.

Hinata Tezuka and Naoya Inoue. The transfer neurons hypothesis: An underlying mechanism for language latent space transitions in multilingual llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31730–31780, 2025.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5075–5094, 2025.

Shaobo Wang, Cong Wang, Wenjie Fu, Yue Min, Mingquan Feng, Isabel Guan, Xuming Hu, Conghui He, Cunxiang Wang, Kexin Yang, et al. Rethinking llm evaluation: Can we evaluate llms with 200× less data? In *International Conference on Learning Representations*, volume 2026, pp. 109852–109870, 2026.

Zheng Weihua, Xin Huang, Zhengyuan Liu, Tarun Kumar Vangani, Bowei Zou, Xiyan Tao, Yuhao Wu, AiTi Aw, Nancy F Chen, and Roy Ka-Wei Lee. Adamcot: Rethinking cross-lingual factual reasoning through adaptive multilingual chain-of-thought. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33863–33871, 2026.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 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)*, pp. 15366–15394, 2024.

Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers)*, pp. 1112–1122, 2018.

Zhihui Xie, Handong Zhao, Tong Yu, and Shuai Li. Discovering low-rank subspaces for language-agnostic multilingual representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5617–5633, 2022.

Shaoyang Xu, Junzhuo Li, and Deyi Xiong. Language representation projection: Can we transfer factual knowledge across languages in multilingual language models? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 3692–3702, 2023.

Zixiang Xu, Yanbo Wang, Yue Huang, Xiuying Chen, Jieyu Zhao, Meng Jiang, and Xiangliang Zhang. Cross-lingual pitfalls: Automatic probing cross-lingual weakness of multilingual large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8254–8284, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. The same but different: Structural similarities and differences in multilingual language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 95372–95392, 2025.

Shimao Zhang, Changjiang Gao, Wenhao Zhu, Jiajun Chen, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. Getting more from less: Large language models are good spontaneous multilingual learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8037–8051, 2024.

Shimao Zhang, Zhejian Lai, Xiang Liu, Shuaijie She, Xiao Liu, Yeyun Gong, Shujian Huang, and Jiajun Chen. How does alignment enhance llms' multilingual capabilities? a language neurons perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34799–34807, 2026.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, et al. When less language is more: Language-reasoning disentanglement makes llms better multilingual reasoners. *Advances in Neural Information Processing Systems*, 38:38608–38642, 2026.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. How do large language models handle multilingualism? *Advances in Neural Information Processing Systems*, 37:15296–15319, 2024.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*, 2024.

# A Appendix

## A.1 Additional Methodological Details

This section provides additional methodological details for MELD, including the balanced-sampling assumption used in our experiments, its extension to unequal corpus sizes, and the analytical behavior of the within-language normalization.

### A.1.1 Balanced Sampling and Unequal Corpus Sizes

**Balanced sampling.** All reported experiments use balanced sampling, with $N_p = N_q = N$. This includes the non-parallel setting, where the independently sampled English reference corpus is subsampled to match the size of the corresponding target-language corpus. Besides providing a symmetric comparison, balanced sampling makes the pooled covariance in the main formulation equivalent to the conventional pooled within-language covariance:

$$\Sigma_\ell = \frac{(N-1)\widehat{\Sigma}_\ell^p + (N-1)\widehat{\Sigma}_\ell^q}{2N-2} = \frac{1}{2}\left(\widehat{\Sigma}_\ell^p + \widehat{\Sigma}_\ell^q\right). \tag{8}$$

