

Abstract Activation Spaces for Content-Invariant Reasoning in Large Language Models

Gabriele Maraia^(†) Marco Valentino^(•) Fabio Massimo Zanzotto^(†,‡) Leonardo Ranaldi^(⊕,†)

(⊕) ILCC, School of Informatics, University of Edinburgh, United Kingdom

(•) School of Computer Science, University of Sheffield, United Kingdom

(†) Human Centric ART & OMNIA Lab, University of Rome Tor Vergata, (‡) Almax Labs

{first_name.last_name}@uniroma2.it

Abstract

LLMs often struggle with deductive judgement in syllogistic reasoning, systematically conflating semantic plausibility with formal validity—a phenomenon known as *content effect*. This bias persists even when models generate step-wise explanations, indicating that intermediate rationales may inherit the same semantic shortcuts that affect answers. Recent approaches mitigate this issue by increasing inference-time structural constraints, either by encouraging abstract intermediate representations or by intervening in the model’s internal computations; yet, reliably suppressing semantic interference remains an open challenge.

To make formal deduction less sensitive to semantic content, we introduce a framework for abstraction-guided reasoning that separates structural inference from lexical semantics. We construct paired *content-laden* and *abstract* syllogisms and use the model’s activations on abstract inputs to define an *abstract reasoning space*. We then learn lightweight *Abstractors* that, from content-conditioned residual-stream states, predict representations aligned with this space and integrate these predictions via multi-layer interventions during the forward pass. Using cross-lingual transfer as a test bed, we show that abstraction-aligned steering reduces content-driven errors and improves validity-sensitive performance. Our results position activation-level abstraction as a scalable mechanism for enhancing the robustness of formal reasoning in LLMs against semantic interference.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities across a wide range of complex reasoning tasks. Yet, they frequently privilege semantic intuition over formal logic, systematically struggling to disentangle the *form* of an argument from its *content* (Eisape et al., 2024; Dasgupta et al., 2024; Ranaldi et al., 2026).

This limitation becomes particularly evident in syllogistic reasoning, a classical test bed for deductive competence, where models show *content effect*: a well-documented phenomenon in human cognition whereby the perceived plausibility of a conclusion overrides the logical validity of the premises. The conflict between semantic heuristics and logical rigour is a structural failure mode. Consider the following contrasting syllogistic cases:

Valid Implausible

Premise 1: All things that have fins live in the desert.

Premise 2: Dolphins have fins.

Conclusion: Therefore, dolphins live in the desert.

Invalid Plausible

Premise 1: All flowers need water.

Premise 2: Roses need water.

Conclusion: Therefore, roses are flowers.

In *Valid–Implausible* arguments, the conclusion follows deductively from the premises but conflicts with world knowledge; in *Invalid–Plausible* arguments, the conclusion is factually acceptable yet unsupported by the premises. In both cases, LLMs tend to align their judgements with semantic plausibility, often misclassifying valid arguments as invalid and accepting invalid ones as valid (Valentino et al., 2026a; Maraia et al., 2026). These behaviours indicate that models implicitly introduce semantic constraints into formal deduction, effectively fabricating logical flaws when conclusions contradict prior knowledge.

Existing strategies for mitigating content effect expose a trade-off between cost, modularity, and reliability. Supervised fine-tuning can improve accuracy but is computationally expensive and prone to reinforcing superficial correlations instead of inducing stable logical abstractions. CoT-based approaches offer a cheaper alternative, but they often reproduce the same semantic intuitions within intermediate reasoning steps, failing to resolve cases where world knowledge and logical validity con-

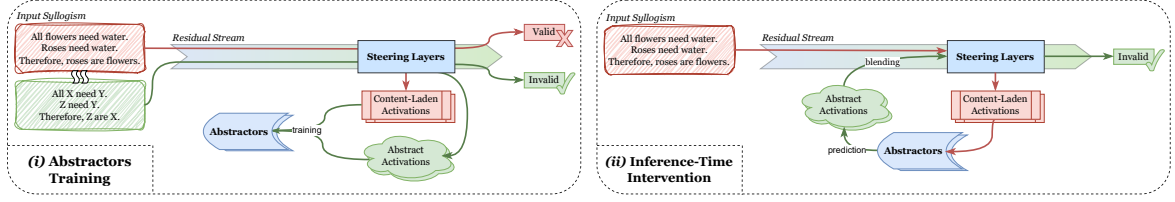


Figure 1: Overview of the abstraction steering framework. (i) The model processes both content-laden ("All **flowers** need **water** ...") and abstract ("All **X** need **Y** ...") syllogisms. Abstractors learn to map content-laden activations into the abstract reasoning space, (ii) integrating predicted targets via multi-layer activation at inference time.

flict (Wan et al., 2025). Moreover, these approaches do not ensure a separation between the logical structure and the semantic content at the internal representation level.

These observations suggest that the core issue lies not in the representational capacity of LLMs, but in how semantic and structural signals are routed during inference. Motivated by insights from mechanistic interpretability, we investigate **activation steering** as an inference-time intervention (Turner et al., 2024; Lucchetti and Guha, 2025). By selectively intervening on the residual stream, activation steering enables the attenuation of semantic content signals while amplifying representations associated with formal structure.

We introduce the **Abstractors**, lightweight Multi-Layer Perceptrons trained to map content-laden activations (specifically, the last-token activation of an input sequence, e.g., "All **flowers** need **water**") onto an *abstract reasoning manifold* (i.e., the activation of its symbolic counterpart, e.g., "All **X** need **Y**"), dynamically predicting the target activation for each input, unlike static steering vectors (see Figure 1). Furthermore, we evaluate the zero-shot, cross-lingual transfer of this manifold by training Abstractors exclusively on English data and evaluating them on nine other languages, both high- (e.g., Chinese) and low-resource (e.g., Telugu). This setting allows us to probe whether abstraction-aligned steering operates independently of linguistic surface form.

Overall, our contributions are as follows:

- We achieve validity-sensitive performance comparable to, and in some cases exceeding, state-of-the-art parameter-efficient methods, without modifying weights.
- We provide evidence that logical structure is encoded in a generalisable subspace, enabling inference-time transfer across unseen languages.
- Our approach offers a modular and inter-

pretable inference-time intervention that can be enabled or disabled on demand, preserving general language-modelling capabilities while specifically targeting belief bias.

Our results position activation-level abstraction as a scalable and language-agnostic mechanism for enhancing the robustness of formal reasoning in LLMs against semantic interference.

