Bridging the Language Gap: Uncovering and Aligning Shared Circuits for Multi-Hop Reasoning in Multilingual LLMs

Chenghao Sun[1], Zhen Huang[2], Yonggang Zhang[3], Xinmei Tian[1]*, Xu Shen[2]*, Jieping Ye[2]

[1]University of Science and Technology of China
[2]Independent Researcher
[3]The Hong Kong University of Science and Technology
{chsun@mail.,*xinmei@}ustc.edu.cn, {yellowtownhz13, *shenxuustc,jieping}@gmail.com, zhangyg@ust.hk

Abstract

Large language models (LLMs) present a paradox: they can correctly answer a multi-hop factual query in a high-resource language like English, yet fail on the identical query in another language. This raises a fundamental question about the nature of multilingual knowledge: are facts missing, or merely inaccessible? The underlying mechanisms for this knowledge gap have remained largely unexplored. In this work, we resolve this question by introducing a mechanistic interpretability framework that traces the causal pathways of multi-hop knowledge reasoning. Our analysis reveals a core, non-obvious finding: cross-lingual inconsistencies do not stem from a knowledge deficit. Instead, factual knowledge is robustly stored in a set of **shared, language-agnostic semantic neurons**. The failure originates from **misaligned attention pathways**, where a common set of critical attention heads fails to correctly route information along the reasoning chain to the appropriate knowledge neurons in lower-resource languages. This mechanistic diagnosis motivates a targeted alignment strategy: a surgical fine-tuning of only these critical heads. Experiments demonstrate that our method achieves significant improvements in multilingual multi-hop factuality—with positive cross-lingual transfer—while uniquely preserving general model capabilities, offering a scalable and mechanistically-grounded approach to building more reliable multilingual models.

1 Introduction

Large Language Models (LLMs), despite significant advances in multilingual training (Guo et al. 2023), exhibit a persistent and troubling disparity in factual reasoning. They can correctly answer a multi-hop query like "What is the capital of the country of citizenship of Anna Faris?" in a high-resource language like English (correctly reasoning: Anna Faris $\rightarrow$ USA $\rightarrow$ Washington), only to fail on the same query in a lower-resource language such as Italian (incorrectly reasoning: Anna Faris $\rightarrow$ Canada $\rightarrow$ Toronto). This performance gap, which can be as wide as 16% on benchmarks like StrategyQA (Geva et al. 2021), undermines the promise of truly global AI and risks reinforcing information inequity (Bender et al. 2021). This predicament raises a fundamental question: do these models lack knowledge in certain languages, or do they simply fail to access it during complex reasoning?

Previous attempts to answer this question have been inconclusive, particularly for multi-hop reasoning, due to methodological limitations. Interpretability research has largely operated at a coarse granularity. For instance, some studies analyze single-hop knowledge gaps by comparing hidden state similarity, identifying that errors often occur in the final layers but without explaining the intermediate reasoning steps crucial for multi-hop queries (Li et al. 2025; Zhao et al. 2024). Other work identifies language-specific versus language-agnostic neurons but stops short of explaining their semantic function or how they connect in a larger circuit (Tang et al. 2024; Zhang et al. 2024b). These methods reveal *that* differences exist, but not *why* they derail a step-by-step reasoning process. While fine-grained methods like Sparse Autoencoders (SAEs) offer deeper feature-level insights (Marks et al. 2025; Cunningham et al. 2024; Bricken et al. 2023), their prohibitive training costs and the difficulty in translating their findings into actionable model fixes make them impractical for targeted repair. Meanwhile, alignment techniques such as Direct Preference Optimization (DPO) (She et al. 2024; Yang et al. 2025), while bypassing the need for mechanistic understanding, treat the model as a black box and suffer from severe catastrophic forgetting (Qi et al. 2024; Winata et al. 2023); for instance, improving Japanese accuracy can eradicate Arabic performance, causing it to plummet from 50% to near zero. This leaves the field at an impasse: coarse analyses are uninformative for complex reasoning, deep analyses are impractical for repair, and black-box fixes are destructive.

To resolve this impasse, we introduce a circuit discovery framework that pinpoints the precise neural mechanisms—*the specific pathways of neurons and attention heads*—underlying cross-lingual knowledge disparities in multi-hop reasoning. As illustrated in Figure 1, our approach moves beyond attribution-based methods that assess components in isolation (Nanda 2024; Wang et al. 2022) by tracing the causal pathways of information flow. It identifies hierarchically organized circuits of neurons and attention heads responsible for reasoning, allowing us not only to identify critical components but also to understand their functional roles and interconnections. Our analysis, visualized in Figure 2, reveals a striking discovery that resolves the central paradox.

The cross-lingual knowledge gap is not a deficit of knowl-

*Corresponding author.