**Extension to unequal corpus sizes.** MELD does not require equal numbers of samples. Let $\mathcal{D}^a = \{s_i^a\}_{i=1}^{N_a}$ denote the corpus for language $a \in \{p, q\}$, containing $N_a$ sentences. At layer $\ell$, the corresponding centroid and sample covariance are

$$\boldsymbol{\mu}_\ell^a = \frac{1}{N_a} \sum_{i=1}^{N_a} \mathbf{h}_{i,\ell}^a, \qquad \widehat{\Sigma}_\ell^a = \frac{1}{N_a - 1} \sum_{i=1}^{N_a} (\mathbf{h}_{i,\ell}^a - \boldsymbol{\mu}_\ell^a)(\mathbf{h}_{i,\ell}^a - \boldsymbol{\mu}_\ell^a)^\top, \tag{9}$$

where $\mathbf{h}_{i,\ell}^a \in \mathbb{R}^d$ is a sentence-level hidden representation. A conventional observation-weighted pooled covariance is

$$\Sigma_\ell^{\text{pool}} = \frac{(N_p - 1)\widehat{\Sigma}_\ell^p + (N_q - 1)\widehat{\Sigma}_\ell^q}{N_p + N_q - 2}. \tag{10}$$

This estimator gives greater influence to the language with more observations. However, for MELD, the two language-conditioned distributions are the objects being compared. We therefore retain equal weighting between languages and define

$$\Sigma_\ell = \frac{1}{2}\left(\widehat{\Sigma}_\ell^p + \widehat{\Sigma}_\ell^q\right). \tag{11}$$

This preserves equal nominal contribution from each language regardless of corpus size. Then, the corresponding within-language dispersion is

$$D_{\text{intra}}^{(\ell)} = \frac{1}{2} \sum_{a \in \{p,q\}} \frac{1}{N_a} \sum_{i=1}^{N_a} d_{\Sigma_\ell}^2 \left(\mathbf{h}_{i,\ell}^a, \boldsymbol{\mu}_\ell^a\right), \tag{12}$$

where $d_{\Sigma_\ell}^2$ has been defined in Equation 3. Therefore, both the covariance geometry and the within-language averaging give equal nominal importance to the two languages.

### A.1.2 Behavior of the Within-language Normalization

The use of the same covariance geometry in the Mahalanobis distance and within-language dispersion gives $D_{\text{intra}}^{(\ell)}$ a useful analytical form. Let $G_\ell = (\Sigma_\ell + \epsilon I)^{-1}$, under balanced sampling,

$$D_{\text{intra}}^{(\ell)} = \frac{N-1}{N} \operatorname{tr}(G_\ell \Sigma_\ell) = \frac{N-1}{N} \sum_{k=1}^{d} \frac{\lambda_{\ell,k}}{\lambda_{\ell,k} + \epsilon}, \tag{13}$$

where $\lambda_{\ell,k}$ are the eigenvalues of $\Sigma_\ell$. Defining the regularized effective dimensionality, $d_{\text{eff}}$, as

$$d_{\text{eff}}^{(\ell)}(\epsilon) = \sum_{k=1}^{d} \frac{\lambda_{\ell,k}}{\lambda_{\ell,k} + \epsilon} \quad \Rightarrow \quad D_{\text{intra}}^{(\ell)} = \frac{N-1}{N} d_{\text{eff}}^{(\ell)}(\epsilon), \tag{14}$$

shows that the within-language term primarily provides a reference scale determined by the effective dimensionality of the representation space. For the diagonal approximation used by default in MELD,

$$D_{\text{intra}}^{(\ell)} = \frac{N-1}{N} \sum_{k=1}^{d} \frac{v_{\ell,k}}{v_{\ell,k} + \epsilon} \approx \frac{N-1}{N} d. \tag{15}$$

Thus, most language-dependent variation in MELD arises from the variance-normalized between-language separation, while $D_{\text{intra}}^{(\ell)}$ provides a reference scale that is approximately determined by the hidden dimension $d$. For sufficiently large $N$, substituting $D_{\text{intra}}^{(\ell)} \approx d$ into the layer-wise MELD neutrality score gives

$$M_\ell \approx \frac{1}{1 + D_{\text{inter}}^{(\ell)}/d}. \tag{16}$$

Hence, in this regime, MELD can be interpreted as a monotonic transformation of the variance-normalized between-language separation relative to the hidden dimension. For completeness, under unequal sample sizes define,

$$\alpha_a = 1 - \frac{1}{N_a}, \qquad T_a^{(\ell)} = \operatorname{tr}\left[(\Sigma_\ell + \epsilon I)^{-1} \widehat{\Sigma}_\ell^a\right], \qquad a \in \{p, q\}. \tag{17}$$