2 Method

2.1 Overview & Problem Formulation

Let x denote the complete input sequence containing the instruction and the syllogism, and let $y(x) \in \{\text{valid}, \text{invalid}\}$ denote the syllogism’s formal validity. Let $\mathcal{L} = \{0, \dots, n_{\text{layers}} - 1\}$ denote the set of zero-indexed transformer layers. For layer $\ell \in \mathcal{L}$ and token position $t \in \{1, \dots, T_x\}$, let

$$a_{\ell,t}(x) \in \mathbb{R}^d, \quad (1)$$

denote the residual-stream activation, where T_x is the number of tokens in x and d is the model’s hidden dimension. Let $\mathcal{L}^* \subseteq \mathcal{L}$ denote the selected steering layers. For every $\ell \in \mathcal{L}^*$, the sequence representation is the final-token activation

$$a_{\ell}(x) := a_{\ell,T_x}(x). \quad (2)$$

The intervention operates at inference time without modifying the model weights. We propose a method to dynamically predict a target representation $\hat{a}_{\ell}(x)$ that encodes logical structure independently of semantic content.

2.2 Abstract Reasoning Space as Target

The objective is to steer the model’s activations towards representations that encode the logical structures while abstracting away semantic noise.

To define these ideal representations, we construct a parallel dataset. Each content-laden syllogism x^{con} is paired with an abstract version x^{abs}

that preserves the logical form but replaces content words with abstract symbols (e.g., "All A are B "). The activations produced by the model when processing these abstract versions, $a_\ell(x^{\text{abs}})$, serve as our **target space**. We hypothesise that mapping $a_\ell(x^{\text{con}})$ onto this *abstract reasoning space* reduces reliance on semantic heuristics.

2.3 Unified Activation Prediction via Contrastive Learning

At inference time, the abstract counterpart x^{abs} is not available. Therefore, for each $\ell \in \mathcal{L}^*$, we train a single **Abstractor** to predict the corresponding abstract activation:

$$\hat{a}_\ell(x^{\text{con}}) = f_\ell(a_\ell(x^{\text{con}})) \approx a_\ell(x^{\text{abs}}). \quad (3)$$

Architecture. For each layer $\ell \in \mathcal{L}^*$, we train a unified Multi-Layer Perceptron (MLP) f_ℓ that handles both valid and invalid syllogisms.

- **Shared Backbone:** A deep feed-forward network h_ℓ extracts features from the input activation:

$$z_\ell(x^{\text{con}}) = h_\ell(a_\ell(x^{\text{con}})). \quad (4)$$

- **Direction Head:** Predicts the normalised unit vector

$$\hat{d}_\ell(x^{\text{con}}) = \frac{g_{d,\ell}(z_\ell(x^{\text{con}}))}{\max(\|g_{d,\ell}(z_\ell(x^{\text{con}}))\|_2, \varepsilon)}. \quad (5)$$

- **Magnitude Head:** Predicts the non-negative scalar magnitude

$$\hat{m}_\ell(x^{\text{con}}) = \text{softplus}(g_{m,\ell}(z_\ell(x^{\text{con}}))). \quad (6)$$

The final predicted abstract activation is:

$$\hat{a}_\ell(x^{\text{con}}) = \hat{m}_\ell(x^{\text{con}}) \hat{d}_\ell(x^{\text{con}}) = f_\ell(a_\ell(x^{\text{con}})). \quad (7)$$

Contrastive training strategy. To train the unified model to correctly handle both validity classes, we employ a contrastive learning approach with triplet construction. For each layer $\ell \in \mathcal{L}^*$ and training example x_i^{con} , let

$$y_i := y(x_i^{\text{con}}) \in \{\text{valid}, \text{invalid}\}, \quad (8)$$

and define the complementary validity label as

$$\bar{y}_i = \begin{cases} \text{invalid}, & \text{if } y_i = \text{valid}, \\ \text{valid}, & \text{if } y_i = \text{invalid}. \end{cases} \quad (9)$$

For each layer $\ell \in \mathcal{L}^*$, we build triplets of the form

$$(x_i^{\text{con}}, x_{i,\ell}^+, x_{i,\ell}^-), \quad (10)$$

where $x_{i,\ell}^+$ is an abstract example with validity y_i , and $x_{i,\ell}^-$ is an abstract example with validity $\bar{y}_i$.

Adaptive matching for triplet construction. To ensure high-quality training data, we employ a schema-based matching process. Let $\tilde{y}(x)$ denote the model's predicted validity for x , and let

$$\mathcal{C}_{\text{correct}} := \left\{ x^{\text{abs}} \mid \tilde{y}(x^{\text{abs}}) = y(x^{\text{abs}}) \right\} \quad (11)$$

denote the set of correctly answered abstract examples. For each $\ell \in \mathcal{L}^*$ and each x_i^{con} :

Positive Target Selection ($x_{i,\ell}^+$):

1. **Direct Match:** If the paired $x_i^{\text{abs}} \in \mathcal{C}_{\text{correct}}$ and has validity y_i , use it directly.
2. **Schema-Based Fallback:** Otherwise, select the nearest neighbour, using $\cos(a_\ell(x_i^{\text{con}}), a_\ell(x^{\text{abs}}))$, from $\mathcal{C}_{\text{correct}}$ with the same logical schema and validity y_i .
3. **Validity-Based Fallback:** If no schema match exists, select the nearest neighbour, using the same layer- ℓ cosine similarity, from $\mathcal{C}_{\text{correct}}$ with validity y_i .

Negative Counterfactual Selection ($x_{i,\ell}^-$): Select the nearest neighbour, using $\cos(a_\ell(x_i^{\text{con}}), a_\ell(x^{\text{abs}}))$, from $\mathcal{C}_{\text{correct}}$ with opposite validity $\bar{y}_i$, without schema constraints. A plain-text example of a triplet is shown in Table 1.

Combined loss function. For each layer $\ell \in \mathcal{L}^*$, the training objective combines three losses:

$$\mathcal{L}_{\text{total}}^{(\ell)} = \mathcal{L}_{\text{attract}}^{(\ell)} + \lambda_{\text{repel}} \mathcal{L}_{\text{repel}}^{(\ell)} + \lambda_{\text{mag}} \mathcal{L}_{\text{mag}}^{(\ell)}. \quad (12)$$

The three terms are defined as follows:

- **Attraction Loss** aligns the predicted direction with the positive target:

$$\mathcal{L}_{\text{attract}}^{(\ell)} = \mathbb{E}_i \left[1 - \cos(\hat{d}_\ell(x_i^{\text{con}}), d_\ell(x_{i,\ell}^+)) \right], \quad (13)$$

where

$$d_\ell(x_{i,\ell}^+) = \frac{a_\ell(x_{i,\ell}^+)}{\|a_\ell(x_{i,\ell}^+)\|_2}. \quad (14)$$

Content-Laden Input (x_i^{con}) <i>Validity:</i> valid <i>Plausibility:</i> True <i>Schema:</i> AI2 Premise 1: All cruisers are watercraft. Premise 2: Some warships are cruisers. Conclusion: Some warships are watercraft.
Positive Target ($x_{i,\ell}^+$) <i>Validity:</i> valid <i>Abstract:</i> True <i>Schema:</i> AI2 Premise 1: All X are Z. Premise 2: Some Y are X. Conclusion: Some Y are Z.
Negative Counterfactual ($x_{i,\ell}^-$) <i>Validity:</i> invalid <i>Abstract:</i> True <i>Schema:</i> Two Negative Premises (Matched via Layer- ℓ Latent Cosine Similarity) <i>Layer-ℓ Latent Cosine Similarity to Input:</i> 0.8750 Premise 1: No M are P. Premise 2: No S are M. Conclusion: No S are P.

