From Representation to Choice: Tracing Decision Emergence Across Languages in LLMs

Divyanshu Bhatt[¶*] **Abhinav Joshi**[◇*] **Ujjaval Patel**[◇] **Ashutosh Modi**[◇]
[◇]Indian Institute of Technology Kanpur (IIT Kanpur)
[¶]Samsung Research & Development Institute, Bangalore
divyanshu.bh@samsung.com,
{ajoshi,ashutoshm}@cse.iitk.ac.in

Abstract

Multilingual large language models (LLMs) can answer questions in many languages, but how they internally reason across languages remains poorly understood. In this work, we study multilingual reasoning through a decision-making perspective to investigate how multilingual reasoning unfolds in multilingual LLMs using aligned multiple-choice questions from the mMMLU benchmark. By formulating a controlled setup, presenting the same question in different languages, and tracking the model's decision from the first token to the final answer choice, we can directly compare how reasoning trajectories evolve across languages. We first demonstrate that, at the representation level, different languages share highly similar activation spaces; however, subtle divergences emerge as decisions propagate through the transformer layers. We then model answer selection as a stepwise trajectory, revealing where language-specific signals arise. These patterns are further confirmed by quantifying deviations along these trajectories, highlighting layers where multilingual processing deviates or converges. Our work provides a controlled, layer-resolved view of multilingual reasoning, shedding light on how LLMs balance shared conceptual understanding with language-specific decision-making.

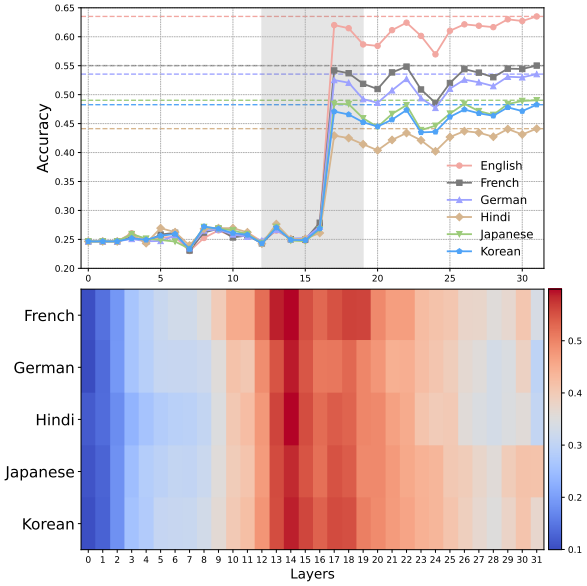


Figure 1: The figure shows the performance (Accuracy) of the Llama-3 model on the MMLU dataset in multiple languages. We observe that the model performance remains near-random (25%, 4-options) for initial layers and starts to rise from layer 16 and saturates at layer 17, with minor changes in the 17-32 layers. However, the change varies across different languages, highlighting the model's use of different spaces for the same knowledge in different languages.

1 Introduction

Recent advances in large language models (LLMs) have made them highly multilingual, with strong performance across dozens of languages on tasks ranging from factual question answering to complex instruction following (Touvron et al., 2023; OpenAI et al., 2024; Gemini et al., 2024; Grattafiori et al., 2024). Despite these advances, the performance still varies across languages, and a fundamental question remains largely unexplored (Resck et al., 2025): how do these models internally reason when presented with semantically equivalent inputs in different languages? Understanding this is not only imperative for interpretability but also has implications for improving performance in low-resource languages, diagnosing potential biases, and guiding the design of more robust multilingual systems. Prior work on understanding multilingual reasoning (Wendler et al., 2024; Dumas et al., 2025; Lindsey et al., 2025; Lu et al., 2025) has shown that multilingual LLMs often encode inputs into shared conceptual representations, and intermediate reasoning can occur through English, even when the input is in another language. However, these studies largely focus on free-form generation or factual recall, leaving structured reasoning

[*]Equal Contributions

Findings of the Association for Computational Linguistics: ACL 2026, pages 42423–42442

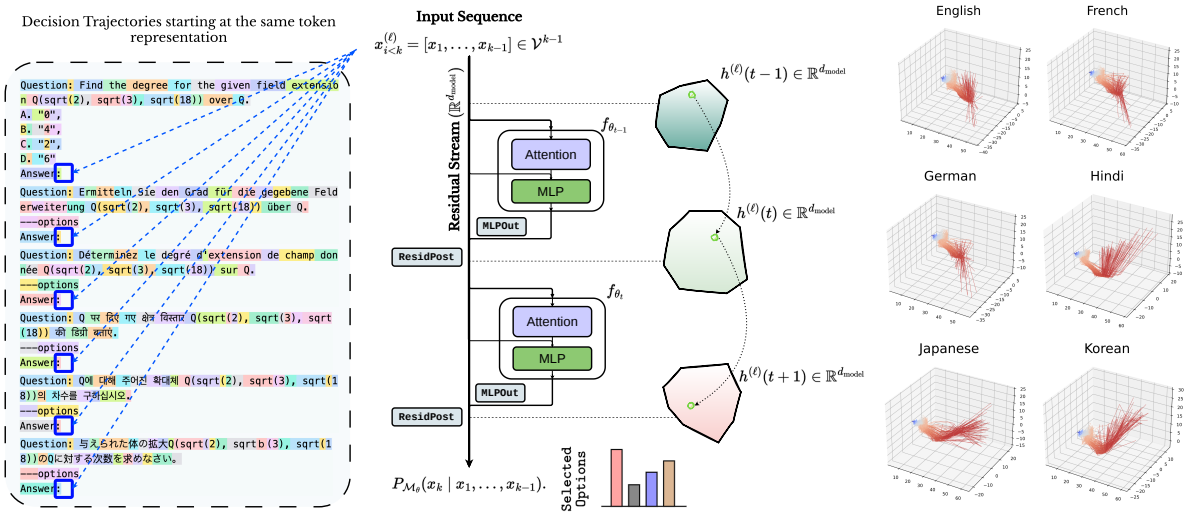


Figure 2: Multilingual transformers show low-dimensional decision trajectories across layers. We analyze how the representation of a decision token evolves through transformer layers for semantically identical multiple-choice questions translated into different languages. All inputs begin with the same embedding (blue box with the ":" token), allowing direct comparison of internal computations. As the representation passes through stacked attention and MLP blocks, it forms a trajectory in representation space towards the final prediction. The right panel visualizes these trajectories across layers for multiple languages, illustrating both low-dimensional structure and divergence in internal reasoning. We use these to understand language-specific paths that converge toward similar answer regions.

tasks such as multiple-choice question answering (MCQA) largely unexplored. MCQA provides a controlled and interpretable setting in which the model must select a discrete answer from a fixed set of options (Robinson et al., 2023; Wiegreffe et al., 2025; Joshi et al., 2025a,b,c). This makes it possible to isolate decision-making dynamics from confounding factors such as token frequency, fluency, or stylistic variation.

To investigate multilingual reasoning in this controlled context, we leverage the Multilingual MMLU (mMMLU) dataset[1] (a multilingual extension of the MMLU dataset Hendrycks et al. (2020)), which contains aligned questions in 14 languages. Each question preserves the same semantic content while varying in surface/language form, enabling a direct, language-to-language comparison of internal representations and decision-making trajectories. This setup allows us to ask precise questions: **1)** ***Do*** internal representations converge toward shared conceptual spaces early in processing? **2)** ***At what point*** do language-specific divergences emerge in the decision process? And **3)** ***How*** do these divergences shape the final answer selection? Importantly, our goal is not to compare static representational similarity (Wendler et al., 2024; Zeng et al., 2025; Wu et al., 2025b), but to analyze the dynamics of how decisions are formed layer-by-layer, focusing on when and how cross-lingual predictability breaks down. We adopt a stepwise view of decision-making within the model. Rather than observing only the final output, we track the evolution of token probabilities and hidden representations for the next token prediction, going from the first transformer block to the final answer choice in the last transformer block, reconstructing the layer-wise trajectory of the decision-making process. This allows us to identify the layers where the model's decision begins to stabilize/commit, and to compare these trajectories across languages. Preliminary observations from Fig. 1 show that performance is near-random in early layers, rises sharply around layer 16, and saturates by layer 17. Notably, though the timing of the sudden jump is the same, the magnitude of this rise varies across languages, suggesting that while knowledge is encoded in a largely shared space, the process of transforming knowledge into a discrete decision is strongly language-specific.

Building on these observations, we adopt a mechanistic, layer-resolved view to understand how representations are changed in multilingual models. By examining hidden representations, token probabilities, and cross-lingual trajectory metrics, we are

[1]Available at https://huggingface.co/datasets/openai/MMMLU

not only able to quantify language-specific divergences in decision-making across transformer layers but also compare trajectories across languages, revealing how shared representations give way to language-specific decision paths in stacked transformer layers. These analyses provide several key insights. First, starting from the same token embedding at Layer 0 (as explained in Fig. 2, also see prompt templates in App. F, where the last token ":" is the same for all the languages), the representations proceed and suddenly shift towards the actual performance at specific layers 16-17. Second, as the decision proceeds from initial representation to next token prediction, layers show representational divergence, with language-specific patterns modifying the model toward its final decision. Third, even when inputs are semantically identical, the decision-making trajectories are distinct, demonstrating that multilingual LLMs do not simply translate knowledge but actively process it in a language-dependent manner. Our contributions are threefold

- We introduce a layer-resolved, stepwise methodology for analyzing decision emergence in multilingual LLMs using MCQA.
- We provide a rigorous empirical analysis on the mMMLU benchmark, showing how decision trajectories diverge across languages and identifying the layers where language-specific signals emerge.
- We propose a stochastic view for understanding/comparing representation trajectories to quantify decisiveness and cross-lingual similarity, formulating a principled method for studying multilingual reasoning dynamics.