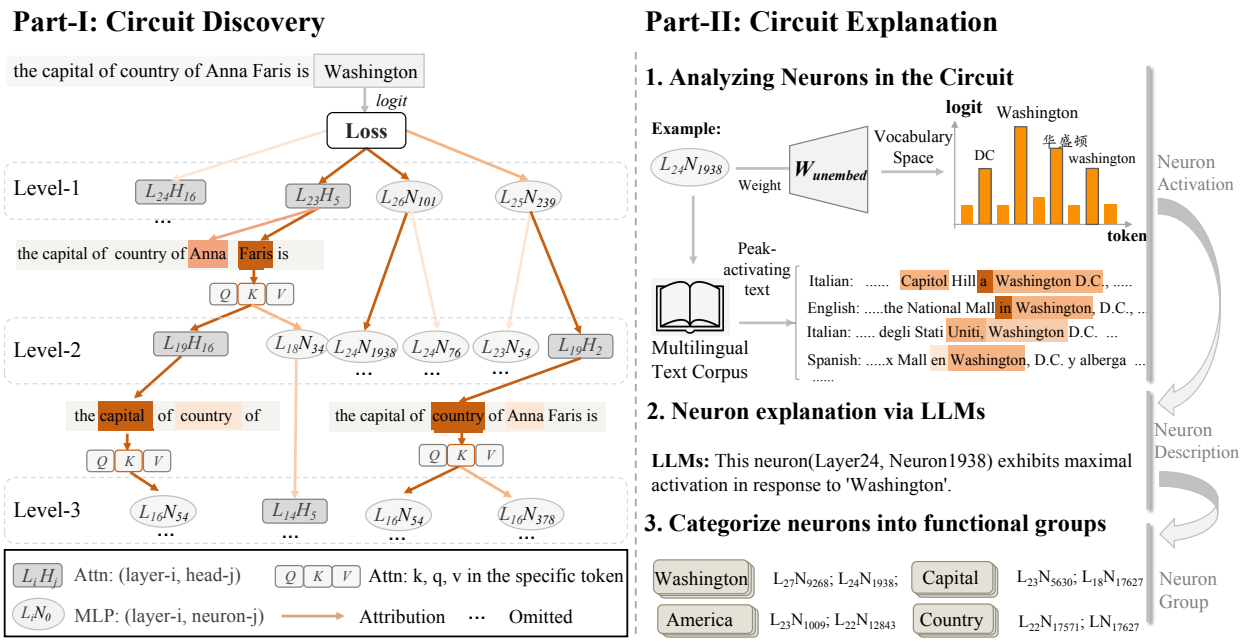


Figure 1: Overview of Our Mechanistic Interpretability Framework. Left (Circuit Discovery): Starting from the output logit, we use attribution patching to recursively identify a causal circuit of high-importance neurons (L_iN_j) and attention heads (L_iH_j). The process traces critical connections, including multi-token interactions via QK-circuits. Right (Circuit Interpretation): (1) analyze each key neuron's role via vocabulary projection and its top activating text examples, (2) synthesize these into a functional description, (3) cluster related neurons into interpretable groups.

edge, but a failure of access. We find that factual knowledge is stored in a set of **shared, language-agnostic semantic neurons**. As shown in Figure 4, the same neurons representing "United States of America" are activated across English, Chinese, and Arabic. The point of failure lies in the **misaligned attention pathways** that are supposed to route information to these neurons. Figure 3 provides a clear example: a single, shared attention head correctly attends to the "America" neuron group when processing an English query, but incorrectly attends to the "Canada" group for the same query in Italian. This insight is causally verifiable: by manually amplifying the "America" neuron's activation, we can correct the model's reasoning in Italian, forcing it to output "Washington" instead of "Toronto" (Figure 2).

This mechanistic insight directly motivates a novel and surgical solution. If knowledge is present and only access is broken, large-scale retraining is unnecessary. Instead, we propose **selectively fine-tuning only the small set of critical attention heads** responsible for these misaligned pathways. Our experiments confirm the efficacy of this mechanistically-informed approach. By fine-tuning just the top-40 most critical heads, we achieve a 15% relative improvement in factual accuracy on our XHopReason test set, with positive cross-lingual transfer to unseen languages. Crucially, this targeted intervention avoids the catastrophic forgetting that plagues traditional methods, preserving general model capabilities on benchmarks like MMLU (Table 3).

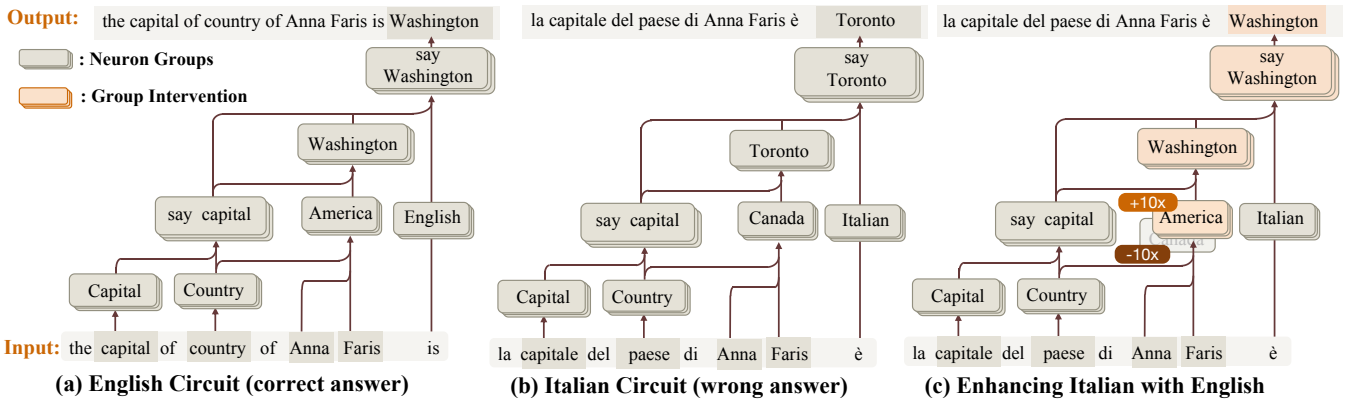
Our contributions are threefold:

1. We introduce a circuit discovery framework that provides a comprehensive, neuron-level explanation of multilingual knowledge reasoning, revealing the precise mechanisms of information flow.
2. We present a novel mechanistic insight: cross-lingual inconsistencies in multi-hop reasoning arise not from knowledge deficits but from misaligned attention pathways connecting to a shared, language-agnostic knowledge base.
3. We develop a highly effective and scalable alignment method that surgically fine-tunes critical attention heads, resolving knowledge gaps while uniquely preserving the model's general capabilities.

2 Methodology

Our goal is to uncover the precise neural mechanisms responsible for cross-lingual reasoning failures and leverage this understanding to develop a targeted alignment strategy. To achieve this, we first formally define the problem and introduce our core analytical tool (Section 2.1). We then detail our circuit discovery framework (Section 2.2) and the functional interpretation process (Section 2.3). Finally, we present the mechanistically-informed alignment method derived from our findings (Section 2.4).

[1]The query shown, "the capital of country of Anna Faris is", is a simplified version of the actual prompt ("the capital of the country of citizenship of Anna Faris is") used for clearer visualization.