Table 1: An example of a contrastive training triplet extracted during the matching phase. Given x_i^{con} , the positive target $x_{i,\ell}^+$ with validity y_i is matched via direct logical schema, while the negative counterfactual $x_{i,\ell}^-$ is selected by finding an example in $\mathcal{C}_{\text{correct}}$ with validity $\bar{y}_i$ and highest layer- ℓ latent cosine similarity to x_i^{con} .

- **Repulsion Loss** pushes the prediction away from the negative counterfactual:

$$\mathcal{L}_{\text{repel}}^{(\ell)} = \mathbb{E}_i \left[\text{ReLU} \left(\cos \left(\hat{d}_\ell(x_i^{\text{con}}), d_\ell(x_{i,\ell}^-) \right) - \mu \right) \right] \quad (15)$$

where $\mu \geq 0$ is the repulsion-margin threshold (we set $\mu = 0.2$), and

$$d_\ell(x_{i,\ell}^-) = \frac{a_\ell(x_{i,\ell}^-)}{\|a_\ell(x_{i,\ell}^-)\|_2}. \quad (16)$$

- **Magnitude Loss** matches the predicted magnitude to the positive target:

$$\mathcal{L}_{\text{mag}}^{(\ell)} = \mathbb{E}_i \left[\left(\hat{m}_\ell(x_i^{\text{con}}) - \|a_\ell(x_{i,\ell}^+)\|_2 \right)^2 \right]. \quad (17)$$

This formulation enables the unified model to learn a representation space where valid and invalid syllogisms are mapped to distinct regions, without requiring explicit class labels at inference time.

2.4 Inference-Time Steering

At inference time, given an input syllogism x , the target abstraction $\hat{a}_\ell(x)$ depends on the final token’s activation. Therefore, we employ a **two-pass inference strategy**:

1. **Pre-computation Pass:** We perform a standard forward pass on input x to extract the unsteered activation of the final token, $a_{\ell,T_x}^{(1)}(x)$, and compute the target

$$\hat{a}_\ell(x) = f_\ell \left(a_{\ell,T_x}^{(1)}(x) \right) \quad (18)$$

for each $\ell \in \mathcal{L}^*$, using the trained Abstractor.

2. **Steered Pass:** We re-process x . During this second pass, the pre-computed target $\hat{a}_\ell(x)$ is blended into the residual stream as described below. We use $a_{\ell,t}^{(2)}(x)$ to denote the second-pass activation immediately before the intervention at layer ℓ and token position t .

Steering layers. We steer multiple contiguous layers rather than a single layer to ensure robust propagation of the abstract representation throughout the forward pass. Target layers $\mathcal{L}^*$ are selected by identifying regions where positive and negative abstract targets exhibit maximal separation (lowest cosine similarity), typically occurring in the middle layers of the network where abstract concepts have formed but remain distinct. Appendix K visualises this layer-wise separation analysis across all models, with the selected steering layers corresponding to regions of minimal positive-negative similarity (full details in Appendix A.1).

Positional blending. The intervention intensity is modulated by a positional coefficient α_t that increases linearly along the sequence. Let $t_{\text{start}} \in \{1, \dots, T_x - 1\}$ be the first token position where the instruction ends and the actual syllogism begins. For $t \in \{1, \dots, T_x\}$, define

$$\alpha_t = \begin{cases} 0, & t < t_{\text{start}}, \\ \alpha \frac{t - t_{\text{start}}}{T_x - t_{\text{start}}}, & t \geq t_{\text{start}}, \end{cases} \quad (19)$$

where $\alpha \in [0, 1]$ is the maximum steering strength. Thus, $\alpha_{t_{\text{start}}} = 0$ and $\alpha_{T_x} = \alpha$.

Activation blending. The steered activation for token t at layer ℓ is computed as:

$$a_{\ell,t}^{\text{steer},(2)}(x) = (1 - \alpha_t) a_{\ell,t}^{(2)}(x) + \alpha_t \hat{a}_\ell(x). \quad (20)$$

For every $\ell \in \mathcal{L}^*$, the intervention is applied as defined above. During the Steered Pass, the single target vector $\hat{a}_\ell(x)$ (computed from the Pre-computation Pass) is broadcast to all content tokens $t \geq t_{\text{start}}$ to align the entire reasoning sequence with the abstract manifold.

3 Experimental Setup

3.1 Models

We evaluate the steering pipeline on six open-weight LLMs from three different model families: Qwen-2.5-7B and Qwen-3-14B; Gemma-2-9B and Gemma-3-12B; and Mistral-7B-Instruct-v0.3 and Ministral-3-14B-Instruct.

3.2 Dataset

We extend the syllogistic reasoning corpus from [Valentino et al. \(2026b\)](#), comprising 24 logical forms instantiated via taxonomic relations from WordNet. Each instance includes two premises, a conclusion, and is annotated with logical **validity** and conclusion **plausibility**.

The original dataset is in English. To assess cross-lingual transfer, we evaluate on nine additional languages spanning diverse families, scripts, and resource levels: French, Spanish, Italian, German, Russian, and Chinese as **High-Resource Languages (HRLs)**, and Bengali, Swahili, and Telugu as **Low-Resource Languages (LRLs)**. The English dataset contains 2,780 examples, each paired with its corresponding abstract, while the other languages’ datasets contain 960 examples each. All datasets are balanced across validity and plausibility categories. Data are generated via automatic translation with back-translation quality control (details in Appendix B).

3.3 Metrics

We introduce metrics that quantify heuristic reliance, penalise it in a holistic score, and measure steering effectiveness.

Belief Bias (Δ_{belief}) measures how strongly the plausibility of the conclusion influences accuracy. Let $p(x) \in \{\text{plausible}, \text{implausible}\}$ is the plausibility of the conclusion of example x , while $y(x)$ is its formal validity (§2.1). We divide the dataset examples into two groups:

- **Belief-Consistent:** inputs that are either Valid–Plausible or Invalid–Implausible.
- **Belief-Conflict:** inputs that are either Valid–Implausible or Invalid–Plausible.

Let $\mathcal{D}_{\text{consistent}}$ and $\mathcal{D}_{\text{conflict}}$ denote these two subsets, respectively. Belief bias is the performance

gap between them:

$$\text{Acc}_{\text{consistent}} = \text{Acc}_{x \in \mathcal{D}_{\text{consistent}}}, \quad (21)$$

$$\text{Acc}_{\text{conflict}} = \text{Acc}_{x \in \mathcal{D}_{\text{conflict}}}, \quad (22)$$

$$\Delta_{\text{belief}} = |\text{Acc}_{\text{consistent}} - \text{Acc}_{\text{conflict}}|. \quad (23)$$

Bias-Penalised Accuracy (BPA). Global accuracy alone can be misleading, as a model may achieve high performance by exploiting semantic heuristics. We therefore scale global accuracy by robustness to belief bias:

$$\text{BPA} = \text{Acc}_{\text{global}} (1 - \Delta_{\text{belief}}). \quad (24)$$

All accuracy and bias terms in these definitions are expressed on the unit interval $[0, 1]$. Reported percentage values are obtained by multiplying the resulting scores by 100.