We hope this work provides a different perspective on comparing multilingual decision trajectories, revealing how LLMs balance shared conceptual spaces with language-specific decision-making. By establishing a setup to trace decisions layer by layer, we pave the way for deeper studies of cross-lingual reasoning, performance improvement in low-resource languages, and interpretable multilingual LLM design. We release the codebase for the experiments on https://github.com/Exploration-Lab/Representation-to-Choice.

2 Related Work

Multilingual LLMs have been the subject of extensive study in recent years, with researchers investigating both their cross-lingual capabilities (Huang et al., 2023; Lai et al., 2024; Wu et al., 2025a) and the internal mechanisms that support multilingual reasoning (Wang et al., 2025; Maraia et al., 2026; Zhang et al., 2025; Fierro et al., 2025; Sundar et al., 2025; Bandarkar et al., 2026). Prior work has established that many multilingual LLMs rely on shared latent spaces to encode semantically equivalent information across languages (Schut et al., 2025; Etxaniz et al., 2024; Fierro et al., 2025; Wilie et al., 2025; Lim et al., 2026). For instance, Wendler et al. (2024) shows that LLaMA models internally traverse an English-aligned conceptual space even when processing non-English inputs, suggesting that intermediate reasoning may partially proceed through pretraining-dominated language (English). Similarly, Zeng et al. (2025) demonstrates that LLMs map semantically identical sentences from different languages into a "Lingua Franca", a shared semantic latent space whose alignment strengthens with increased model size. These findings are further supported by region-level analyses (Zhang et al., 2024), which identify core linguistic regions in LLMs responsible for maintaining cross-lingual competence, as well as distinct monolingual subregions whose perturbation significantly affects language-specific performance.

Apart from shared representations, there is growing evidence that multilingual LLMs leverage language-specific neurons and layerwise processing dynamics to fine-tune decisions in each language. Kojima et al. (2024) identify neurons that are selectively active for individual languages, concentrated in the first and last layers, and show that even minimal interventions on these neurons can drastically alter generation probabilities in the target language. Gurgurov et al. (2025) extend these observations by systematically identifying and manipulating neurons, highlighting the role of language-specific components/circuits in modulating model outputs. However, other studies, such as Mondal et al. (2025), caution that language-specific neurons alone may be insufficient for effective cross-lingual transfer, especially in low-resource languages, indicating that shared conceptual spaces remain critical for generalization. Mechanistic analyses also reveal that language-specific divergences emerge in later layers, where decisions crys-

tallize/form along language-dependent trajectories (Nie et al., 2025). Several surveys and broader analyses contextualize these findings within the larger multilingual LLM landscape. Resck et al. (2025) provides a comprehensive overview of explainability and interpretability methods for multilingual LLMs, categorizing existing work by task, language, and modeling technique, and highlighting persistent challenges such as English-centric biases and cross-lingual performance disparities. Complementing these, Zhao et al. (2024) formalize the MWork hypothesis, which suggests that multilingual LLMs encode knowledge in a shared conceptual space, then project through English-aligned representations before decoding into the target language, a workflow that resonates with mechanistic observations from stepwise decision analyses. Our work builds on this foundation by framing a layer-resolved, trajectory-based view of multilingual decision-making in multiple-choice tasks. Unlike prior studies that focus primarily on free-form generation or factual recall, we examine how LLMs incrementally construct decisions in mMMLU (Hendrycks et al., 2020), isolating the emergence of language-specific divergences while maintaining a shared conceptual backbone. By integrating insights from shared latent spaces, language-specific neurons, and mechanistic interventions, our analysis situates the layerwise crystallization of decisions within the broader context of multilingual reasoning, providing a framework that complements and extends the current literature.

3 Background

Transformer-Based Language Models A language model defines a conditional distribution over sequences of discrete tokens drawn from a vocabulary $\mathcal{V}$. In this work, we focus on transformer-based, decoder-only language models trained in an autoregressive manner, which underlie most contemporary large language models (Radford et al., 2019; OpenAI et al., 2024; Gemini et al., 2024). Given a sequence of tokens $x = [x_1, \ldots, x_{k-1}] \in \mathcal{V}^{k-1}$, a language model $\mathcal{M}_\theta$ parameterized by θ computes a probability distribution over the next token $P_{\mathcal{M}_\theta}(x_k \mid x_1, \ldots, x_{k-1})$. Internally, transformer-based language models consist of a stack of T transformer blocks, each of which reads from and writes onto a shared *residual stream* (Elhage et al., 2021). Denoting the residual stream representation for the whole sequence at layer t as $\mathbf{h}(t) \in \mathbb{R}^{k \times d_{\text{model}}}$ and $h(t)$ corresponding to the last token, the computation proceeds as a sequence of transformations $h(t+1) = h(t) + f_{\theta_t}(\mathbf{h}(t))$, where f_{θ_t} represents the transformation performed by the t^{th} transformer block (also see Fig. 2). After the final transformer block, the residual stream is normalized and projected onto the vocabulary using an unembedding matrix $\mathbf{W}_U \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$, producing logits over the vocabulary. Applying a softmax yields the final probability distribution over the next token. We use this as a basis for an Itô-process-inspired view of layer-wise representation dynamics (explained in more detail in App. A).

4 Experimental Setup

Our goal is to study *how and when* a multilingual large language model forms (or goes towards) a discrete decision during inference, and how this process differs when the same semantic content is expressed in different languages. To isolate decision-making from surface-level linguistic variation or language-dependent information, we design a controlled experimental setup based on aligned multilingual MCQA queries/prompts, where the same query is asked in different languages. This setting allows for comparing internal model behavior across languages while keeping task semantics constant.

Multilingual MCQA as a Controlled Setup We make use of the mMMLU dataset, which provides professionally translated versions of the MMLU benchmark (Hendrycks et al., 2020) in 14 languages. The presence of the same query in multiple languages enables a counterfactual-style comparison, i.e., the same underlying knowledge and reasoning requirements are presented to the model under different linguistic encodings. MCQA provides a principled and constrained setting for studying decision-making because it reduces reasoning to a single discrete commitment among a fixed set of alternatives (Robinson et al., 2023; Wiegreffe et al., 2024). Unlike open-ended generation, it minimizes confounds arising from token frequency, verbosity, or fluency, and yields a well-defined decision point that can be traced through the model's internal representations.

Each question contains four answer options, yielding a random-guess accuracy of 25%. To mitigate positional biases, we randomly shuffle the order of answer choices for each instance. The

model is instructed to respond by generating the identifier of the selected option (e.g., "A", "B", "C", or "D"), ensuring that the final prediction corresponds to a single token. This discrete output makes MCQA particularly well-suited for layer-wise and trajectory-based analysis of decision emergence. We explain the model input-related details, including the prompt structure, notations, and templates, in the App. F.

5 Representation Extraction and Decision Emergence

To study how multilingual language models form decisions, we examine the evolution of internal representations associated with the model's final answer (i.e., the predicted token) selection. Our analysis focuses on the *decision token*, i.e., the token corresponding to the chosen answer option, and how its representation is constructed and refined across transformer layers. This token provides a precise and interpretable locus for studying decision-making, as it integrates information from the entire prompt directly determines the model's output. In autoregressive transformer-based models, predictions are generated sequentially, and the hidden representation of the final token determines the distribution of the next token. In the MCQA setting, this next token corresponds to the identifier of the answer choice. As a result, the representation associated with this token serves as the model's *point of commitment*: once generated, the model has effectively completed its reasoning and selected an answer. Note that the commitment here is defined as the earliest layer at which the model's top-logit prediction over {"_A","_B","_C", "_D"} remains stable for all subsequent layers, which we also refer as an *emergence of decisiveness* in model representations.

Let $x^{(\ell)}_{i<k}$ denote the input token sequence for a question expressed in language ℓ. As the prompt propagates through a transformer model $\mathcal{M}_\theta = \{f_{\theta_1}, \ldots, f_{\theta_T}\}$ with T layers, each layer updates a shared residual stream. We extract the hidden representation corresponding to the decision token after each layer, enabling a layer-wise view of how the model incrementally refines its decision.

Layer-wise Representation Extraction For a given input in language ℓ, let $h^{(\ell)}(t) \in \mathbb{R}^{d_{\text{model}}}$ denote the hidden state of the decision token after the t-th transformer layer. These vectors capture how contextual information and task-relevant features are progressively accumulated as the model processes the input. By collecting $\{h^{(\ell)}(t)\}_{t=1}^{T}$, we obtain a *layer-wise trajectory* that reflects the internal evolution of the model's decision for a given language and question. We also observe cluster formation in the representational space using UMAP projection (see App. Fig. 14, 15). Unless stated otherwise, we analyze post-residual representations; details on isolating attention and MLP contributions are provided in App. C.