(a) English Circuit (correct answer) **(b) Italian Circuit (wrong answer)** **(c) Enhancing Italian with English**

Figure 2: Cross-lingual knowledge circuits reveal misaligned information routing. Knowledge circuits for the same factual query across languages.[1]Gray boxes represent neuron groups; orange highlights show interventions. (a) English correctly routes to "America" neurons, producing "Washington." (b) Italian erroneously routes to "Canada" neurons, producing "Toronto." (c) Manual intervention corrects the Italian pathway.

2.1 Preliminaries: Problem Formulation and Causal Analysis

Problem Definition. The central problem is to identify the causal circuit underlying a specific multi-hop reasoning task and pinpoint why it functions correctly in a high-resource language but fails in a low-resource one. Given a model M and a factual query, we consider two inputs: a high-resource language input x_h (e.g., English) for which the model produces the correct answer A_r, and a low-resource language input x_l (e.g., Italian) for which it produces an incorrect answer A_c. Our objective is to locate the sub-network of parameters $\mathcal{C} \subset M$—comprising specific neurons and attention heads—whose divergent behavior between processing x_h and x_l explains the difference in output.

Background: Identifying Critical Components via Attribution Patching. To identify these critical components, our methodology builds upon attribution patching (Nanda 2024; Kramár et al. 2024), a gradient-based approximation of activation patching (Wang et al. 2023, 2022). This technique allows us to efficiently estimate the causal effect of a component on the model's output. We construct a contrastive pair of inputs: a *reference input* x_r that elicits the correct reasoning pathway and a *corrupted input* x_c that does not. The model's performance is measured by a scoring function, typically the logit difference $f(x) = \text{logit}(A_r|x) - \text{logit}(A_c|x)$.

The core idea is to quantify how patching a component's activation from the reference forward pass ($a_n(x_r)$) to the corrupted pass ($a_n(x_c)$) restores the correct output. This effect is measured by a normalized metric (Nanda 2024):

$$\text{Metric}(x_p) = \frac{f(x_p) - f(x_c)}{f(x_r) - f(x_c)}, \tag{1}$$

where a score near 1 indicates a component is critical. Attribution patching approximates this score for each component n in a single backward pass using a first-order Taylor expansion:

$$A(n) \approx \nabla_{a_n(x_c)}\text{Metric} \cdot (a_n(x_r) - a_n(x_c)), \tag{2}$$

where $a_n(x)$ is the activation of component n. This forms the foundational tool for our circuit discovery process.

2.2 Causal Circuit Discovery

Building on this foundation, we trace the full causal pathway from input to output, as illustrated in **Figure 1 (left)**.

Recursively Tracing Causal Pathways. After using attribution patching (Eq. 2) to identify components with high direct influence on the final output, we recursively trace their causal predecessors. The influence of an upstream component n_2 on a downstream component n_1 is computed as:

$$A(n_1, n_2) \approx \nabla_{a_{n_1}(x_c)}\text{Metric} \cdot (a_{n_2}(x_r) - a_{n_2}(x_c)). \tag{3}$$

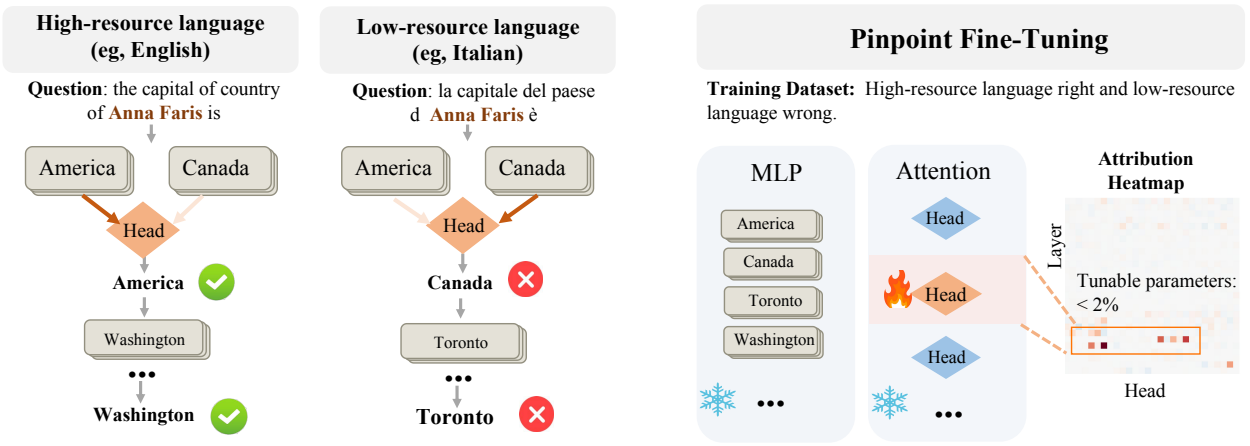
This quantifies how activation changes propagate through the network.[2] By iteratively identifying upstream components whose attribution $A(n_{l,i}, n_{l-1,j})$ exceeds a threshold τ, we construct a directed acyclic graph representing the full causal circuit.

Uncovering Cross-Token Information Flow via QK Circuits. A key limitation of prior work(Wang et al. 2023; Jack Lindsey and et al 2023) is its failure to explain how attention heads gather information from distant tokens. To address this, we analyze the complete QK circuit. For each critical attention head $h_{l,i}$, we first identify its most influential source token, t_{focus}, by applying attribution to its attention probabilities. We then recursively trace the origins of the key vector $K_{t_{\text{focus}}}$ by treating it as an upstream component and computing its attribution on the head's output. This cross-token analysis, visually represented by the Q-K-V blocks in **Figure 1 (left)**, reveals the complete causal chain of how information is selected and moved.

2.3 Functional Interpretation of Circuits

Discovering a circuit is insufficient; we must understand its function. As outlined in **Figure 1 (right)**, we employ a

[2]This is a first-order approximation. To handle dimensional differences, calculations are projected into the residual stream. See Appendix H.