Abstract Alignment (η). To validate our hypothesis that steering maps content-laden activations into an abstract manifold, we compare the steered model’s performance on content-laden syllogisms with its performance on abstract syllogisms:

$$\eta = \frac{\text{Acc}_{\text{steered}}}{\text{Acc}_{\text{abstract}}}. \quad (25)$$

Here, $\eta = 1.0$ indicates that the steered model matches its performance on abstract inputs, $\eta < 1.0$ indicates lower performance, and $\eta > 1.0$ indicates better performance on the steered content-laden inputs than on the abstract ones.

3.4 Evaluation

Cross-validation. Abstractors are trained and evaluated using three-fold stratified cross-validation. In each run, two folds are used for training and the remaining fold is used for evaluation. The folds are stratified by syllogism validity. Unless otherwise stated, steering results are averaged across the three held-out folds. The SFT and CoT comparison uses only one shared held-out fold, as detailed in Appendix F.

Cross-lingual transfer. Our primary hypothesis is that Abstractors learn a generalisable representation of logic across languages. To test this hypothesis, we train Abstractors exclusively on the English training folds and evaluate them zero-shot on English, French, Spanish, Italian, German, Russian, Chinese, Bengali, Swahili, and Telugu.

Steering strength ablation. We conduct an ablation on the steering strength hyperparameter α , testing $\alpha \in \{0.1, 0.2, \dots, 0.9, 1.0\}$. Detailed results for each α value are shown in Appendix H.

Baselines. To evaluate the effectiveness of the inference-time intervention, we compare our approach with three baselines:

- **Baseline (No Steering):** the original, unsteered model’s zero-shot performance.
- **Supervised Fine-Tuning (SFT):** a parameter-efficient fine-tuning baseline. To ensure a fair comparison, we employ PiSSA (PrIncipal Singular values and Singular vectors Adaptation) (Meng et al., 2025), a PEFT method that outperforms standard LoRA by initialising adapters via SVD (details in Appendix A.3).
- **Chain-of-Thought (CoT):** used to evaluate whether the model’s logical reasoning can be elicited through explicit step-by-step prompting (details in Appendix A.4).

4 Results & Analysis

4.1 Main Results: Steering Enhances Accuracy and Reduces Bias

We first evaluate the Abstractors on the English test set. Table 2 reports Bias-Penalised Accuracy (BPA), while Figure 2 provides global accuracy decomposition and BPA before and after steering.

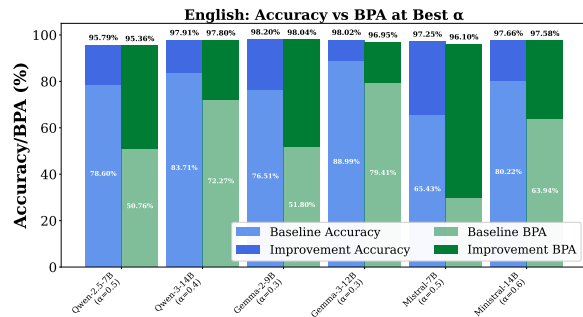


Figure 2: Performance on the English test set. Global accuracy (blue) and Bias-Penalised Accuracy (green); darker shades indicate gains from activation steering.

The results show improvements across all model families. Steering content-laden activations towards the abstract reasoning space results in BPA gains ranging from +17.54 pp (Gemma-3-12B) to +65.94 pp (Mistral-7B). Notably, Mistral-7B rises from 30.16 to 96.10 BPA.

BPA gains are consistently larger than gains in raw global accuracy. For example, Qwen-2.5-7B

improves by +17.19 pp in accuracy and by +44.60 pp in BPA. Since BPA penalises performance gaps between belief-consistent and belief-conflict examples, these results suggest that the Abstractors reduce reliance on semantic plausibility and make predictions more grounded in formal validity at the same time.

4.2 Cross-Lingual Generalisation

A central hypothesis of our work is that English-trained Abstractors capture a transformation that generalises across languages. We group evaluation languages into HRLs and LRLs (§3.2) to distinguish the steering effect from differences in the quality of the underlying representations.

HRLs: near-perfect transfer. Table 2 and Figure 3 reveal robust transfer across all model families. Steering achieves percentage increases in BPA that often exceed those observed in English.

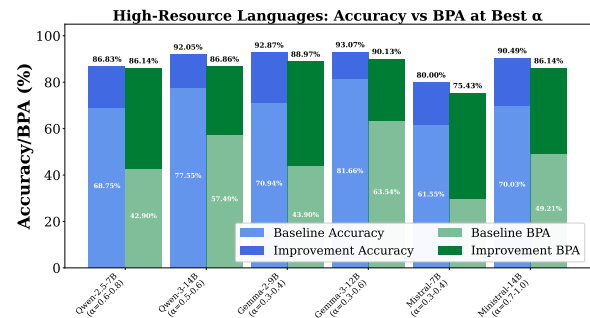


Figure 3: Performance comparison on HRLs (French, Spanish, Italian, German, Russian, Chinese).

The consistency of these results is significant. For instance, Gemma-3-12B achieves an average BPA of 90.13 across HRLs, only slightly below its English performance of 96.95. This transfer extends even to languages with fundamentally different scripts and tokenisation schemes such as Chinese, which uses logographic characters and requires fundamentally different subword segmentation. This suggests that the Abstractors target a shared latent geometry that encodes logical structure, generalising across typologically diverse languages.

LRLs: limited gains. Steering consistently improves performance, but Table 2 and Figure 4 show that the absolute gains are noticeably smaller. BPA improvements in LRLs range from +5.39 pp (Mistral-7B) to +26.55 pp (Gemma-2-9B). Whilst substantial in relative terms, corresponding to increases up to +56.38%, they fall short of the near-perfect alignment observed in HRLs. This suggests

Model	English					HRLs					LRLs				
	Base Model		SFT		CoT	Base Model		SFT		CoT	Base Model		SFT		CoT
	BPA	→Steer	BPA	Δ_{BPA}	BPA	BPA	→Steer	BPA	Δ_{BPA}	BPA	BPA	→Steer	BPA	Δ_{BPA}	BPA
Qwen-2.5-7B	50.8	95.4	98.6	+3.2	81.7	42.9	86.1	84.6	-1.5	57.7	55.2	64.8	70.1	+5.3	55.0
Qwen-3-14B	72.3	97.8	—	—	—	57.5	86.9	—	—	—	54.0	70.8	—	—	—
Gemma-2-9B	51.8	98.0	98.8	+0.8	69.9	43.9	89.0	93.4	+4.4	58.5	47.0	73.5	77.0	+3.5	53.6
Gemma-3-12B	79.4	97.0	—	—	—	63.5	90.1	—	—	—	58.0	79.0	—	—	—
Mistral-7B	30.2	96.1	88.1	-8.0	34.3	30.0	75.4	78.2	+2.8	34.2	47.6	53.0	53.7	+0.7	50.0
Ministral-3-14B	63.9	97.6	—	—	—	49.2	86.1	—	—	—	50.4	69.2	—	—	—
Average	58.1	97.0	95.2	-1.8	62.0	47.8	85.6	85.4	-0.2	50.1	52.0	68.4	66.9	-1.5	52.9

Table 2: Comparison of Bias-Penalised Accuracy (BPA) across all methods and language groups. **Steering** (bold) shows average performance across 3 folds for all models. Δ_{BPA} shows the pp difference in BPA relative to steering (green = improvement, red = degradation).