Decisiveness via Layer-wise Logit Projection To quantify when a representation becomes *decisive*, we adopt the Logit Lens approach (nostalgebraist, 2020; Haviv et al., 2023). At each layer t, we project the decision-token representation $h^{(\ell)}(t)$ directly to the vocabulary space using the model's unembedding matrix, i.e., $\text{logits}^{(\ell)}(t) = \mathbf{W}_U \,\text{LayerNorm}(h^{(\ell)}(t))$. This projection yields a hypothetical next-token distribution as if the model were forced to make a prediction at layer t. By measuring whether the correct answer token has the highest logit (among options) at each layer, we obtain a layer-wise estimate of decision accuracy. Fig. 1 illustrates this process for LLaMA-3 on mMMLU across multiple languages. We observe that performance remains near chance (25% for four options) in early layers and then rises sharply over a narrow range of layers (16 - 17). Specifically, while the overall pattern is consistent across languages, the rate and sharpness of this transition vary substantially. This variation provides early evidence that, although representations may be globally similar, the point at which the model commits to a specific answer is language-specific. The layer-wise extraction also forms the foundation for the trajectory-based analyses in the next section, where we treat $\{h^{(\ell)}(t)\}_{t=1}^{T}$ as a continuous decision path, updated by the stacked transformer blocks, and quantify how multilingual trajectories diverge, align, and evolve over the course of inference.

5.1 Constructing and Analyzing Decision Trajectories

Layer-wise representations provide snapshots of a model's internal state, but reasoning in transformers is inherently sequential across layers. To capture how decisions are formed over time (stacked computational blocks), we model inference as a *trajectory* in representation space, i.e., a stepwise path traced by the decision token as it propagates through the transformer stack. This trajectory-based view allows us to study not only *when* a

model becomes decisive, but *how* its internal representations evolve toward a commitment, and how this process differs across languages.

Decision Trajectories For a given multiple-choice question q_i expressed in language ℓ, let $h_i^{(\ell)}(t) \in \mathbb{R}^{d_{\text{model}}}$ denote the hidden representation of the decision token after layer $t \in \{1, \ldots, T\}$. The full decision trajectory is defined as

$$\Gamma_i^{(\ell)} = \left[h_i^{(\ell)}(1), h_i^{(\ell)}(2), \ldots, h_i^{(\ell)}(T)\right] \in \mathbb{R}^{d_{\text{model}} \times T}$$

Each trajectory represents not just representations, but the evolving decision state of the model, which we treat as the object of comparison across languages. Because we use aligned multilingual prompts, trajectories corresponding to the same question across different languages encode identical task semantics but differ in linguistic realization. This enables direct, controlled comparison of decision dynamics across languages.

Trajectory Projection and Visualization Raw decision trajectories live in a high-dimensional space and are difficult to interpret directly. To facilitate comparison and visualization, we project trajectories into a shared low-dimensional subspace. For each layer t, we form a matrix

$$M^{(l)}(t) = \left[h_1^{(\ell)}(t), \ldots, h_N^{(\ell)}(t)\right] \in \mathbb{R}^{d_{\text{model}} \times N}$$

containing decision-token representations for all queries of language ℓ, where N is the number of datapoints/queries. We compute the singular value decomposition

$$M^{(\ell)}(t) = U^{(\ell)}(t)\, \Sigma^{(\ell)}(t)\, V^{(\ell)}(t)^\top$$

and retain the top three left singular vectors $\mathcal{U}^{(\ell)}(t) = [u_1^{(\ell)}(t), u_2^{(\ell)}(t), u_3^{(\ell)}(t)]$ as a shared basis. Each trajectory is projected to $\mathbb{R}^{3\times T}$ as

$$\hat{\Gamma}_i^{(\ell)} = \begin{bmatrix} u_1^{(l)}(1)^\top h_i^{(\ell)}(1) & \cdots & u_1^{(l)}(L)^\top h_i^{(\ell)}(T) \\ u_2^{(l)}(1)^\top h_i^{(\ell)}(1) & \cdots & u_2^{(l)}(L)^\top h_i^{(\ell)}(T) \\ u_3^{(l)}(1)^\top h_i^{(\ell)}(1) & \cdots & u_3^{(l)}(L)^\top h_i^{(\ell)}(T) \end{bmatrix}$$

This projection captures the dominant directions of variation in decision-making while preserving the temporal structure across layers. Note that the SVD helps to test/validate whether decision trajectories lie in a low-dimensional subspace (see App. Fig. 5, where we observe singular values decaying quickly and monotonically over several orders of magnitude, pointing towards the existence of low-dimensionality). If trajectories are highly structured, a small number of directions should explain most of the variance; if not, trajectory comparisons across languages would be ill-posed. Fig. 2 right side shows the evolution trajectory of last token representation across transformer blocks (also see App. Fig. 12, 13). Despite language-specific variations, trajectories show a consistent curved but low-dimensional structure, supporting the hypothesis that decision formation could be understood by a shared underlying geometric subspace that can be useful for comparison.

To quantify cross-lingual divergence during inference, we compare decision trajectories across languages for semantically aligned questions. We measure the average layer-wise deviation between trajectories, yielding a counterfactual distance that captures how the model's decision process changes under alternative linguistic realizations (see App. D for detailed definitions/formulations and App. Fig. 6 for the obtained results).

Since we now know that the trajectories lie in a low-dimensional subspace, we now see if a direct comparison could be made to understand the decision-making/formation happening between the stacked transformer blocks. In the next section, we formalize this comparison by modeling trajectory evolution as a stochastic process, allowing us to formulate a predictive setting, i.e., testing whether trajectories in one language can forecast/predict those in another and quantify where cross-lingual predictability breaks down.

6 Decision Trajectories as a Stochastic Process

The trajectory formulation introduced in the previous section treats decision-making as a sequence of layer-wise updates on the residual stream. While this view captures the geometry of reasoning, it does not yet provide a principled model for how trajectories/representations evolve across layers, nor how trajectories in different languages relate to one another. In this section, we model decision trajectories as a stochastic dynamical process and use this formulation to quantify cross-lingual deviations in decision-making. We provide the intuitive motivation for the presented Stochastic View of decision making in App. B. Note that we do not assume that transformer layers are generated by a true stochastic Itô process. Instead, we use this formulation as a linearized analytical surrogate for layer-to-layer representation changes. The goal is to quantify cross-lingual predictability of updates

rather than to model the full generative dynamics of the network.

Itô-Style Approximation of Layer-wise Dynamics Our formulation is motivated by recent findings on linear approximations of transformer computations (Sarfati et al., 2025; Park et al., 2024), which suggest that, locally, a transformer block can be approximated as a linear operator acting on the previous layer's representations. More specifically, to approximate a representation $h(t+\tau)$ from $h(t)$, one can project $h(t)$ onto an intrinsic basis at layer t, rescale the coordinates along principal directions, and rotate them to account for changes in the basis between layers. This decomposition provides a natural way to take a linearized view of layer-wise transformations. Following this intuition and building on recent findings that hidden-state evolution can be well approximated by a low-rank linear structure with residual noise, we model the evolution of decision-token representations as a special case of a discrete-time Itô process. Let $h(t) \in \mathbb{R}^{d_{\text{model}}}$ denote the decision-token representation at layer t. The evolution from layer t to $t+\tau$ is modeled as:

$$h(t+\tau) = \underbrace{\mu(t,\tau)}_{\text{drift}} h(t) + \underbrace{w(t,\tau)}_{\text{diffusion}}$$
$$= R(t+\tau)\,\Lambda(t,\tau)\,R(t)^\top h(t) + w(t,\tau)$$

where $R(t)$ is an orthogonal matrix capturing the dominant basis of variation at layer t, $\Lambda(t,\tau)$ is a diagonal scaling matrix governing contraction or expansion along principal directions, and $w(t,\tau)$ is a noise term capturing unmodeled nonlinear effects. In practice, we estimate these quantities using the singular value decomposition of the matrix of decision-token representations at each layer, $M(t) = U(t)\,\Sigma(t)\,V(t)^\top$ and set $R(t) = U(t)$, $\Lambda(t,\tau) = \Sigma(t+\tau)\Sigma(t)^{-1}$. This approximation captures the dominant linear component of trajectory evolution while allowing residual stochasticity.

Cross-Lingual Trajectory Prediction We extend this hypothesis formulation to a multilingual setting by using trajectories in one language to predict trajectories in another. Let ℓ_a and ℓ_b denote two languages, and $h^{(\ell_a)}(t) \in \mathbb{R}^{d_{\text{model}}}$ be the decision-token representation for a given question at layer t in language ℓ_a. We construct a predicted representation for language ℓ_b at layer $t+\tau$ as:

$$\hat{h}^{(\ell_b)}(t+\tau) = \mu^{(l_a,l_b)}(t,\tau)h^{(\ell_a)}(t)$$
$$\mu^{(l_a,l_b)}(t,\tau) = R^{(\ell_b)}(t+\tau)\,\Lambda^{(\ell_a,\ell_b)}(t,\tau)\,R^{(\ell_a)}(t)^\top$$

where $R^{(\ell)}(t) = U^{(\ell)}(t)$ and $\Lambda^{(\ell_a,\ell_b)}(t,\tau) = \Sigma^{(\ell_b)}(t+\tau)\Sigma^{(\ell_a)}(t)^{-1}$. This construction provides a cross-lingual predictive approximation to test whether trajectories in one language can predict those in another, i.e., if the model were reasoning in language ℓ_b, how would its internal decision state evolve/deviate given the state observed in language ℓ_a?