Then

$$D_{\text{intra}}^{(\ell)} = \frac{1}{2}\left(\alpha_p T_p^{(\ell)} + \alpha_q T_q^{(\ell)}\right), \qquad T_p^{(\ell)} + T_q^{(\ell)} = 2d_{\text{eff}}^{(\ell)}(\epsilon), \tag{18}$$

which implies

$$\min(\alpha_p, \alpha_q) d_{\text{eff}}^{(\ell)}(\epsilon) \le D_{\text{intra}}^{(\ell)} \le \max(\alpha_p, \alpha_q) d_{\text{eff}}^{(\ell)}(\epsilon). \tag{19}$$

Therefore, once both $N_p$ and $N_q$ are sufficiently large, $D_{\text{intra}}^{(\ell)} \approx d_{\text{eff}}^{(\ell)}(\epsilon)$, and differences in corpus size have only a small effect on the normalization.

### A.2 Ablation Studies

We conduct two ablation studies to examine key design choices in MELD. First, we evaluate whether retaining cross-dimensional covariance through PCA-based dimensionality reduction changes predictive performance relative to the default diagonal Mahalanobis approximation. Second, we assess the sensitivity of MELD to the fraction of middle layers used for aggregation. Together, these analyses test the robustness of MELD to its covariance approximation and layer-selection strategy.

#### A.2.1 Full-Covariance Mahalanobis Distance via PCA

To examine the effect of the diagonal covariance approximation, we compare the default MELD formulation with variants that apply PCA before computing the full Mahalanobis distance in a reduced representation space. We retain either 256 or 512 principal components and evaluate the resulting scores using the same within-model language-ranking protocol as in Table 2. We choose 256 and 512 as moderate reduced dimensionalities that remain well below the available sample size, improving the stability of full-covariance estimation while retaining substantial representational structure.

Table 5 shows that both PCA-based variants retain strong correlations with downstream multilingual performance. However, the default diagonal approximation achieves the highest average correlations across benchmarks and corpora. The PCA variants are particularly competitive on FLORES+, where PCA-512 slightly improves upon the diagonal formulation on selected benchmarks. On NTREX-128, however, both PCA variants yield lower average correlations. Increasing the number of retained components from 256 to 512 does not consistently improve performance. These results suggest that the default diagonal approximation provides a favorable practical trade-off: it avoids PCA-based dimensionality reduction and full-covariance estimation while achieving stronger predictive performance on average.

#### A.2.2 Sensitivity to the Middle-Layer Fraction

The default MELD configuration aggregates language-neutrality scores over the central 10% of model layers. To assess sensitivity to this choice, we vary the fraction of layers included in the aggregation while keeping the selected region centered at the midpoint of the network. For each fraction, we recompute MELD and evaluate the same within-model language-ranking correlations reported in Table 2, averaging Pearson $r$ and Spearman $\rho$ across models for FLORES+ and NTREX-128.

Figure 2 shows that MELD is relatively insensitive to the precise size of the middle-layer region over a broad range of settings. Correlations remain strong as the fraction is increased beyond the default 10%, with modest improvements for some intermediate fractions. However, performance begins to decline when a large proportion of the network is included, particularly as the aggregation approaches the full model depth. This pattern is consistent with the motivation for MELD: intermediate layers exhibit the strongest shared multilingual structure, whereas layers closer to the input and output boundaries contain more language-specific information.

Therefore, we use the central 10% of layers as the default setting. It achieves predictive performance close to the best observed across the tested fractions while requiring hidden-state extraction from only a small subset of the network, providing a favorable trade-off between predictive performance and computational cost.