(a) Cross-Lingual Inconsistency from Divergent Attention Head Behavior

(b) Resolving Inconsistency with Precise Attention Head Fine-Tuning

Figure 3: Identifying and Correcting Cross-Lingual Failures via Attention Head Tuning. (a) Cross-Lingual Inconsistency from Divergent Attention Head Behavior. Shared attention heads exhibit different behaviors across languages. In English (high-resource), the head correctly attends to the "America" neuron group. In Italian (low-resource), the same head incorrectly attends to the "Canada" group, causing a factual error. (b) Resolving Inconsistency with Precise Attention Head Fine-Tuning. A precisely fine-tuning approach corrects the misbehaving attention head. By adjusting a small subset of parameters, the head learns to route information correctly in the low-resource language, resolving the inconsistency.

multi-pronged strategy to interpret the roles of its constituent neurons. **Neuron Function Analysis.** For each neuron, we synthesize two sources of information: **(a) Vocabulary Projection**: We project neuron weights onto the vocabulary space ($P_{\text{vocab}} = W_{\text{out}} \cdot W_{l,i}$) to identify which output tokens it most strongly promotes or suppresses (Geva et al. 2022). **(b) Activation Analysis**: We identify text examples from a diverse corpus that maximally activate the neuron, revealing the contextual patterns it responds to (Choi et al. 2024).

LLM-based Interpretation and Grouping. We then leverage an LLM (GPT-4o) to synthesize these two data streams into a concise, natural-language description of the neuron's function. This scalable approach allows us to move from low-level data to high-level concepts. Finally, to create a more abstract and manageable view of the circuit, we aggregate semantically related neurons into functional groups or "super-neurons" (e.g., "America-related", "Capital-city-related"), the final step illustrated in **Figure 1 (right)**. This grouping is inspired by methodologies in (Ester et al. 1998).

2.4 Mechanistically-Informed Cross-Lingual Alignment

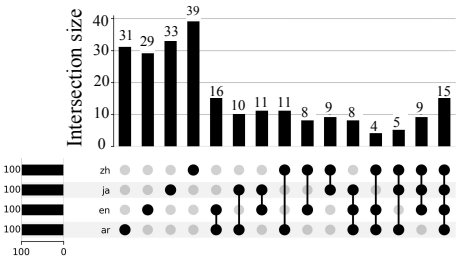
Our circuit analysis revealed that cross-lingual failures stem from misaligned attention pathways rather than a fundamental knowledge deficit. This diagnosis motivates a surgical intervention. Inspired by recent work on tracing and editing internal model components (Wang et al. 2025; Chen et al. 2024; Xiao et al. 2024), we propose mechanistically-informed fine-tuning (MFT) as an alternative to costly, full-model retraining. As detailed in Algorithm 1 (Appendix), MFT first identifies a small set of critical attention heads via attribution scores. Then, we fine-tune *only* the parameters of these top-k heads (W_Q, W_K, W_V, W_O), keeping the rest of the model frozen. By targeting the precise locus of failure (typically top-40 heads), this approach efficiently realigns faulty attention patterns while causing minimal disruption to the model's vast repository of learned capabilities.

3 Experimental Analysis

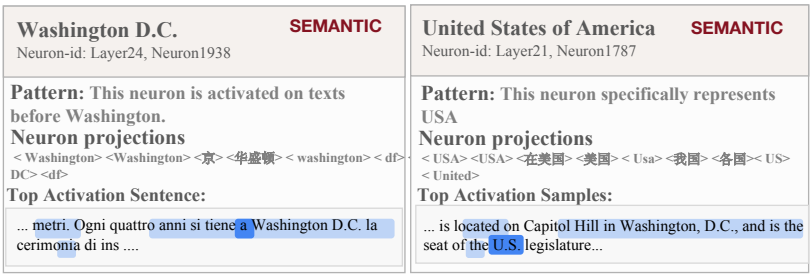
In this section, we empirically validate our hypotheses. We first detail the experimental setup, then present a mechanistic analysis that (1) identifies a shared, language-agnostic knowledge representation layer, and (2) pinpoints misaligned attention heads as the root cause of cross-lingual errors. We then causally validate these discovered circuits before demonstrating that a mechanistically-informed alignment method, targeting these specific heads, effectively bridges cross-lingual knowledge gaps while preserving general model capabilities.

3.1 Experimental Setup

Dataset. We construct XHOPREASON, a multi-hop reasoning dataset, by translating 82,020 queries from (Biran et al. 2024) into seven languages, ensuring semantic fidelity through a multi-stage validation process. This overcomes limitations of existing benchmarks like BMLAMA17 (Qi, Fernández, and Bisazza 2023) and KLAR (Li et al. 2025), which are unsuitable for analyzing multi-step reasoning. We partition data for circuit discovery (100 samples), alignment training (2,000 inconsistent pairs per language), and testing (1,000 non-overlapping samples per language). General capabilities are evaluated on MMLU (Hendrycks et al. 2020), StrategyQA (Geva et al. 2021), INCLUDE (Team 2023), and Belebele (Bandarkar et al. 2023). Further details are in Appendix D.



(a) Cross-lingual Knowledge Neuron Overlap



(b) Semantic Neuron Functionality Analysis

Figure 4: Structure of a Cross-lingual Knowledge Circuit. (a) An UpSet plot, based on 100 reasoning queries, reveals a large core of shared knowledge neurons alongside distinct language-specific neurons. (b) Functional analysis of two shared neurons confirms their language-agnostic semantic roles (e.g., "Washington D.C." and "United States of America") through activation patterns and vocabulary projections.