Prediction Error To evaluate the quality of these cross-lingual predictions, we measure the deviation between the predicted and true decision trajectories in language ℓ_b. To ensure comparability across layers, we use a normalized, direction-based error metric that isolates differences in trajectory direction rather than magnitude. Full definitions and implementation details are provided in App. E.

Interpreting Error as Language-Specific Decision Dynamics Low prediction error indicates that decision evolution in language ℓ_b can be accurately inferred from language ℓ_a, suggesting shared decision dynamics. Conversely, high error reveals layers where multilingual trajectories diverge in a way that cannot be explained by a shared linear process. Specifically in Fig. 1, we find that prediction error peaks in the same layers where logit-based decisiveness emerges and where trajectory divergence is highest (for all subcategories in the MMLU dataset, see App. Fig. 7a - 7d, also see App. Fig. 16). This alignment supports a primary finding of the paper, language-specific deviations are more pronounced during the decision-making phase of inference. Early layers admit a largely shared stochastic evolution, while later layers exhibit language-dependent deviations that reflect how the model commits to an answer. Fig. 3 shows the prediction error from English trajectory to other languages in more detail. Note that this should be interpreted as a correlational alignment between divergence and decision formation, rather than a strict causal attribution.

In the next section, we leverage this insight to identify dominant directions associated with English-centric decision-making and test whether steering representations along these directions can improve performance in other languages.

6.1 Languages-to-English Directions and Cross-Lingual Steering

Our trajectory and prediction-error analyses reveal that while early-layer processing is largely shared across languages, late-layer decision trajectories diverge in a structured manner. Languages with

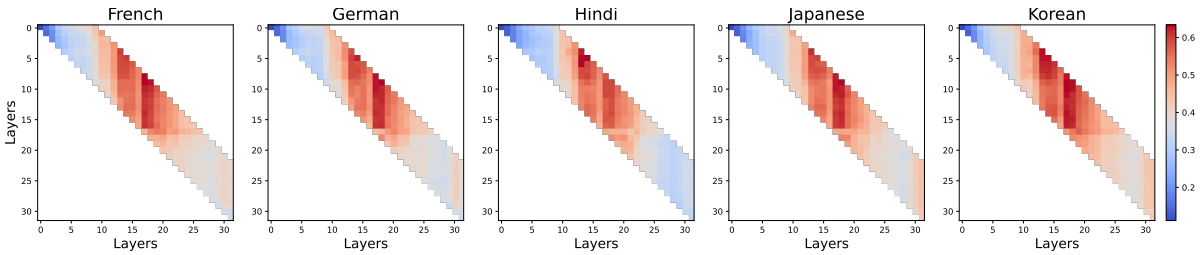


Figure 3: The Figure shows trajectory prediction error (defined in App. E), from English trajectory to other languages. For each layer, we also consider predicting the representations for the next 10 layers (rows in the matrix), that directly highlights the divergence increasing in the middle layer where the decision starts to form.

higher downstream accuracy, most notably English, occupy a distinct region of the decision space near the point of commitment. In this section, we ask whether this geometric advantage plays a crucial role in decision quality.

English as a Reference Decision Space Prior work suggests that multilingual LLMs may implicitly rely on English as a reference language during reasoning (Wendler et al., 2024; Zhao et al., 2024). Our analyses refine this view, rather than reasoning explicitly in English throughout inference, the model's late-layer decision states for different languages are systematically displaced relative to English along consistent directions.

Languages-to-English Direction For a fixed layer t and question q, we define an *Languages-to-English Direction* as the average normalized displacement from non-English decision representations toward the English representation:

$$\hat{d}_q^{(\mathrm{en})}(t) = \frac{1}{|L|-1} \sum_{\ell \neq \mathrm{en}} \frac{h_q^{(\mathrm{en})}(t) - h_q^{(\ell)}(t)}{\|h_q^{(\mathrm{en})}(t) - h_q^{(\ell)}(t)\|_2}$$

This direction ($d_q^{(\mathrm{en})}(t) = \hat{d}_q^{(\mathrm{en})}(t)/\|\hat{d}_q^{(\mathrm{en})}(t)\|$) captures a systematic axis along which English decision states differ from those of other languages. We find that the stability and magnitude of this direction increase sharply in the same layers where decision trajectories diverge and logit-based accuracy saturates, suggesting its relevance to the decision-making phase of inference.

Steering Decision Trajectories To test the impact/contribution of this direction, we further intervene on non-English decision representations during inference. For language $\ell \neq \mathrm{en}$ at layer t, we modify the residual stream as $\tilde{h}^{(\ell)}(t) = h^{(\ell)}(t)/\|h^{(\ell)}(t)\|_2 + d^{(\mathrm{en})}(t)$. We normalize and scale by the representation norm ($\|h^{(\ell)}(t)\|_2$) to preserve layer-dependent magnitude and avoid artificially collapsing the activation scale when applying the intervention. This intervention nudges the internal state toward the English decision subspace without changing the input or model parameters.

Effects on Multilingual Performance Fig. 4 shows the effect of Languages-to-English steering on multiple-choice accuracy across layers for several non-English languages. Across languages, we note that steering narrows, but does not eliminate, the performance gap to English, suggesting that English-aligned directions encode a shared yet incomplete decision subspace. These results provide some evidence that late-layer English-aligned directions directly contribute to decision quality, rather than merely reflecting correlational differences between languages. Note that we do not claim that this intervention isolates a uniquely English subspace; rather, it probes whether a consistent high-performance direction exists across languages in late-layer decision space.

Interpretation and Limitations These results suggest that multilingual LLMs encode a shared decision subspace that is unevenly accessed across languages. English, by virtue of its prevalence during training, appears to lie closer to a high-utility region of this subspace. Steering other languages toward this region improves decision quality, but does not eliminate language-specific structure entirely. We would like to emphasize that steering is used here as a diagnostic tool rather than a proposed deployment strategy. By identifying and intervening on English-aligned directions in representation space, we provide interventional evidence that multilingual performance disparities arise from language-specific decision dynamics rather than a lack of shared conceptual understanding. This finding complements our earlier mechanistic analyses and underscores the importance of studying decision-making, not just representation similarity, when evaluating multilingual language models.

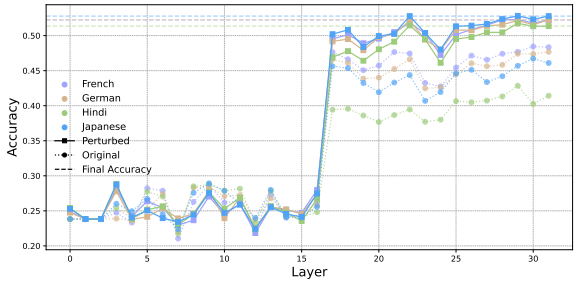


Figure 4: Adding English steering vector to the latent representations of each languages results in an increase in the accuracy once the decision making phase starts.

7 Discussion

Our analysis of multilingual LLMs reveals several interconnected phenomena that illuminate how these models form decisions and manage cross-lingual reasoning. By combining trajectory-based analyses, logit-lens accuracy, and interventions along English-aligned directions, the proposed experimental setup helps uncover the dynamics and structure of decision-making in transformer-based language models.

Language-Specific Trajectory Divergence Aligned multilingual MCQA prompts reveal that decision trajectories are not fully language-invariant. Importantly, our analysis shows that this divergence is not gradual from early layers, but emerges sharply near the same layers where decisiveness occurs. While shared early-layer representations exist, late-layer deviations are systematic, potentially reflecting differences in training data frequency, tokenization, and syntactic conventions. Multiple trajectory-based metrics (e.g., Counterfactual Distance, and CKA (Park et al., 2024); see App. G, Fig. 6, 8, 9)) consistently show that higher divergence correlates with increased prediction error across languages. These results highlight that multilingual LLMs encode a shared conceptual subspace but access it via language-specific pathways. Understanding these pathways is helpful for diagnosing and improving cross-lingual performance.

English-Aligned Decision Directions and Steering By identifying English-aligned directions in the decision subspace, we provide interventional evidence that shifting representations toward these directions improves performance. We also note that these directions are derived from differences between languages on the same task and may partially encode correctness signals in addition to language-specific structure. This finding suggests that higher-performing languages (often English in our setup) tend to lie closer to regions associated with correct answer selection. Importantly, this analysis frames Language-to-English directions as a tool for understanding cross-lingual decision geometry rather than as a prescriptive solution. It demonstrates that language disparities are not merely an artifact of representation similarity but emerge dynamically during decision formation.

Mechanistic Insights and Broader Implications Our work emphasizes the importance of studying decision dynamics rather than static embeddings alone. The combination of trajectory analysis, logit-lens probing, and Itô-process modeling provides a complementary set of analytical tools for tracking how internal representations evolve toward a discrete choice. Each component captures a different aspect, where logit lens (nostalgebraist, 2020) identifies when decisions emerge, trajectory analysis captures how representations evolve, and prediction-based analysis tests cross-lingual predictability of these dynamics.