Table 5: Effect of retaining cross-dimensional covariance through PCA-based dimensionality reduction. We compare the default diagonal approximation with full-covariance, computed after projection to 256 or 512 principal components. Values are within-model Pearson $r$ and Spearman $\rho$ correlations with downstream multilingual performance, averaged across models. Best results are shown in bold.

| Corpus | $\text{MELD}_{\text{variant}}$ | m-MMLU | | m-ARC | | m-HellaSwag | | MMMLU | | BELEBELE | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ | $r$ | $\rho$ |
| FLORES | Diagonal | 0.940 | 0.934 | 0.956 | 0.941 | 0.961 | 0.946 | 0.961 | 0.942 | 0.887 | 0.886 | **0.941** | **0.930** |
| | PCA-256 | 0.924 | 0.932 | 0.947 | 0.941 | 0.950 | 0.944 | 0.933 | 0.934 | 0.829 | 0.880 | 0.917 | 0.926 |
| | PCA-512 | 0.944 | 0.937 | 0.958 | 0.945 | 0.961 | 0.953 | 0.927 | 0.923 | 0.827 | 0.878 | 0.923 | 0.927 |
| NTREX | Diagonal | 0.948 | 0.940 | 0.961 | 0.944 | 0.965 | 0.947 | 0.955 | 0.942 | 0.873 | 0.886 | **0.940** | **0.932** |
| | PCA-256 | 0.926 | 0.934 | 0.947 | 0.942 | 0.954 | 0.952 | 0.902 | 0.901 | 0.763 | 0.860 | 0.898 | 0.917 |
| | PCA-512 | 0.919 | 0.918 | 0.940 | 0.931 | 0.944 | 0.942 | 0.892 | 0.881 | 0.762 | 0.846 | 0.891 | 0.903 |
| Average | Diagonal | 0.944 | 0.937 | 0.959 | 0.943 | 0.963 | 0.947 | 0.958 | 0.942 | 0.880 | 0.886 | **0.941** | **0.931** |
| | PCA-256 | 0.925 | 0.933 | 0.947 | 0.942 | 0.952 | 0.948 | 0.918 | 0.917 | 0.796 | 0.870 | 0.908 | 0.922 |
| | PCA-512 | 0.932 | 0.928 | 0.949 | 0.938 | 0.952 | 0.948 | 0.909 | 0.902 | 0.794 | 0.862 | 0.907 | 0.915 |

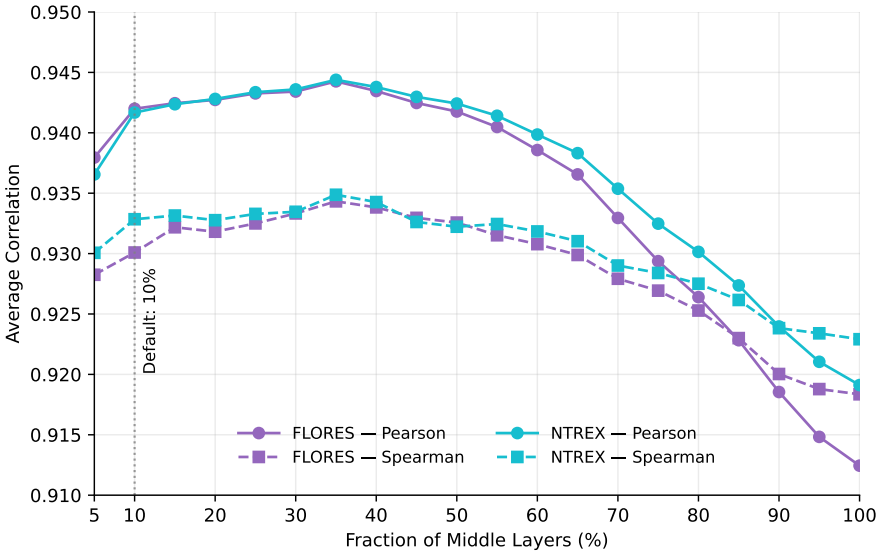


Figure 2: Sensitivity of MELD to the fraction of middle layers used for aggregation. Curves report the within-model language-ranking correlations, averaged across models and benchmark families for FLORES and NTREX. The selected region is centered at the midpoint of each model, and the vertical dotted line marks the default 10% setting.

### A.3 Datasets and Evaluation Details

This section provides additional details on the representation corpora, non-parallel reference data, downstream multilingual benchmarks, and language coverage used in our experiments.