Alignment Method	English (en)	Japanese (ja)	Chinese (zh)	Arabic (ar)	Polish (pl)	Hebrew (he)	Vietnamese (vi)	Italian (it)	Average
Baseline (No-Op)	23.02	9.99	19.36	11.57	11.94	8.77	18.15	19.96	15.34
Full Fine-tuning	12.30 -10.72	22.35 +12.36	18.21 -1.15	13.27 +1.70	12.30 +0.36	7.31 -1.46	16.86 -1.29	18.82 -1.14	15.17 -0.17
DPO	15.50 -7.52	24.10 +14.11	18.50 -0.86	13.40 +1.83	12.50 +0.56	7.50 -1.27	17.20 -0.95	19.10 -0.86	15.98 +0.64
Mechanistically-Informed Head-Tuning (Ours)									
SFT Bottom-100	26.67 +3.65	17.05 +7.06	20.29 +0.93	12.67 +1.10	13.27 +1.33	8.86 +0.09	20.31 +2.16	21.53 +1.57	17.58 +2.24
Top-20 Heads	27.28 +4.26	19.61 +9.62	20.95 +1.59	13.39 +1.82	13.28 +1.34	9.98 +1.21	20.71 +2.56	20.04 +0.08	18.16 +2.82
Top-40 Heads	**27.41** +4.39	**20.28** +10.29	21.19 +1.83	13.58 +2.01	**14.01** +2.07	9.20 +0.43	21.38 +3.23	**22.05** +2.09	**18.64** +3.30
Top-80 Heads	**27.41** +4.39	19.12 +9.13	**22.29** +2.93	**14.12** +2.55	13.89 +1.95	8.04 -0.73	20.34 +2.19	21.92 +1.96	18.39 +3.05
Top-160 Heads	27.28 +4.26	19.61 +9.62	21.19 +1.83	13.27 +1.70	12.30 +0.36	**9.20** +0.43	20.28 +2.13	20.95 +0.99	18.01 +2.67

Table 1: Comparison of Alignment Methods on the XHopReason Test Set. All models were trained on 2,000 Japanese examples. Each cell shows accuracy (%) with the change from baseline underneath (gain / loss). Our head-tuning method significantly outperforms baselines, with Top-40 head-tuning achieving the best results. Best performance in each column is bolded.

Models. Our primary experiments use Qwen2.5-7B-Instruct (Yang et al. 2024) for its strong multilingual performance and manageable scale for interpretability. We also validate our method on Llama-3-8B-Instruct (Grattafiori et al. 2024) (see Appendix G).

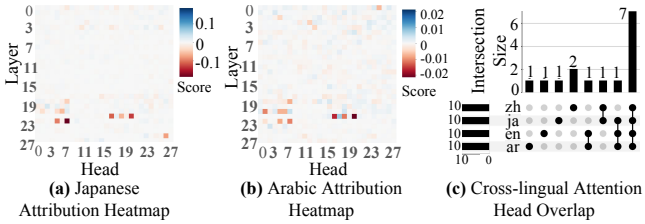
Implementation Details. Experiments were run on a single NVIDIA A100 GPU, with circuit discovery and analysis taking approximately 4 hours.

3.2 Cross-Lingual Knowledge Resides in Shared Semantic Neurons

Our first major finding is that factual knowledge is represented by a common set of language-agnostic semantic neurons. Analyzing multi-hop reasoning circuits, we found a substantial overlap in the critical neurons activated for the same task across languages. This is demonstrated in Figure 4(a), which aggregates results from 100 distinct queries, and the phenomenon was consistent across all tested reasoning types.

Crucially, our analysis also reveals that the complete circuit includes a complementary set of language-specific neurons. These are not a contradiction but are responsible for language-specific processing, enabling the precise control over output language that we demonstrate in Section 3.4 (**Functional Specificity**).

Functional analysis confirms the role of these shared neurons: they encode abstract concepts like "United States of America" irrespective of the input language (Figure 4(b)) and are concentrated in deeper model layers (Figure 10 in Appendix). We further validate their causal role in factual recall. As abstracted in Figure 2, manually amplifying the "America" neuron group in an Italian query corrects a factual error, providing decisive evidence that knowledge is stored in a shared, functionally specific substrate. This is further illustrated in our interactive visualization (Appendix 7).



(a) Japanese Attribution Heatmap **(b)** Arabic Attribution Heatmap **(c)** Cross-lingual Attention Head Overlap

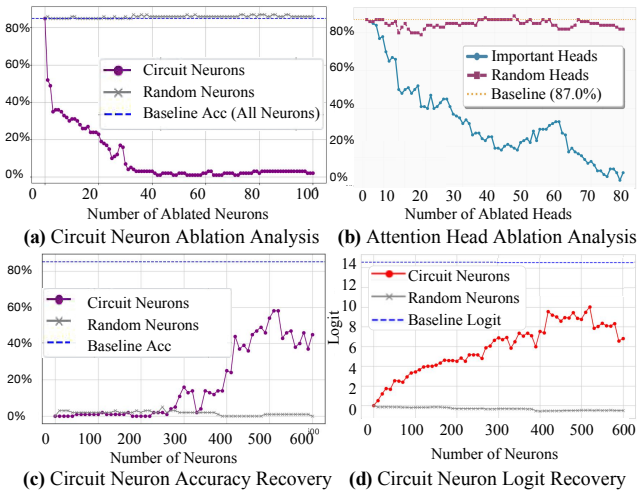
Figure 5: Cross-lingual Consistency of Critical Attention Heads. Critical heads, identified on 100 diverse queries per language, show remarkable consistency. (a, b) Attribution patterns are highly similar even for distant languages like Japanese and Arabic. (c) An UpSet plot reveals that 70% of the top-10 critical heads are shared across four languages, indicating a common reasoning circuit.

Training Language	Japanese (Ja)	Chinese (Zh)	Arabic (Ar)	Polish (Pl)	Hebrew (He)	Vietnamese (Vi)	Italian (It)
Baseline	9.99	19.36	11.57	11.94	8.77	18.15	19.96
Ja	**20.28**	21.19	13.58	14.01	9.20	21.38	22.05
	+10.29	+1.83	+2.01	+2.07	+0.43	+3.23	+2.09
Zh	13.39	**25.91**	13.52	14.74	**11.21**	20.82	22.04
	+3.40	+6.55	+1.95	+2.80	+2.44	+2.67	+2.08
Ar	14.01	21.01	**15.22**	15.10	10.23	**23.75**	22.06
	+4.02	+1.65	+3.65	+3.16	+1.46	+5.60	+2.10
Pl	12.42	20.23	13.15	14.98	10.48	21.19	23.39
	+2.43	+0.87	+1.58	+3.04	+1.71	+3.04	+3.43
He	12.67	22.17	13.03	13.64	9.99	20.46	20.59
	+2.68	+2.81	+1.46	+1.70	+1.22	+2.31	+0.63
Vi	12.67	21.32	14.13	14.62	10.60	21.92	21.92
	+2.68	+1.96	+2.56	+2.68	+1.83	+3.77	+1.96
It	13.39	21.92	14.34	**16.08**	9.99	21.44	**24.24**
	+3.40	+2.56	+2.77	+4.14	+1.22	+3.29	+4.28

Table 2: Mechanistically-informed head-tuning shows strong positive cross-lingual transfer. We train the top-40 critical heads using data from a single source language (rows) and evaluate on all target languages (columns). Nearly all interventions improve accuracy on unseen languages, not just the training language. Each cell shows accuracy (%), with the gain over baseline in gray.