8 Conclusion

Our study demonstrates that multilingual LLMs form decisions through the structured, layerwise accumulation of information in a residual stream, resulting in language-specific yet partially aligned decision trajectories. We find that higher-performing languages lie closer to regions associated with correct answer selection, and that interventions along these directions provide interventional evidence linking representation geometry to cross-lingual performance differences. Overall, our results suggest that multilingual reasoning is neither fully shared nor entirely language-specific, but emerges from shared representations that diverge during decision formation. This provides a layer-resolved perspective on how representational geometry and decision dynamics interact in multilingual LLMs. More importantly, the multilingual setup we used (sharing the same knowledge across multiple languages and decision-making starting from the same token) will be interesting for future analysis to pin down the multilingual behavior of language models.

Limitations

Several limitations warrant further discussion. First, our analyses focus on multiple-choice question answering; other tasks, especially open-ended

generation, may exhibit different trajectory dynamics. Second, English-aligned steering provides only a partial correction for cross-lingual disparities; it does not resolve structural biases arising from tokenization, pretraining corpora, or cultural context. Finally, while trajectory-based and Itô-process modeling provide a rich lens for mechanistic understanding, these methods rely on low-dimensional projections and linear approximations, which may miss finer-grained nonlinear interactions in the residual stream. Future work could extend these analyses to: 1) Broader task categories beyond MCQA, including reasoning, translation, and dialogue. 2) Mixed-language and code-switching prompts to study real-world multilingual scenarios. 3) Training-time interventions that encourage cross-lingual alignment of decision trajectories without reliance on a single pivot language.

Ethical Considerations

The current research focuses on understanding the inner workings of LLMs. To the best of our knowledge, the current research does not have direct ethical implications. We have used openly available datasets in this research.

Acknowledgements

We would like to thank the anonymous reviewers and the meta-reviewer for their insightful comments and suggestions. This research work was partially supported by the Research-I Foundation of the Department of CSE at IIT Kanpur.

References

Lucas Bandarkar, Chenyuan Yang, Mohsen Fayyaz, Junlin Hu, and Nanyun Peng. 2026. Multilingual routing in mixture-of-experts. In *The Fourteenth International Conference on Learning Representations*.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2025. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31822–31841, Vienna, Austria. Association for Computational Linguistics.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. How do multilingual language models remember facts? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16052–16106, Vienna, Austria. Association for Computational Linguistics.

Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniil Gurgurov, Katharina Trinley, Yusser Al Ghussin, Tanja Baeumel, Josef van Genabith, and Simon Ostermann. 2025. Language arithmetics: Towards systematic language neuron identification and manipulation. *Preprint*, arXiv:2507.22608.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. *Preprint*, arXiv:2210.03588.

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

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. 2025a. Calibration across layers: Understanding calibration evolution in LLMs. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

Abhinav Joshi, Areeb Ahmad, Divyaksh Shukla, and Ashutosh Modi. 2025b. Towards quantifying commonsense reasoning with mechanistic insights. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9633–9660, Albuquerque, New Mexico. Association for Computational Linguistics.

Abhinav Joshi, Divyanshu Bhatt, and Ashutosh Modi. 2025c. Geometry of decision making in language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. 2025. Resi: A comprehensive benchmark for representational similarity measures. In *The Thirteenth International Conference on Learning Representations*.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

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

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8186–8213, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn. 2026. Language-specific latent process hinders cross-lingual performance.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Meng Lu, Ruochen Zhang, Carsten Eickhoff, and Ellie Pavlick. 2025. Paths not taken: Understanding and mending the multilingual factual recall pipeline. *Preprint*, arXiv:2505.20546.

Gabriele Maraia, Leonardo Ranaldi, Marco Valentino, and Fabio Massimo Zanzotto. 2026. Can activation steering generalize across languages? a study on syllogistic reasoning in language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2739–2753, Rabat, Morocco. Association for Computational Linguistics.

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

Neel Nanda and Joseph Bloom. 2022. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`.

Ercong Nie, Helmut Schmid, and Hinrich Schuetze. 2025. Mechanistic understanding and mitigation of language confusion in English-centric large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 690–706, Suzhou, China. Association for Computational Linguistics.

nostalgebraist. 2020. interpreting GPT: the logit lens, LessWrong. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. [Accessed 26-01-2025].

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20454–20486, Suzhou, China. Association for Computational Linguistics.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering. *Preprint*, arXiv:2210.12353.

Raphaël Sarfati, Toni J.B. Liu, Nicolas Boulle, and Christopher Earls. 2025. Lines of thought in large language models. In *The Thirteenth International Conference on Learning Representations*.

Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. Do multilingual LLMs think in english? In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*.

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. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2375–2401, Vienna, Austria. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

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, Bangkok, Thailand. Association for Computational Linguistics.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hanna Hajishirzi, and Ashish Sabharwal. 2024. Answer, assemble, ace: Understanding how lms answer multiple choice questions. In *International Conference on Learning Representations*.

Sarah Wiegreffe, Oyvind Tafjord, Yonatan Belinkov, Hannaneh Hajishirzi, and Ashish Sabharwal. 2025. Answer, assemble, ace: Understanding how LMs answer multiple choice questions. In *The Thirteenth International Conference on Learning Representations*.

Bryan Wilie, Samuel Cahyawijaya, Junxian He, and Pascale Fung. 2025. High-dimensional interlingual representations of large language models. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 122–155, Vienna, Austria. Association for Computational Linguistics.

Linjuan Wu, Hao-Ran Wei, Huan Lin, Tianhao Li, Baosong Yang, Fei Huang, and Weiming Lu. 2025a. Enhancing LLM language adaption through cross-lingual in-context pre-training. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27152–27166, Suzhou, China. Association for Computational Linguistics.

Zhaofeng Wu, Xinyan Velocity Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim. 2025b. The semantic hub hypothesis: Language models share semantic representations across languages and modalities. In *The Thirteenth International Conference on Learning Representations*.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2025. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10602–10617, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Unveiling linguistic regions in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6228–6247, Bangkok, Thailand. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Appendix

Table of Contents

List of Figures

A An Itô-Process-Inspired View of Layer-Wise Representation Dynamics

To formalize the intuition, we adopt an *Itô-process-inspired* perspective on the evolution of token representations across transformer layers. In stochastic calculus, an Itô process describes the evolution of a state as a sequence of small, incremental updates driven by deterministic and stochastic components. While transformer models are deterministic at inference time, the residual stream update $h(t+1) = h(t) + f_{\theta_t}(\mathbf{h}(t))$ naturally lends itself to an analogous interpretation: each layer applies a small, structured update that gradually steers the representation toward a final outcome.

We emphasize that this is an *interpretive lens*, not a literal stochastic differential equation. The utility of this perspective lies in framing decision-making as a *progressive accumulation of evidence* across layers. Under this view, early layers correspond to low signal-to-noise regimes where answer probabilities remain near-uniform, while later layers amplify specific directions in representation space that correspond to particular answer choices. This formulation motivates our trajectory-based analyses in later sections. By comparing the layer-wise evolution of decision-token representations across languages, we can quantify when trajectories diverge, how decisiveness emerges, and which layers play a dominant role in multilingual decision-making.

B Why a Stochastic View?

Transformer layers repeatedly apply similar operations, attention, non-linear mixing, and residual updates, suggesting that inference proceeds through incremental transformations rather than abrupt state changes. Recent work has shown that, when viewed at a sufficiently coarse scale, hidden-state trajectories in transformers can be well approximated by linear stochastic processes (Sarfati et al., 2025). This perspective models inference as a noisy evolution through representation space, where each layer applies a small, structured update to the current state. Considering this view allows us to move beyond descriptive trajectory visualization and ask a sharper question: *to what extent can the evolution of a decision trajectory in one language predict the trajectory in another language?* If multilingual reasoning were fully language-agnostic, such predictions should be accurate. Systematic prediction error, by contrast, would indicate language-specific decision dynamics. We approximate layer-wise evolution using a linear stochastic model that captures dominant directions of variation while treating residual effects as noise.

C Residual Stream and MLP Contributions

Within each transformer layer, information is written to the residual stream twice: once by the self-attention block and once by the feed-forward (MLP) block. To disentangle these contributions, we extract representations at two points: 1) **MLP output**: the contribution of the non-linear feed-forward network before it is added to the residual stream; 2) **Residual post (Resid-Post)**: the updated residual stream after both attention and MLP updates.

Note that unless stated otherwise, our analysis focuses on the Resid-Post representations, as they reflect the full accumulated state used by subsequent layers and by the final unembedding operation. Focusing on the decision token's Resid-Post state ensures that the extracted representation integrates all information relevant to answer selection. All representations are extracted using `TransformerLens` (Nanda and Bloom, 2022), which provides precise access to internal activations at each layer.

D Counterfactual Distance Metrics for Trajectory Comparison

Decision trajectories provide a natural object for quantifying how multilingual reasoning unfolds. Given two languages ℓ_a and ℓ_b and a shared question $q \sim \mathcal{Q}$, we compare their trajectories by measuring the average layer-wise distance:

$$\mathrm{CDM}(q; \ell_a, \ell_b) = \frac{1}{T} \sum_{t=1}^{T} \left\| h_q^{(\ell_a)}(t) - h_q^{(\ell_b)}(t) \right\|_2^2$$

We refer to this quantity as the *counterfactual distance metric (CDM)*: one language serves as the factual input, while the other is a counterfactual realization of the same semantics. Low values indicate that the model follows similar internal decision paths across languages, whereas high values reflect language-dependent divergence.