#### A.3.1 Parallel Representation Corpora

We use two multilingual parallel corpora, FLORES+ (Burchell et al., 2024) and NTREX-128 (Federmann et al., 2022), to construct language-conditioned representation sets under sentence-level semantic correspondence. FLORES+ contains professionally translated, multilingual-aligned sentences drawn from sources including Wikinews, Wikijunior, and Wikivoyage. The corpus provides 997 sentences in the dev split and 1,012 in the devtest split. We use the 1,012-sentence devtest split for all FLORES+ experiments. NTREX-128 is a multilingual machine-translation evaluation corpus consisting of 123 news documents containing 1,997 sentences, professionally translated from English into 128 target languages. We use all 1,997 sentences per language.

Table 6: Language coverage used in our experiments. A check mark indicates that the language was selected for the corresponding resource. Evaluations use the intersection of the relevant representation-corpus and benchmark language sets.

| Language | FLORES | NTREX | m-MMLU | m-ARC | m-HellaSwag | MMMLU | Belebele |
|---|---|---|---|---|---|---|---|
| Afrikaans (afr) | ✓ | ✓ | | | | | ✓ |
| Arabic (arb) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Bengali (ben) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Catalan (cat) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Czech (ces) | ✓ | ✓ | | | | | ✓ |
| Chinese (cmn) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Danish (dan) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| German (deu) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| English (eng) | ✓ | ✓ | | | | | |
| Basque (eus) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| French (fra) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Gujarati (guj) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Hindi (hin) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Croatian (hrv) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Hungarian (hun) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Armenian (hye) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Indonesian (ind) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Italian (ita) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Japanese (jpn) | ✓ | ✓ | | | | ✓ | ✓ |
| Kannada (kan) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Korean (kor) | ✓ | ✓ | | | | ✓ | ✓ |
| Malayalam (mal) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Marathi (mar) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Dutch (nld) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Nepali (npi) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Portuguese (por) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Romanian (ron) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Russian (rus) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Slovak (slk) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Spanish (spa) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Serbian (srp) | ✓ | | ✓ | ✓ | ✓ | | |
| Swedish (swe) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Swahili (swh) | ✓ | ✓ | | | | ✓ | ✓ |
| Tamil (tam) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Telugu (tel) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Thai (tha) | ✓ | ✓ | | | | | ✓ |
| Ukrainian (ukr) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Urdu (urd) | ✓ | ✓ | | | | | ✓ |
| Vietnamese (vie) | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Wolof (wol) | ✓ | ✓ | | | | | ✓ |
| Yoruba (yor) | ✓ | ✓ | | | | ✓ | ✓ |
| Zulu (zul) | ✓ | ✓ | | | | | ✓ |
| **No. languages** | **42** | **41** | **31** | **31** | **31** | **14** | **40** |

Therefore, both resources provide sentence-aligned multilingual data, allowing MELD and the pair-based baselines to be evaluated under the same parallel-data setting. The languages from each corpus included in our experiments are listed in Table 6.

### A.3.2 Non-Parallel Reference Corpus

To evaluate MELD without sentence-level semantic correspondence, we construct an independent English reference corpus from five GLUE tasks (Wang et al., 2018). We use the full SST-2 test set (Socher et al., 2013) (1,821 sentences) and the full MRPC test set (Dolan & Brockett, 2005) (1,725 sentences). For the larger GLUE tasks, we randomly sample 2,000 sentences from QQP (Sharma et al., 2019) and 2,500 sentences each from QNLI (Rajpurkar et al., 2016) and MNLI (Williams et al.,