3.3 Misaligned Attention Pathways Cause Cross-Lingual Errors

Having established that knowledge is shared, we found that access to it is mediated by a consistent set of attention heads that act as information routers. Across diverse languages, attribution heatmaps for critical heads are remarkably similar (Figure 5(a-b) and Appendix 8), with 70% of the top-10 heads being shared across all tested languages (Figure 5(c)). The crux of cross-lingual failure lies in the divergent behavior of these shared components. As illustrated in Figure 3, the same reasoning head that correctly attends to the "America" neuron group in English incorrectly attends to the "Canada" group in Italian, causing a factual error. This demonstrates that inconsistencies arise not from a knowledge deficit, but from misaligned attentional pathways, a mechanistic insight



(a) Circuit Neuron Ablation Analysis **(b)** Attention Head Ablation Analysis **(c)** Circuit Neuron Accuracy Recovery **(d)** Circuit Neuron Logit Recovery

Figure 6: Causal Validation of the Knowledge Circuit's Necessity and Sufficiency. (a, b) Necessity: Ablating the top 50 circuit neurons (a) or top 40 heads (b) nearly eliminates task accuracy, while ablating random components has no effect. (c, d) Sufficiency: After full model ablation, restoring only the circuit neurons progressively recovers task accuracy (c) and correct token logits (d), unlike restoring random neurons.

that pinpoints the routing behavior of a small, identifiable set of heads as the precise point of failure.

3.4 Causal Validation of Discovered Circuits

To confirm that our discovered circuits are not merely correlational, we conducted rigorous validation experiments.

Necessity and Sufficiency: We first established causal necessity via targeted ablation. As shown in Figure 6 (a,b), ablating just 50 identified knowledge neurons or the top 40 critical attention heads collapses performance to near-zero, while random ablation has a negligible effect. We then confirmed sufficiency by ablating the entire model and restoring only the neurons from the "Washington" reasoning circuit. Restoring just 500 neurons (≈0.2%) recovers accuracy to over 60% (Figure 6(c)), demonstrating that the identified subnetwork is both necessary and sufficient for the task.

Functional Specificity: We verified the distinct roles of circuit components through intervention. First, as shown in Figure 2, amplifying specific knowledge neurons corrects factual errors, confirming their role in storing specific facts. Second, we demonstrated a clear separation of knowledge and language processing. By activating Chinese-specific neurons while suppressing their Italian counterparts—leveraging the language-specific neurons identified in our circuit analysis—we induced the model to answer an Italian query with a factually correct response in Chinese (Figure 11 in Appendix and Appendix 9). This confirms that our method successfully disentangles language-specific circuits from the shared knowledge circuits.

Alignment Method	**StrategyQA** (Complex Reasoning) en	ja	zh	**MMLU** (General Knowledge) en	zh	ja	**INCLUDE** (Commonsense) ar	id	ja	**Belebele** (Reading Comp.) ar	ja	vi	**Overall** Avg. Δ
Baseline (No-Op)	56.83	57.21	55.89	74.70	66.95	62.05	57.97	61.82	72.85	84.00	82.89	83.33	–
Full Fine-tuning	30.31 -26.52	47.60 -9.61	44.98 -10.91	6.10 -68.60	2.87 -64.08	0.07 -61.98	50.36 -7.61	46.36 -15.46	66.47 -6.38	83.22 -0.78	79.78 -3.11	53.33 -30.00	-24.58
DPO	35.15 -21.68	50.40 -6.81	48.20 -7.69	10.50 -64.20	5.70 -61.25	2.10 -59.95	52.80 -5.17	51.20 -10.62	68.50 -4.35	83.50 -0.50	80.10 -2.79	55.70 -27.63	-22.72
Mechanistically-Informed Head-Tuning (Ours)													
Top-40 Heads	**61.13** +4.30	**54.58** -2.63	**54.15** -1.74	**74.06** -0.64	**65.47** -1.48	**58.16** -3.89	**56.34** -1.63	**62.91** +1.09	**73.65** +0.80	**82.89** -1.11	**80.11** -2.78	**83.78** +0.45	-0.85

Table 3: General Capability Preservation on Multilingual Benchmarks. All models were trained on 2,000 Japanese examples from XHOPREASON. Each cell shows accuracy (%) with the change from baseline underneath (gain / loss). Our head-tuning method (Top-40) successfully preserves general capabilities, while conventional methods (Full Fine-tuning, DPO) suffer from catastrophic forgetting. The best result in each column is bolded.

3.5 Mechanistically-Informed Alignment Corrects Cross-Lingual Gaps

Our analysis motivates a precise alignment strategy: selectively fine-tuning the critical heads identified in Section 3.3. This surgical intervention repairs the faulty information flow without costly full-model retraining.

Effectiveness and Generalization: We benchmarked our head-tuning approach against full fine-tuning and DPO on the XHOPREASON dataset. As shown in Table 1, conventional methods fail to generalize. In contrast, our approach, which trains only the top-40 heads (a number informed by our ablation study, Figure 6), delivers robust improvements across all languages. For instance, training only on Japanese data not only improves Japanese accuracy but also yields gains in Italian and other languages (Table 2). This positive cross-lingual transfer provides strong evidence that our method repairs a shared, underlying reasoning mechanism rather than merely memorizing language-specific corrections.