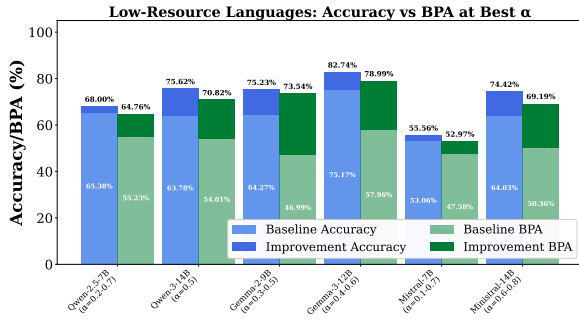


Figure 4: Performance comparison on LRLs (Bengali, Swahili, Telugu).

a *geometry gap*: LRLs receive less pre-training exposure, leading to noisier activations. Since steering amplifies existing latent features, it cannot induce robust reasoning where the underlying representation quality is poor. Yet, improving LRLs without language-specific training remains non-trivial. The Abstractors still amplify logical reasoning in these languages, suggesting a shared manifold even in under-resourced settings.

4.3 Abstract Alignment Analysis

Beyond the primary performance metrics, we evaluate our intervention using Abstract Alignment (η), which compares accuracy on steered content-laden inputs with accuracy on abstract inputs. Detailed results are reported in Appendix D (Tables 8 and 9).

English Alignment. Models exhibit strong alignment in English, with η values consistently exceeding 1.0. These results are consistent with the hypothesis that Abstractors reduce content-related interference while improving reasoning performance.

High-Resource Languages (HRLs). Despite Abstractors being trained in English, η scores remain clustered around 1.0 across all HRLs, demonstrating that they share a broadly isomorphic reasoning geometry with English in the model’s latent space.

Content-Laden Input

Validity: False | Plausibility: True

Premise 1: All cows are mammals.

Premise 2: Some mammals are not birds.

Conclusion: No birds are cows.

Model Response

"The given syllogism is a classic example of a valid deductive argument. [...] Since cows are mammals and there are mammals that are not birds, it is impossible for a bird to be a cow because a bird is not a mammal. Therefore, the given syllogism is valid."

Table 3: An example of a plausibility-driven reasoning error. The model accepts a plausible conclusion even though it does not logically follow from the premises.

Low-Resource Languages (LRLs). In LRLs, alignment remains high but shows slight degradation. The variance in η scores suggests that the target abstract manifold is intrinsically noisier due to limited pre-training exposure, consistent with our earlier findings on the *geometry gap* (§4.2).

4.4 Comparative Analysis

We compare the Abstractors with CoT reasoning and SFT; detailed results in Appendix F.

CoT. Contrary to trends in general reasoning tasks, CoT yields the lowest performance among methods and significantly underperforms steering across all scenarios, with relative BPA differences ranging from -11.55% to -63.86% . Qualitative analysis reveals that the generated reasoning chains frequently hallucinate logical validity to align with the semantic plausibility of the conclusion. An illustrative example is shown in Table 3. This failure pattern, where the model generates seemingly logical reasoning that ultimately justifies a semantically plausible but logically unsupported conclusion (or vice versa), appears consistently across

belief-conflict cases, suggesting that CoT prompting alone does not solve the underlying problem.

SFT. SFT via PiSSA is the strongest of the standard interventions we compare against. It performs on par with or better than steering on Qwen and Gemma, with relative BPA differences with respect to steering ranging only from -7.05% to $+7.38\%$. Yet, this comes at the cost of modularity: SFT alters model weights, risking catastrophic forgetting or overfitting to the syllogistic format, while steering can be dynamically toggled during inference. Notably, for Mistral-7B, steering outperforms SFT, suggesting that, for some architectures, reasoning features are better accessed via inference-time geometry than weight modification.

4.5 Fluency Sanity Check

A risk of training Abstractors exclusively in English is that they may implicitly bias the model towards English-centric representations, which could manifest as an increased perplexity (PPL) score on text in other languages. We measure the relative increase in perplexity across all languages on mC4 (Xue et al., 2021) as a function of steering strength α . At the optimal α range identified in §4 ($\alpha \in [0.4, 0.6]$), the PPL increase remains moderate ($\leq 5\%$). More significant degradation ($> 10\%$) is observed only at aggressive steering levels ($\alpha \geq 0.8$), where the abstract representations begin to override the syntactic features necessary for coherent generation. PPL increases uniformly across all evaluated languages, suggesting no English-specific bias and confirming that the intervention is safe: it corrects reasoning without altering the underlying generative capabilities of the models involved (Appendix I).

4.6 OOD Analysis

To assess the limits of our abstraction mechanism, we evaluate across two distinct OOD scenarios: **Near-OOD**, assessing performance under distributional shifts (i.e., changing syllogism phrasing and structure), and **Far-OOD**, evaluating the impact on completely different tasks.

Near-OOD: NeuBAROCO. We evaluate our English-trained Abstractors on the NeuBAROCO NLI dataset (Ozeki et al., 2024) for both English and Japanese across Qwen-2.5-7B, Gemma-2-9B and Mistral-7B. As shown in Table 4, global accuracy increases in both languages. Notably, our

method yields the largest improvements in the *Incongruent* category, which tests content-driven biases. The fact that the Abstractors successfully steer a model even when presented with a Japanese prompt, a language typologically distant from English, strengthens the hypothesis that our framework captures a highly generalisable and language-agnostic logical manifold.

Model	Lang	Global	Incongr.	Other Biases*
Qwen-2.5-7B	EN	88.0 $+2.3$	84.4 $+8.2$	82.8 -8.5
	JA	83.3 $+4.1$	82.0 $+13.1$	81.4 $+3.7$
Gemma-2-9B	EN	84.3 $+3.0$	86.1 $+15.6$	79.1 $+4.6$
	JA	82.3 $+1.8$	81.2 $+1.7$	79.9 $+1.7$
Mistral-7B	EN	83.3 $+8.5$	80.3 $+15.6$	73.1 $+8.9$
	JA	68.8 $+5.2$	69.7 $+13.5$	55.0 $+8.4$

Table 4: NeuBAROCO results. Steered accuracy shown; absolute baseline improvement in parentheses. *Other Biases: mean performance across Congruent, Conversion, and Atmosphere categories.