Averaging over the dataset $\mathcal{Q}$, yields a global measure of cross-lingual divergence:

$$\mathrm{LangDiv}(\ell_a, \ell_b) = \mathbb{E}_{q \sim \mathcal{Q}} \left[\mathrm{CDM}(q; \ell_a, \ell_b) \right]$$

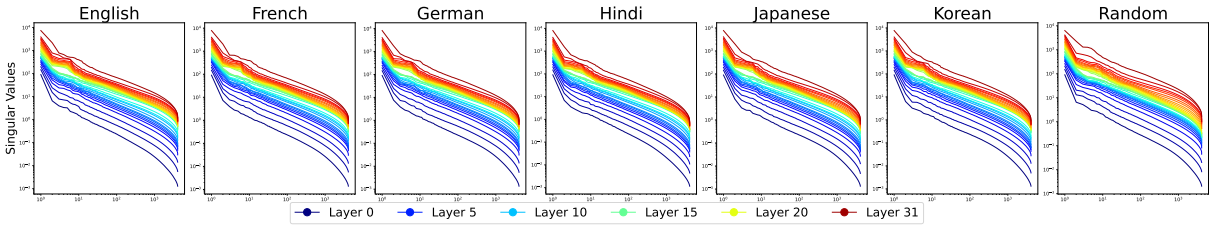


Figure 5: The figure shows singular values increasing monotonically across the layer, highlighting that the corresponding singular values decay quickly over several orders of magnitude, highlighting a low-dimensional curved subspace followed by the decision trajectories, making trajectory comparison suitable for a multilingual decision making setup.

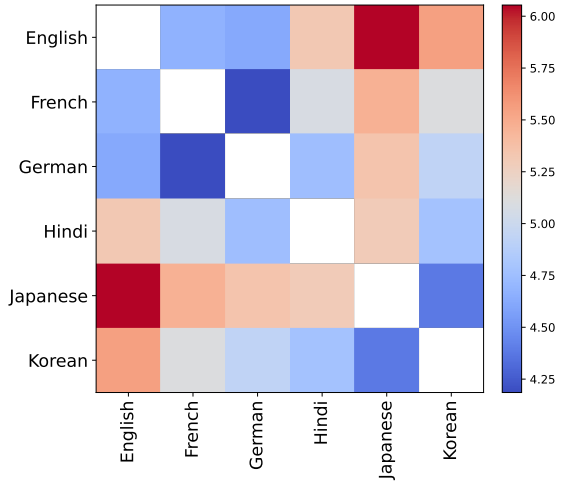


Figure 6: Pairwise language divergence across languages. Heatmap of the language divergence score between pairs of languages, computed by averaging trajectory distances over semantically equivalent inputs. Lower values indicate stronger cross-lingual alignment. The matrix shows a systematic variation in representational similarity across different language pairs.

These metrics allow us to identify layers where multilingual trajectories diverge most strongly, revealing stages of inference where language-specific processing dominates decision-making.

E Normalized Prediction Error

To quantify the quality of these cross-lingual predictions, we compare the predicted representation $\hat{h}^{(\ell_b)}(t)$ with the true representation $h^{(\ell_b)}(t)$. Because representation norms tend to grow across layers, raw mean squared error is not comparable across depth. We therefore use a normalized error metric:

$$\mathrm{Err}(t;\ell_a,\ell_b) = \left\| \frac{h^{(\ell_b)}(t)}{\|h^{(\ell_b)}(t)\|_2} - \frac{\hat{h}^{(\ell_b)}(t)}{\|\hat{h}^{(\ell_b)}(t)\|_2} \right\|_2 .$$

This measure captures angular deviation between predicted and true trajectories, isolating differences in direction rather than magnitude.

F Prompt Details

For a given language ℓ, each MCQA prompt consists of three components: 1) **Query** $x^{(\ell)}_{\text{query}}$, containing the question text in language ℓ; 2) **Choice set** $x^{(\ell)}_{\text{options}} = \{\textbf{A.}\ o^{(\ell)}_1, \textbf{B.}\ o^{(\ell)}_2, \dots\}$, listing the candidate answers in the same or another language; 3) **Prompt template** x_ϵ, which specifies the task instruction (kept fixed across languages unless otherwise stated). The full input sequence is formed by concatenating these components at the token level. Let $x^{(\ell)}_{i<k}$ denote the resulting token sequence for a question expressed in language ℓ. Given an autoregressive transformer-based language model $\mathcal{M}_\theta = \{f_{\theta_1}, \dots, f_{\theta_T}\}$ with T layers, the model defines a probability distribution over the next token:

$$P(x_k \mid x^{(\ell)}_{i<k}, \mathcal{M}_\theta) = P\big(x_k \mid x^{(\ell)}_{\text{query}}, x^{(\ell)}_{\text{options}}, x_\epsilon, \mathcal{M}_\theta\big)$$

In autoregressive decoding, the generated next token x_t corresponds to the model's selected answer option. This token serves as the model's *point of commitment*, integrating information from the entire input context. By presenting the same question in multiple languages and examining the internal computation leading to this final token, we can directly compare how decision-making unfolds across languages. This experimental setup, i.e., aligned multilingual inputs, discrete decision outputs, and a well-defined commitment token, provides a controlled foundation for understanding/investigating the internal dynamics of multilingual reasoning. In the following sections, we build on this setup to extract internal representations, measure layer-wise decisiveness, and analyze decision trajectories across languages.

F.1 Prompt Templates

For all our experiments, we follow a standard prompt template. This section provides the details of the prompt templates used in our multiple-choice question answering (MCQA) evaluations of autoregressive open-weight language models (`LLaMA(-3-8B) in our experiments`). All prompts follow a unified format to ensure consistency across tasks and models. Each prompt begins with an instruction directing the model to select the correct answer from a set of multiple-choice options. In few-shot or in-context learning settings, this instruction is optionally followed by a set of in-context examples. The question is prefaced by a task-specific description, and followed by a list of labeled answer choices (A, B, C, etc.). The final line of the prompt contains the prefix `Answer:`, which serves as the model's response cue. For evaluation, we extract the model's next-token prediction probabilities at this position over the answer option tokens (e.g., `A`, `B`), which we treat as the model's predicted distribution over choices.

Fig. 10 illustrates a fully instantiated example using a question from the mMMLU dataset . The correct answer is provided at the end of the prompt and underlined. The colored annotations in the figure denote fixed template components (in black) and variable elements drawn from the dataset (in orange and teal).

This templated format is applied uniformly across datasets and experimental configurations, enabling controlled comparisons of model behavior across domains and prompting setups.

Randomized Cross-lingual Prompt We also extend the standard controlled MCQA setup to a randomized cross-lingual prompt construction setting, where each prompt component is independently sampled from a shared multilingual pool (available set of languages). Let $\mathcal{L}$ denote the set of available languages (6 languages in our case). Instead of fixing a single language ℓ for all components, we define a stochastic construction process in which the query, options, and prompt template may originate from different languages.

Formally, we sample language indices:

$$\ell_q, \ell_o, \ell_\epsilon \sim \mathcal{L},$$

and construct the corresponding prompt components as:

$$x^{(\ell_q)}_{\text{query}}, \quad x^{(\ell_o)}_{\text{options}}, \quad x^{(\ell_\epsilon)}_{\epsilon}.$$

The resulting input sequence is a compositional multilingual prompt formed by heterogeneous language segments:

$$x^{(\text{mix})}_{i<k} = \text{concat}\big(x^{(\ell_q)}_{\text{query}}, x^{(\ell_o)}_{\text{options}}, x^{(\ell_\epsilon)}_{\epsilon}\big).$$

Under this construction, the autoregressive language model $\mathcal{M}_\theta$ defines the next-token distribution as:

$$P(x_k \mid x^{(\text{mix})}_{i<k}, \mathcal{M}_\theta) = P\big(x_k \mid x^{(\ell_q)}_{\text{query}}, x^{(\ell_o)}_{\text{options}}, x^{(\ell_\epsilon)}_{\epsilon}, \mathcal{M}_\theta\big)$$

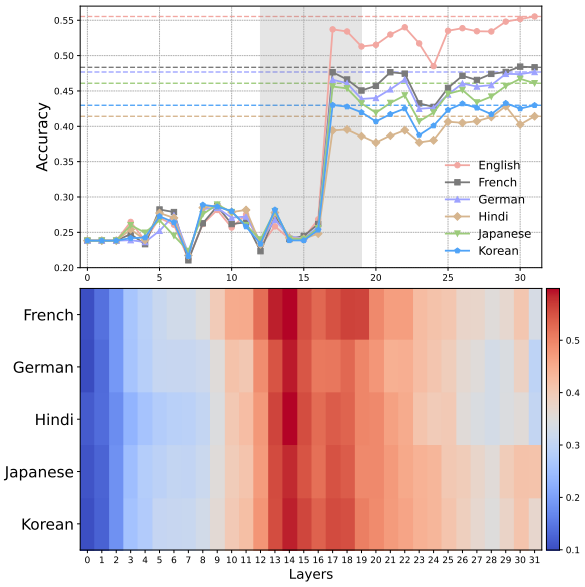
Fig. 11 shows an example of such a constructed prompt. This randomized construction adds another setting and introduces controlled linguistic heterogeneity within a single prompt, helping analyze cross-lingual composition, interference effects, and the robustness of decision-making in autoregressive reasoning models.