Preservation of General Capabilities: A critical test for any alignment technique is avoiding catastrophic forgetting. We evaluated our model on a diverse suite of out-of-domain benchmarks (MMLU, StrategyQA, INCLUDE, Belebele). The results in Table 3 are stark: while full fine-tuning and DPO cause performance to plummet by up to 50 percentage points, our head-tuning method preserves performance at near-baseline levels across all tasks. This confirms our surgical intervention is both an effective and safe alignment strategy, repairing specific reasoning pathways without damaging the model's broader capabilities (see Appendix C for full results).

4 Related Work

Cross-lingual Knowledge Analysis and Enhancement. Recent interpretability work on multilingual gaps, such as (Li et al. 2025), analyzed hidden state similarity for single-hop queries, finding that knowledge is language-agnostic but errors arise in the final language-specific adaptation. This layer-level analysis, however, lacks the granularity to explain multi-hop reasoning failures. Other studies identified language-agnostic and language-specific neurons (Tang et al. 2024; Zhang et al. 2024b; Wang et al. 2024), but stopped at identification without revealing their semantic function or tracing the causal pathways that connect them. To address performance gaps, black-box alignment methods like DPO (She et al. 2024; Yang et al. 2025) have been explored, but they often lead to catastrophic forgetting (Qi et al. 2024; Winata et al. 2023). As our experiments confirm (e.g., training on Japanese can collapse Arabic accuracy), these methods risk harming shared circuits without mechanistic insight.

Mechanistic Interpretability Methods. Our research builds on mechanistic interpretability (Olah et al. 2020; Madsen, Reddy, and Chandar 2023; Räuker et al. 2023). One prominent paradigm uses Sparse Autoencoders (SAEs) to find interpretable features (Marks et al. 2025; Cunningham et al. 2024; Bricken et al. 2023; Gao et al. 2024; Zhang et al. 2024a). While powerful, this approach requires costly, model-specific autoencoders, and its findings are difficult to translate into actionable repairs. Another paradigm, causal tracing (e.g., activation patching) (Wang et al. 2022; Conmy et al. 2023; Nanda 2024; Sun et al. 2025; Miao and Kan 2025; Mondorf, Wold, and Plank 2025), has successfully localized capabilities like induction heads (Olsson et al. 2022). However, these methods often fail to fully trace the cross-token QK circuit—how information from source tokens forms the key and query vectors (Rajamanoharan et al. 2024a,b). This is a critical limitation for multi-hop reasoning. Our approach extends these causal methods to map this entire information flow, from input tokens to QK vector construction and final outputs, enabling actionable analysis without SAEs.

5 Conclusion

In this work, we present a mechanistic explanation for cross-lingual knowledge disparities in LLMs, revealing that factual knowledge resides in shared, language-agnostic semantic neurons, while reasoning failures stem from misaligned attention pathways. By identifying these causal circuits, we developed a targeted fine-tuning strategy that selectively trains only the top-40 critical attention heads. This surgical intervention not only corrects factual errors with positive cross-lingual generalization but also preserves the model's general capabilities, offering a scalable and effective solution to enhance the reliability and equity of multilingual models without the catastrophic forgetting associated with traditional methods.

Acknowledgements

This work was supported in part by NSFC No. 62222117.

References

Bandarkar, L.; Liang, D.; Muller, B.; Artetxe, M.; Shukla, S. N.; Husa, D.; Goyal, N.; Krishnan, A.; Zettlemoyer, L.; and Khabsa, M. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*.

Biran, E.; Gottesman, D.; Yang, S.; Geva, M.; and Globerson, A. 2024. Hopping Too Late: Exploring the Limitations of Large Language Models on Multi-Hop Queries. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 14113–14130. Association for Computational Linguistics.

Bricken, T.; Templeton, A.; Batson, J.; Chen, B.; Jermyn, A.; Conerly, T.; Turner, N.; Anil, C.; Denison, C.; Askell, A.; Lasenby, R.; Wu, Y.; Kravec, S.; Schiefer, N.; Maxwell, T.; Joseph, N.; Hatfield-Dodds, Z.; Tamkin, A.; Nguyen, K.; McLean, B.; Burke, J. E.; Hume, T.; Carter, S.; Henighan, T.; and Olah, C. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Chen, W.; Huang, Z.; Xie, L.; Lin, B.; Li, H.; Lu, L.; Tian, X.; Cai, D.; Zhang, Y.; Wang, W.; et al. 2024. From yes-men to truth-tellers: addressing sycophancy in large language models with pinpoint tuning. *arXiv preprint arXiv:2409.01658*.

Choi, D.; Huang, V.; Meng, K.; Johnson, D. D.; Steinhardt, J.; and Schwettmann, S. 2024. Scaling Automatic Neuron Description. https://transluce.org/neuron-descriptions. Accessed: 2025-11-14.

Conmy, A.; Mavor-Parker, A. N.; Lynch, A.; Heimersheim, S.; and Garriga-Alonso, A. 2023. Towards automated circuit discovery for mechanistic interpretability. *Thirty-seventh Conference on Neural Information Processing Systems*.

Cunningham, H.; Ewart, A.; Riggs, L.; Huben, R.; and Sharkey, L. 2024. Sparse autoencoders find highly interpretable features in language models. *The Twelfth International Conference on Learning Representations*.

Ester, M.; Kriegel, H.; Sander, J.; Wimmer, M.; and Xu, X. 1998. Incremental Clustering for Mining in a Data Warehousing Environment. In Gupta, A.; Shmueli, O.; and Widom, J., eds., *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York City, New York, USA*, 323–333. Morgan Kaufmann.

Gao, L.; Dupre la Tour, T.; Tillman, H.; Goh, G.; Troll, R.; Radford, A.; Sutskever, I.; Leike, J.; and Wu, J. 2024. Scaling and evaluating sparse autoencoders. *Computing Research Repository*.