Far-OOD: MMMLU. We extend our evaluation to the Multilingual Massive Multitask Language Understanding (MMMLU) dataset (Hendrycks et al., 2021) to assess how abstraction impacts model capabilities. We evaluated two task categories: **Factual**, which includes subjects such as *geography* and *world history*; and **Formal Logic**, which includes *abstract algebra* and *formal logic*. Since our method is designed to suppress semantic content in favour of logical form, we hypothesise that this could produce a "feature suppression" effect specifically on factual, knowledge-based tasks. Figure 5 (details Appendix J) shows the average deltas across models and languages as a function of α . The Formal Logic subset shows lower degradation, given the mechanistic nature of the intervention: the steering pipeline’s specificity targets reasoning-related representations while inhibiting factual memory, keeping in mind that the nature of our approach allows steering to be dynamically disabled, set per task, or tuned to balance reasoning enhancement with general knowledge preservation.

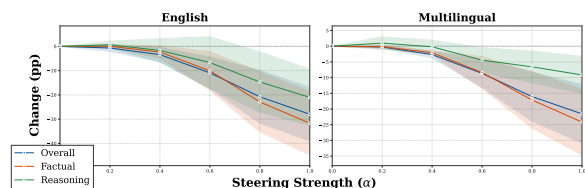


Figure 5: Performance impact on MMMLU. Formal logic subtasks (green) degrade at a lower rate.

5 Related Work

Syllogistic reasoning. Syllogisms provide a controlled setting for probing whether LLMs can track logical entailment independently of lexical and world-knowledge cues (Evans et al., 1983; Newstead, 2003). Although this is a formally clear task, LLMs often fail to disentangle content from form, treating semantic plausibility as evidence of validity and exhibiting biases grounded in semantic content, rejecting valid conclusions when these disagree with background knowledge. Ozeki et al. (2024); Eisape et al. (2024) suggest that scaffolding paradigms yield only marginal improvements, whereas SFT reduces reliance on such heuristics but does not eliminate them.

Activation Steering. A growing body of work investigates inference-time control by intervening on a model’s internal activations. These methods identify subspaces associated with specific behaviours and modulate them to influence downstream predictions (Stoehr et al., 2024; Soo et al., 2025). Activation steering has been applied to reduce hallucinations, mitigate harmful outputs, and improve factual reliability, offering a lightweight and reversible alternative to parameter updates. Activation steering has already been applied to mitigate content effects in syllogistic reasoning as a class-conditional method that pushes representations towards valid or invalid labels (Valentino et al., 2026a). This label-directed intervention risks enforcing statistical shortcuts rather than robust logical reasoning. Concurrently, emerging work has shown that steering mechanisms can effectively generalise across languages (Maraia et al., 2026). The intersection of these findings—the limitations of label-directed interventions and the possibility of cross-lingual transfer—motivated the content-invariant approach proposed in this work.

Abstraction for content-invariant reasoning. A complementary line of work seeks to improve reasoning robustness through higher levels of abstraction. CoT explanations can inherit semantic shortcuts that compromise faithfulness (Lyu et al., 2023), motivating intermediate representations that separate variables and relations from surface content (Ranaldi and Freitas, 2024a; Ranaldi et al., 2025). Recent work has also considered abstraction as a mechanism for compressing experience into reusable pseudo-symbolic representations supporting adaptive reasoning and planning (Ranaldi, 2026), consistent with proposals

for neuro-symbolic reasoning by design, where explicit architectural mechanisms make symbolic structure accessible to neural computation (Zanzotto et al., 2025). Existing approaches primarily construct abstractions via explicit reasoning procedures, architectural mechanisms, or supervision; we instead ask whether abstraction can be induced within the model’s latent geometry and exploited via inference-time intervention.

Reasoning control and adaptation. Recent work has also explored mechanisms to control reasoning beyond direct prompting. One direction uses argumentation and model-based oversight to evaluate reasoning and derive supervision from disagreements between models (Ranaldi and Ranaldi, 2026; Pucci and Ranaldi, 2025). A complementary direction transfers reasoning capabilities through instruction tuning, later refined via preference optimisation and model-generated trajectories (Ranaldi and Freitas, 2024b). These approaches demonstrate the value of additional control over reasoning, though they operate through external oversight or parameter adaptation. Our approach instead investigates controlling reasoning directly through the geometry of internal representations, without requiring model-weight updates.

Our contribution. We unify abstraction and activation steering by treating abstraction as a geometric target. Using paired syllogisms, we define an abstract reasoning space and train lightweight *Abstractors* to map content-laden activations into it. These targets are integrated via weight-free, multi-layer interventions during inference. Finally, cross-lingual evaluation confirms that this steering captures structural regularities beyond language-specific surface cues.

6 Conclusion

We showed that abstraction-guided activation steering can reduce content effect in syllogistic reasoning by separating structural inference from semantic content at the representation level. By dynamically mapping internal states towards an abstract reasoning manifold, we observe substantial improvements in bias-penalised accuracy, alongside zero-shot transfer across typologically diverse languages, without requiring additional language-specific supervision or fine-tuning. These findings support a geometric view of abstraction: a target definable within an LLM’s latent space and controllable directly through activation steering.

Limitations

Task scope. We focused on syllogistic reasoning as a controlled test bed for analysing content effects. While this setting allows for fine-grained control over logical form and semantic plausibility, it limits the generalisability of the findings to more complex reasoning scenarios, such as propositional or first-order logic, multi-hop inference, or open-ended reasoning tasks.

Constructing abstract counterparts. The proposed framework relies on paired content-laden and abstract examples to define a target reasoning manifold. This design assumes that suitable abstract counterparts can be constructed reliably, an assumption that may not hold in domains where abstraction is less well-defined or inherently task-specific.

Model access. Our evaluation is restricted to open-weight models, as the method requires direct access to internal activations. Thus, the approach is not directly applicable to closed-source models, in which activation-level interventions are infeasible. While this does not affect the conceptual validity of abstraction-guided steering, it constrains its practical applicability in settings dominated by proprietary systems.