G Static Representation Similarity

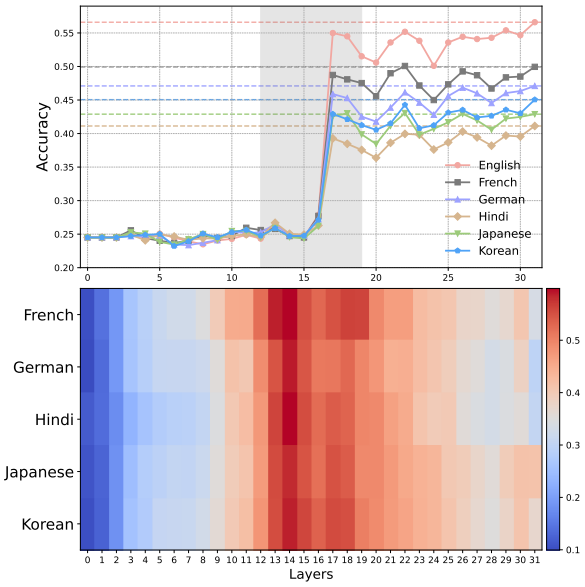
Trajectory-based metrics characterize the *dynamics* of decision-making, but they do not directly measure static representational alignment. To complement our analysis, we compute centered kernel alignment (CKA) (Kornblith et al., 2019) between decision-token representations across languages at each layer. CKA is invariant to isotropic scaling and orthogonal transformations and has been shown to be particularly well suited for comparing representations in language models (Klabunde et al., 2025).

While CKA captures whether representations occupy similar subspaces, our trajectory metrics reveal how these representations are *used over time* to form decisions. Together, these analyses allow us to distinguish shared representational structure from language-specific decision dynamics.

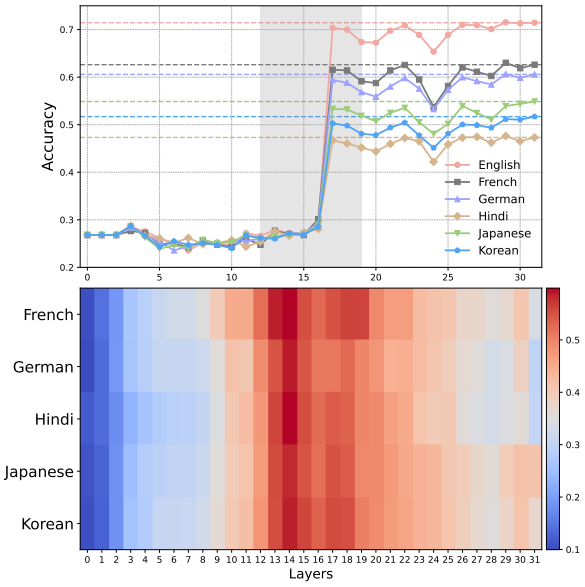
By treating inference as a trajectory through representation space, we obtain a unified framework for analyzing multilingual decision-making. This view reveals that, although representations may be globally aligned across languages, the *paths taken to reach a decision are often language-specific*, particularly in layers where decisiveness emerges. In the next section, we build on this trajectory framework to model decision evolution as a stochastic process and quantify deviations across languages in a principled manner.



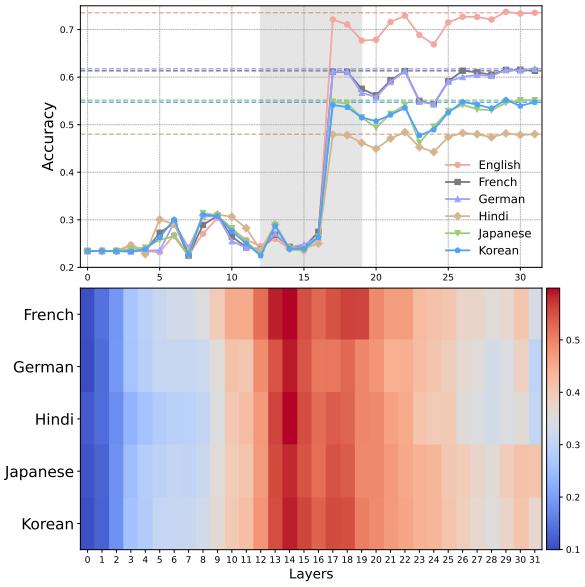
(a) STEM



(b) Humanities



(c) Other



(d) Social Sciences

Figure 7: The figure shows the performance (Accuracy) of the Llama-3 model on the diferent domains of MMLU dataset in multiple languages. We observe that the model performance remains near-random (25%, 4-options) for initial layers and starts to rise from layer 16 and saturates at layer 17, with minor changes in the 17-32 layers. However, the change varies across different languages, highlighting the model's use of different spaces for the same knowledge in different languages. Interestingly the same trend is observed for all the dataset categories present in MMLU.

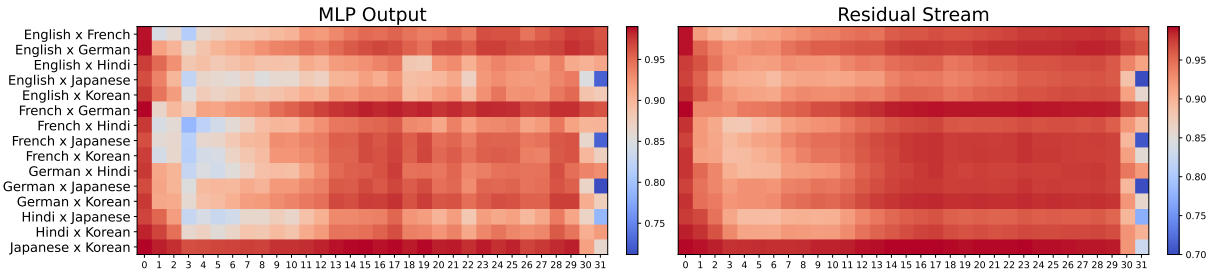


Figure 8: Layer-wise CKA reveals cross-lingual representational alignment. We compute centered kernel alignment (CKA) between trajectories induced by semantically equivalent inputs across language pairs. The MLP output (left) and post-residual stream (right) are compared layerwise. Results show high similarity throughout, with the residual stream maintaining stronger cross-lingual geometry alignment, reflecting preserved shared structure despite language-specific processing.

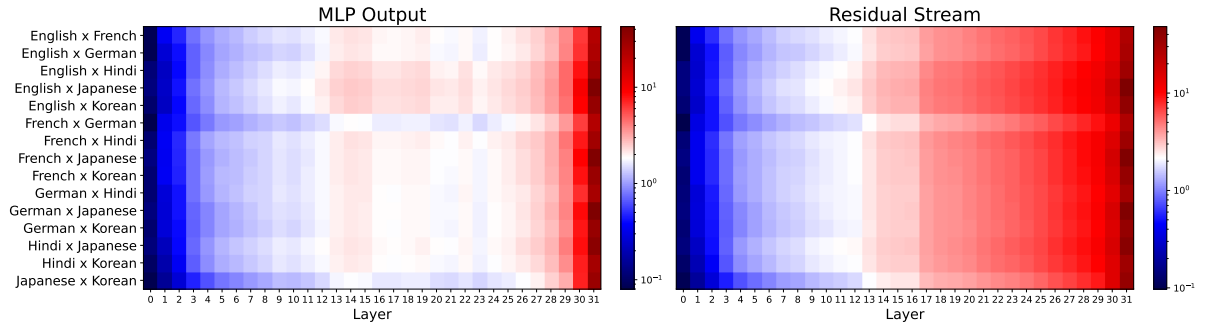


Figure 9: Layer-wise language divergence in residual stream and MLP outputs. Heatmaps show langauge divergence between trajectories induced by semantically equivalent inputs in different languages. Rows denote language pairs and columns denote transformer layers. Divergence is measured at the MLP output (left) and post-residual stream (right). Divergence is low in early layers and increases with depth, with consistently higher values in the residual stream, indicating that language-specific effects accumulate through residual updates and shape downstream representations.

Question: Mars has an atmosphere that is almost entirely carbon dioxide. Why isn't there a strong greenhouse effect keeping the planet warm?
A: the atmosphere on Mars is too thin to trap a significant amount of heat
B: There actually is a strong greenhouse effect and Mars would be 35oC colder than it is now without it.
C: Mars does not have enough internal heat to drive the greenhouse effect
D: the greenhouse effect requires an ozone layer which Mars does not have
Answer: A

Figure 10: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models. The black text is the templated input for all datasets. The blue text is the query from the **mMMLU dataset** and orange text are the respective choice set. The next-token prediction probabilities of the option IDs at the red text are used as the observed prediction distribution.

```
Question: Votre première action après la confirmation d'un arrêt cardiaque est de :
A: sicherstellen, dass das Notfallteam/die Notfalldienste gerufen werden.
B: zwei Atemzüge zur Wiederbelebung geben.
C: 30 Herzdruckmassagen durchführen.
D: die Aufzeichnungen prüfen, um festzustellen, ob der Patient eine DNAR-Anordnung hat.
Answer: A
```

Figure 11: Input prompt formats for the MCQA-based evaluation of autoregressive open-weight models. The `black text` is the templated input for all datasets. The `blue text` is the query from the **Random mMMLU dataset** and `orange text` are the respective choice set, where the query and the choices are probabilistically in different languages. The next-token prediction probabilities of the option IDs at the `red text` are used as the observed prediction distribution.