Geva, M.; Caciularu, A.; Wang, K. R.; and Goldberg, Y. 2022. Transformer Feed-Forward Layers Build Predictions by Promoting Concepts in the Vocabulary Space. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Guo, Y.; Liang, Y.; Zhao, D.; Liu, B.; and Duan, N. 2023. Analyzing and Reducing the Performance Gap in Cross-Lingual Transfer with Fine-tuning Slow and Fast. In *ACL*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jack Lindsey, E. A., Wes Gurnee; and et al. 2023. On the Biology of a Large Language Model. *blog:https://transformer-circuits.pub/2025/attribution-graphs/biology.html*. Accessed: 2025-11-14.

Kramár, J.; Lieberum, T.; Shah, R.; and Nanda, N. 2024. AtP*: An efficient and scalable method for localizing LLM behaviour to components. *CoRR*, abs/2403.00745.

Li, Z.; et al. 2025. 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*.

Madsen, A.; Reddy, S.; and Chandar, S. 2023. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Comput. Surv.*, 55(8): 155:1–155:42.

Marks, S.; Rager, C.; Michaud, E. J.; Belinkov, Y.; Bau, D.; and Mueller, A. 2025. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *International Conference on Learning Representations*.

Miao, Y.; and Kan, M.-Y. 2025. Discursive Circuits: How Do Language Models Understand Discourse Relations? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 32558–32577.

Mondorf, P.; Wold, S.; and Plank, B. 2025. Circuit compositions: Exploring modular structures in transformer-based language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14934–14955.

Nanda, N. 2024. Attribution Patching: Activation Patching At Industrial Scale. *blog:https://www.neelnanda.io/mechanistic-interpretability/attribution-patching*. Accessed: 2025-11-14.

Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3): e00024–001.

Olsson, C.; Elhage, N.; Nanda, N.; Joseph, N.; DasSarma, N.; Henighan, T.; Mann, B.; Askell, A.; Bai, Y.; Chen, A.; Conerly, T.; Drain, D.; Ganguli, D.; Hatfield-Dodds, Z.; Hernandez, D.; Johnston, S.; Jones, A.; Kernion, J.; Lovitt, L.; Ndousse, K.; Amodei, D.; Brown, T.; Clark, J.; Kaplan, J.; McCandlish, S.; and Olah, C. 2022. In-context learning and induction heads. Transformer Circuits Thread.

Qi, B.; Li, P.; Li, F.; Gao, J.; Zhang, K.; and Zhou, B. 2024. Online dpo: Online direct preference optimization with fast-slow chasing. *arXiv preprint arXiv:2406.05534*.

Qi, J.; Fernández, R.; and Bisazza, A. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. In *EMNLP*.

Rajamanoharan, S.; Conmy, A.; Smith, L.; Lieberum, T.; Varma, V.; Kramar, J.; Shah, R.; and Nanda, N. 2024a. Improving dictionary learning with gated sparse autoencoders. *Computing Research Repository*.

Rajamanoharan, S.; Lieberum, T.; Sonnerat, N.; Conmy, A.; Varma, V.; Kramar, J.; and Nanda, N. 2024b. Jumping ahead: Improving reconstruction fidelity with jumprelu sparse autoencoders. *Computing Research Repository*.

Räuker, T.; Ho, A.; Casper, S.; and Hadfield-Menell, D. 2023. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023, Raleigh, NC, USA, February 8-10, 2023*, 464–483. IEEE.

She, S.; Zou, W.; Huang, S.; Zhu, W.; Liu, X.; Geng, X.; and Chen, J. 2024. MAPO: Advancing Multilingual Reasoning through Multilingual-Alignment-as-Preference Optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*.

Sun, C.; Huang, Z.; Zhang, Y.; Lu, L.; Li, H.; Tian, X.; Shen, X.; and Ye, J. 2025. Interpret and Improve In-Context Learning via the Lens of Input-Label Mappings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3873–3895.

Tang, T.; et al. 2024. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.

Team, C. 2023. INCLUDE: Evaluating Multilingual Language Models with Exam-Level Multiple Choice Questions. https://huggingface.co/datasets/Cohere/wikibio-include-base-44. Accessed: 2025-07-28.

Wang, K.; Variengien, A.; Conmy, A.; Shlegeris, B.; and Steinhardt, J. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 small. *CoRR*.

Wang, K. R.; Variengien, A.; Conmy, A.; and et al. 2023. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2 Small. In *ICLR*.

Wang, W.; Haddow, B.; Wu, M.; Peng, W.; and Birch, A. 2024. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Wang, Y.; Wan, C.; Hu, S.; Zhang, Y.; Tian, X.; Chen, Y.; Shen, X.; and Ye, J. 2025. Tracing and Dissecting How LLMs Recall Factual Knowledge for Real World Questions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 23246–23271.

Winata, G. I.; Xie, L.; Radhakrishnan, K.; Wu, S.; Jin, X.; Cheng, P.; Kulkarni, M.; and Preotiuc-Pietro, D. 2023. Overcoming catastrophic forgetting in massively multilingual continual learning. *arXiv preprint arXiv:2305.16252*.

Xiao, Y.; Chaoqun, W.; Zhang, Y.; Wang, W.; Lin, B.; He, X.; Shen, X.; and Ye, J. 2024. Enhancing multiple dimensions of trustworthiness in LLMs via sparse activation control. *Advances in Neural Information Processing Systems*, 37: 15730–15764.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.

Yang, W.; Wu, J.; Wang, C.; Zong, C.; and Zhang, J. 2025. Language Imbalance Driven Rewarding for Multilingual Self-improving. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Zhang, W.; Wan, C.; Zhang, Y.; Cheung, Y.-m.; Tian, X.; Shen, X.; and Ye, J. 2024a. Interpreting and improving large language models in arithmetic calculation. *arXiv preprint arXiv:2409.01659*.

Zhang, Z.; Zhao, J.; Zhang, Q.; Gui, T.; and Huang, X. 2024b. Unveiling Linguistic Regions in Large Language Models. In Ku, L.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, 6228–6247. Association for Computational Linguistics.

Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024. How do Large Language Models Handle Multilingualism? In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.