References

- Peter Clark, Øyvind Tafjord, and Kyle Richardson. 2020. [Transformers as soft reasoners over language](#). *Preprint*, arXiv:2002.05867.
- Ishita Dasgupta, Andrew K. Lampinen, Stephanie C. Y. Chan, Hannah R. Sheahan, Antonia Creswell, Dharshan Kumaran, James L. McClelland, and Felix Hill. 2024. [Language models show human-like content effects on reasoning tasks](#). *Preprint*, arXiv:2207.07051.
- Boyi Deng, Yu Wan, Baosong Yang, Yidan Zhang, and Fuli Feng. 2025. [Unveiling language-specific features in large language models via sparse autoencoders](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4563–4608, Vienna, Austria. Association for Computational Linguistics.
- Tiwalayo Eisape, Michael Tessler, Ishita Dasgupta, Fei Sha, Sjoerd Steenkiste, and Tal Linzen. 2024. [A systematic comparison of syllogistic reasoning in humans and language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8425–8444, Mexico City, Mexico. Association for Computational Linguistics.
- J. St. B. T. Evans, Julie L. Barston, and Paul Pollard. 1983. [On the conflict between logic and belief in syllogistic reasoning](#). *Memory Cognition*, 11(3):295–306.
- Changjiang Gao, Xu Huang, Wenhao Zhu, Shujian Huang, Lei Li, and Fei Yuan. 2025. [Could thinking multilingually empower llm reasoning?](#) *Preprint*, arXiv:2504.11833.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Gemma Team and 1 others. 2024. [Gemma 2: Improving open language models at a practical size](#). *Preprint*, arXiv:2408.00118.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. [Transformer feed-forward layers are key-value memories](#). *Preprint*, arXiv:2012.14913.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Preprint*, arXiv:2009.03300.
- Ruixin Hong, Hongming Zhang, Xiaoman Pan, Dong Yu, and Changshui Zhang. 2024. [Abstraction-of-thought makes language models better reasoners](#). In *Conference on Empirical Methods in Natural Language Processing*.
- Wenyue Hua, Kaijie Zhu, Lingyao Li, Lizhou Fan, Shuhang Lin, Mingyu Jin, Haochen Xue, Zelong Li, Jindong Wang, and Yongfeng Zhang. 2024. [Disentangling logic: The role of context in large language model reasoning capabilities](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Geonhee Kim, Marco Valentino, and Andre Freitas. 2025. [Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10074–10095, Vienna, Austria. Association for Computational Linguistics.
- Francesca Lucchetti and Arjun Guha. 2025. [Understanding how codellms \(mis\)predict types with activation steering](#). *Preprint*, arXiv:2404.01903.
- Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. [Faithful chain-of-thought reasoning](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific*

- Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 305–329, Nusa Dua, Bali. Association for Computational Linguistics.
- 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.
- Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2025. [Pissa: Principal singular values and singular vectors adaptation of large language models](#). *Preprint*, arXiv:2404.02948.
- Mistral AI. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Mistral AI. 2025. [Minstral-3 14b instruct](#). <https://huggingface.co/mistralai/Minstral-3-14B-Instruct-2512>. Large Language Model, Instruct-tuned.
- Jacqueline L. Mitchell, Brian Hyeonseok Kim, Chenyu Zhou, and Chao Wang. 2025. [Understanding formal reasoning failures in llms as abstract interpreters](#). *Proceedings of the 1st ACM SIGPLAN International Workshop on Language Models and Programming Languages*.
- Stephen E Newstead. 2003. [Can natural language semantics explain syllogistic reasoning?](#) *Cognition*, 90(2):193–199.
- Kentaro Ozeki, Risako Ando, Takanobu Morishita, Hirohiko Abe, Koji Mineshima, and Mitsuhiro Okada. 2024. [Exploring reasoning biases in large language models through syllogism: Insights from the NeuBAROCO dataset](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16063–16077, Bangkok, Thailand. Association for Computational Linguistics.
- David Pomerence, Jonas Nothnagel, and Simon Ostermann. 2025. [The ai language proficiency monitor – tracking the progress of llms on multilingual benchmarks](#). *Preprint*, arXiv:2507.08538.
- Giulia Pucci and Leonardo Ranaldi. 2025. [Advancing oversight reasoning across languages for audit sycophantic behaviour via X-agent](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12949–12965, Suzhou, China. Association for Computational Linguistics.
- Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. [Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.
- Qwen, :, An Yang, An Feng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, and 43 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Federico Ranaldi, Leonardo Ranaldi, Fabio Massimo Zanzotto, and Shay B Cohen. 2026. [Thinking in schemas: Robust syllogistic reasoning in LLMs](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40891–40909, San Diego, California, United States. Association for Computational Linguistics.
- Leonardo Ranaldi. 2026. [Evolving agents](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16561–16569, San Diego, California, United States. Association for Computational Linguistics.
- Leonardo Ranaldi and Andre Freitas. 2024a. [Aligning large and small language models via chain-of-thought reasoning](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1812–1827, St. Julian’s, Malta. Association for Computational Linguistics.
- Leonardo Ranaldi and Andre Freitas. 2024b. [Self-refine instruction-tuning for aligning reasoning in language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2347, Miami, Florida, USA. Association for Computational Linguistics.
- Leonardo Ranaldi and Giulia Pucci. 2025. [Multilingual reasoning via self-training](#). 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 11566–11582, Albuquerque, New Mexico. Association for Computational Linguistics.
- Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024. [A tree-of-thoughts to broaden multi-step reasoning across languages](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241, Mexico City, Mexico. Association for Computational Linguistics.
- Leonardo Ranaldi and Federico Ranaldi. 2026. [Agentic oversight via dialectic reasoning](#). In *Proceedings of the 64th Annual Meeting of the Association for*

- Computational Linguistics (Volume 1: Long Papers)*, pages 24924–24944, San Diego, California, United States. Association for Computational Linguistics.
- Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025. [Improving chain-of-thought reasoning via quasi-symbolic abstractions](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17222–17240, Vienna, Austria. Association for Computational Linguistics.
- Nina Rimskey, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- S Seals and Valerie Shalin. 2024. [Evaluating the deductive competence of large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8614–8630, Mexico City, Mexico. Association for Computational Linguistics.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language models are multilingual chain-of-thought reasoners](#). *Preprint*, arXiv:2210.03057.
- Samuel Soo, Chen Guang, Wesley Teng, Chandrasekaran Balaganesh, Tan Guoxian, and Yan Ming. 2025. [Interpretable steering of large language models with feature guided activation additions](#). *Preprint*, arXiv:2501.09929.
- Niklas Stoeck, Kevin Du, Vésteinn Snæbjarnarson, Robert West, Ryan Cotterell, and Aaron Schein. 2024. [Activation scaling for steering and interpreting language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8189–8200, Miami, Florida, USA. Association for Computational Linguistics.
- Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. 2024. [Steering language models with activation engineering](#). *Preprint*, arXiv:2308.10248.
- Marco Valentino, Geonhee Kim, Dhairya Dalal, Zhixue Zhao, and André Freitas. 2026a. [Mitigating content effects on reasoning in language models through fine-grained activation steering](#). *Preprint*, arXiv:2505.12189.
- Marco Valentino, Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, and André Freitas. 2026b. [SemEval-2026 task 11: Disentangling content and formal reasoning in large language models](#). In *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pages 3716–3730, San Diego, California, USA. Association for Computational Linguistics.
- Wentao Wan, Zhuojie Yang, Yongcan Chen, Chenglin Luo, Ruilin Wang, Kehao Cai, Nan Kang, Liang Lin, and Keze Wang. 2025. [Sr-fot: A syllogistic-reasoning framework of thought for large language models tackling knowledge-based reasoning tasks](#). *Preprint*, arXiv:2501.11599.
- Yiming Wang, Zhuosheng Zhang, and Rui Wang. 2023. [Meta-reasoning: Semantics-symbol deconstruction for large language models](#). *ArXiv*, abs/2306.17820.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models](#). *Preprint*, arXiv:2201.11903.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mT5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.
- Fabio Massimo Zanzotto, Elena Sofia Ruzzetti, Giancarlo A. Kompero, Leonardo Ranaldi, Davide Venditti, Federico Ranaldi, Cristina Giannone, Andrea Favalli, and Raniero Romagnoli. 2025. [Position paper: MeMo: Towards language models with associative memory mechanisms](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 15169–15180, Vienna, Austria. Association for Computational Linguistics.