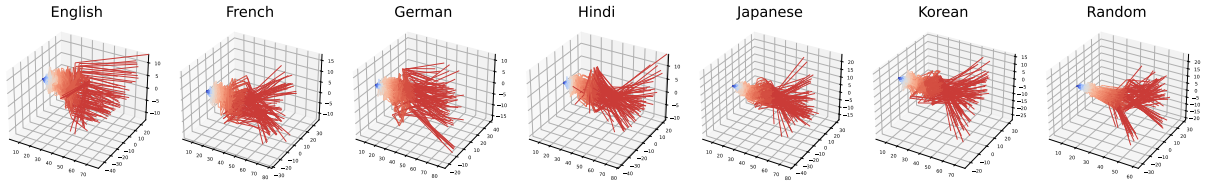


Figure 12: The figure shows trajectory visualizations obtained for decision-making across languages with a language-specific basis. Each trajectory is projected into a shared low-dimensional basis to reveal consistent geometric structure across languages.

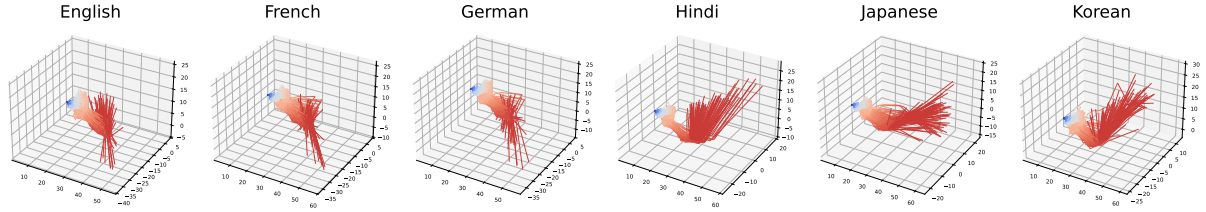


Figure 13: The figure shows trajectory visualizations obtained for decision making across languages where the basis is obtained for the Random MMLU dataset

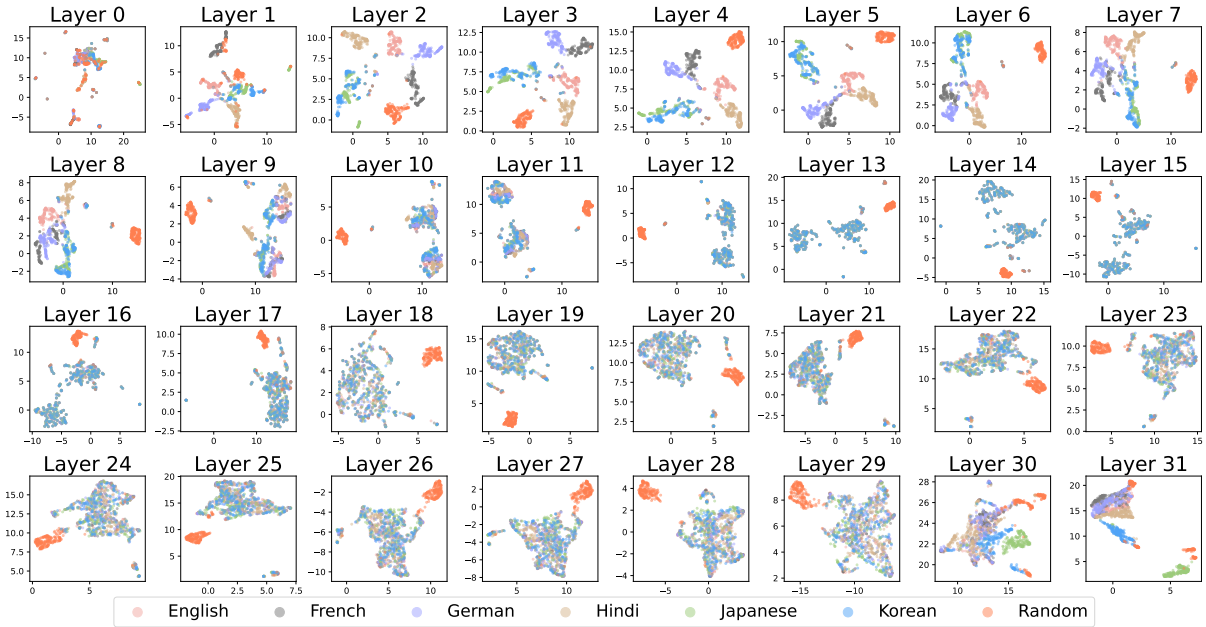


Figure 14: The figure shows the UMAP of latent representations across the model for different languages.

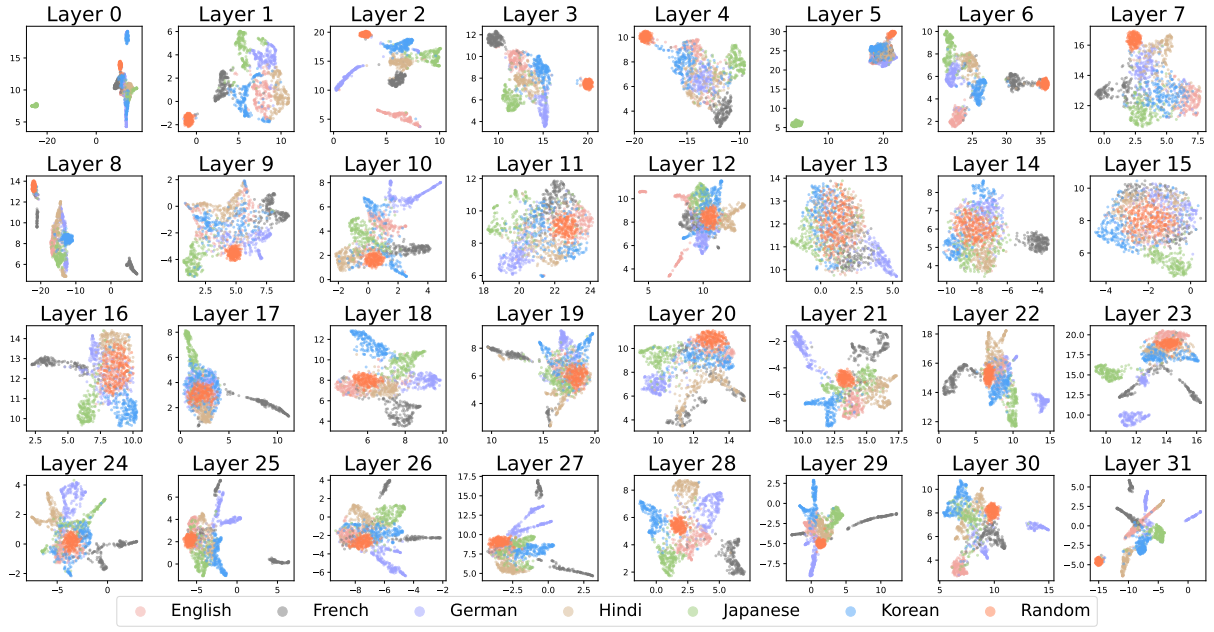


Figure 15: The figure shows the UMAP corresponding to the scaled trajectories (scaled by $\Sigma(t)^{-1}$) for different languages.

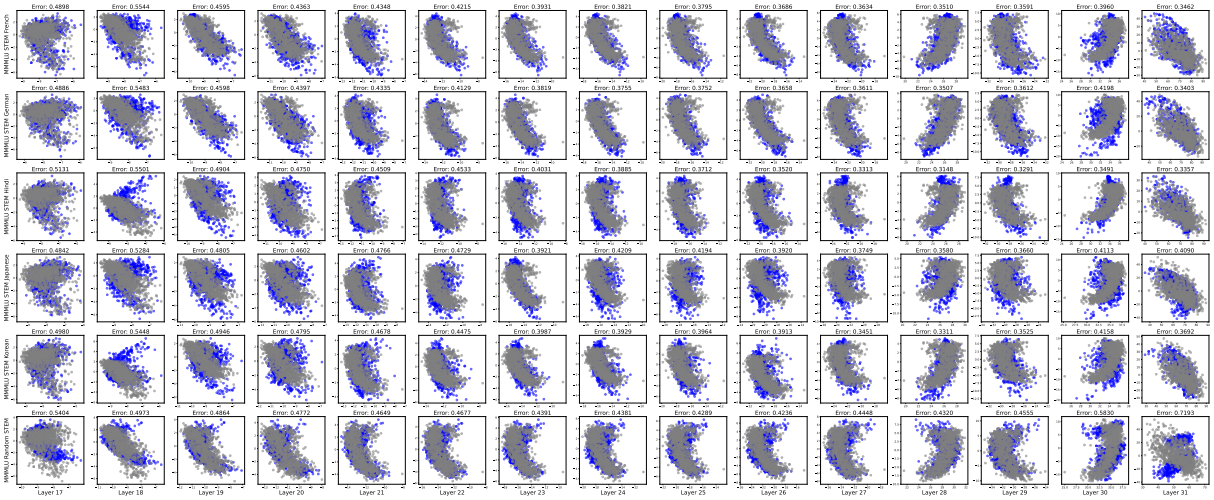


Figure 16: The figure shows the scatter plot corresponding to the Itô process predictions for MMLU STEM dataset, from English to other languages