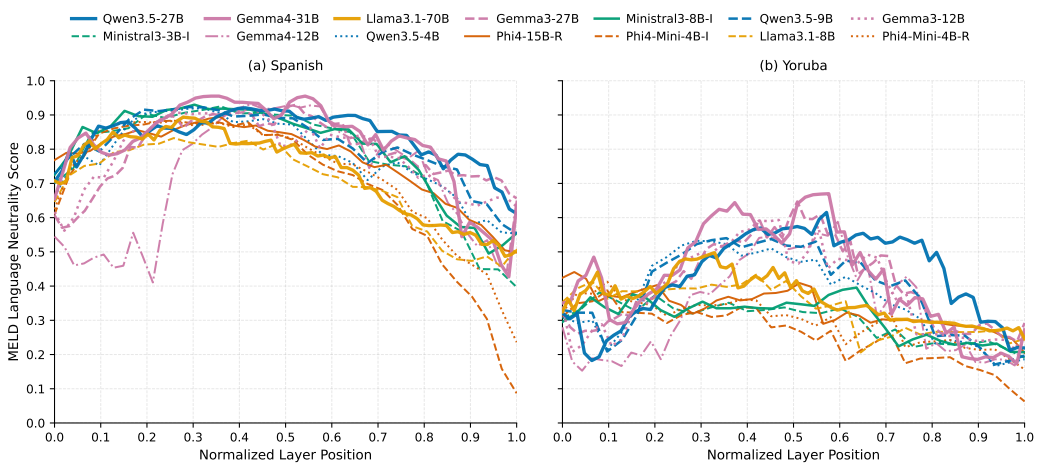


Figure 3: Layer-wise MELD language-neutrality profiles across all evaluated models on NTREX-128 for Spanish and Yoruba, representing higher- and lower-resource languages, respectively. Layer positions are normalized by model depth to enable comparison across architectures. Models are ordered by their overall downstream multilingual performance, from stronger to weaker.

2018). These samples are combined into a single English pool that is unrelated to the FLORES+ and NTREX-128 target-language sentences.

Then, we randomly partition this pool into five approximately equal, disjoint reference subsets. For each evaluation run, we further randomly subsample the selected reference subset to match the number of target-language sentences: 1,012 sentences for FLORES+ and 1,997 for NTREX-128. This preserves the balanced-sampling setting used by MELD while ensuring that the English reference and target-language corpora contain no sentence-level semantic correspondence. We repeat the evaluation using each of the five reference subsets, recomputing the intrinsic scores in every run. Table 3 reports the mean and standard deviation of the resulting correlations with downstream multilingual performance.

### A.3.3 Multilingual Benchmarks

We evaluate downstream multilingual capability using five benchmark families: m-MMLU, m-ARC, and m-HellaSwag from the Okapi suite (Lai et al., 2023), OpenAI MMMLU, and BELEBELE (Bandarkar et al., 2024). The three Okapi benchmarks are machine-translated multilingual versions of MMLU (Hendrycks et al., 2020), ARC (Clark et al., 2018), and HellaSwag (Zellers et al., 2019), each covering 31 languages. m-MMLU evaluates knowledge and reasoning across a broad range of academic and professional domains, m-ARC focuses on grade-school science reasoning, and m-HellaSwag evaluates commonsense inference. OpenAI MMMLU is a professionally translated version of MMLU covering 14 languages. BELEBELE is a parallel multilingual reading-comprehension benchmark spanning 122 language variants.

All downstream evaluations are performed using the EleutherAI `lm-evaluation-harness` (Gao et al., 2024). For multiple-choice benchmarks, candidate answers are scored using their model log-likelihood and accuracy is used as the downstream performance measure. We use five-shot evaluation for instruction-tuned and reasoning-oriented models and zero-shot evaluation for all other models.

### A.3.4 Language Coverage

For each downstream benchmark, we evaluate only those target languages that are present in both the representation corpus and the benchmark. Thus, the language set used for a given corpus–benchmark pair is the intersection of their available languages. English is used as the reference language and is not included as a target language in the within-model language-ranking evaluation.

Table 6 summarizes the resulting language coverage for each representation corpus and downstream benchmark. The reported counts therefore correspond to the languages used in our experiments rather than the full language coverage of the original resources.

### A.4 Additional Layer-Wise Results

Figure 3 extends the layer-wise analysis in Figure 1 to all evaluated models. We show profiles for Spanish and Yoruba as representative higher- and lower-resource languages, respectively. Across architectures, MELD generally exhibits greater language neutrality in intermediate layers than near the model boundaries, although the absolute level and shape of the profile vary across models and languages. This supports the use of a proportionally defined middle-layer region rather than fixed layer indices across architectures.