A Implementation Details

Software and hardware. All experiments were conducted using the Hugging Face Transformers library. Both training and inference were performed using half-precision (FP16) arithmetic on an NVIDIA RTX A6000 GPU.

Models. In our experimental setting, as introduced in §3.1, we evaluate several models (detailed in Table 5). We choose the generation temperature for (mostly) deterministic outputs. The other parameters are left unchanged as recommended by the official resources.

Model	Version
Qwen-2.5-7B	Qwen/Qwen2.5-7B-Instruct
Qwen-3-14B	Qwen/Qwen3-14B
Gemma-2-9B	google/gemma-2-9b-it
Gemma-3-12B	google/gemma-3-12b-it
Mistral-7B	mistralai/Mistral-7B-Instruct-v0.3
Ministral-3-14B	mistralai/Ministral-3-14B-Instruct-2512

Table 5: Model versions used in the experiments, which can be found on huggingface.co. Unless otherwise stated, we used all the default configurations provided in the repositories for each model.

A.1 Steering Layer Selection

Optimal steering layers are identified empirically by analysing the geometric separation between positive and negative target abstract representations. For each content-laden syllogism, we identify its positive target (an abstract syllogism with the same validity) and negative counterfactual (an abstract syllogism with opposite validity) using the matching procedure described in §2.3.

For each layer ℓ , we compute the average cosine similarity between these positive and negative target pairs:

$$s_\ell = \frac{1}{N} \sum_{i=1}^N \cos(a_\ell(x_{i,\ell}^+), a_\ell(x_{i,\ell}^-)). \quad (26)$$

Here, N is the number of content-laden examples, $x_{i,\ell}^+$ is the positive abstract target selected for example i at layer ℓ , and $x_{i,\ell}^-$ is the corresponding negative counterfactual.

Layers with the lowest similarity s_ℓ exhibit the clearest separation between validity classes in the abstract reasoning space, making them optimal targets for steering. These layers are typically found in the middle layers (second or third quarter) of the network, where abstract concepts have formed but have not yet collapsed into task-specific outputs.

Model	Selected Layers
Qwen-2.5-7B	[18, 19, 20, 21, 22]
Qwen-3-14B	[24, 25, 26, 27, 28, 29]
Gemma-2-9B	[21, 22, 23, 24, 25]
Gemma-3-12B	[24, 25, 26, 27, 28, 29]
Mistral-7B	[13, 14, 15, 16, 17]
Ministral-3-14B	[20, 21, 22, 23, 24, 25]

Table 6: Optimal steering layers selected via similarity analysis (0-indexed).

Appendix K, Figure 15 visualises this layer-wise analysis across all models.

A.2 Abstractor Architecture Details

Each Abstractor f_ℓ is implemented as a two-headed Multi-Layer Perceptron as follows:

Shared backbone. A three-layer feed-forward network processes the input activation:

- Layer 1: Linear(d , 1024) + LayerNorm + LeakyReLU(0.01) + Dropout(0.1)
- Layer 2: Linear(1024, 1024) + LayerNorm + LeakyReLU(0.01) + Dropout(0.1)
- Layer 3: Linear(1024, 1024) + LayerNorm + LeakyReLU(0.01)

Direction head. Predicts the normalised direction vector:

- Linear(1024, 512) + ReLU
- Linear(512, d) + ℓ_2 Normalisation

Magnitude head. Predicts the scalar magnitude:

- Linear(1024, 512) + ReLU
- Linear(512, 1) + Softplus

The final prediction combines both components:

$$\hat{a}_\ell(x) = \hat{m}_\ell(x) \hat{d}_\ell(x). \quad (27)$$

Training configuration. Abstractors are trained using the AdamW optimiser with a learning rate of 5×10^{-4} and a weight decay of 10^{-3} . We use batch size 128 with gradient clipping (max norm 1.0) and ReduceLROnPlateau scheduling (patience=10, factor=0.5). Training runs for up to 150 epochs with early stopping (patience=20). The contrastive margin threshold is set to 0.2, with loss weights $\lambda_{\text{repel}} = 0.75$ and $\lambda_{\text{mag}} = 1.0$.

A.3 SFT Implementation Details

We employ PiSSA (Principal Singular values and Singular vectors Adaptation) for parameter-efficient fine-tuning, implemented as follows:

LoRA parameters.

- Rank: $r = 16$
- Alpha: $\text{lora}_\alpha = 16$
- Dropout: 0.05
- Target modules: query, key, value, and output projection layers
- Initialisation: PiSSA (SVD-based)

Training data. We train on the English training set, using both content-laden and abstract examples. For n content-abstract pairs in a fold, we create $2n$ training examples by including each variant separately. These examples are shuffled together during training, ensuring the model learns to handle both content-rich and symbolic inputs. This doubles the effective training set size compared to approaches that use only content-laden examples.

Training hyperparameters. Learning rate and epochs vary by model to account for differences in architecture and baseline performance:

- Qwen-2.5-7B: $\text{lr}=10^{-5}$, 2 epochs
- Gemma-2-9B: $\text{lr}=10^{-5}$, 1 epoch
- Mistral-7B: $\text{lr}=10^{-5}$, 5 epochs

All models use batch size 4 with gradient accumulation over 2 steps (effective batch size 8). We apply gradient clipping (max norm 1.0) and early stopping based on validation loss (patience=10 epochs). The training objective is the standard cross-entropy loss on ground-truth labels.

Label masking. To train the model to predict only the answer token, we mask all prompt tokens in the loss calculation. The system identifies where the answer begins by searching for "Valid" or "Invalid" tokens in the tokenised sequence, then sets all preceding positions to -100 (ignored index). If the answer position cannot be located, we conservatively mask the first 90% of the sequence.

A.4 CoT Implementation Details

CoT prompting encourages models to reason step-by-step before providing a final answer. We use the following prompt:

"Evaluate the logical validity of the following syllogism by reasoning step-by-step. First, analyse

each premise carefully. Then, determine whether the conclusion logically follows from the premises. Finally, state your answer as either 'valid' or 'invalid'. Ignore the meaning, realism, or plausibility of the statements. Let's think step by step (be concise and conclude within a few sentences):"

Generation parameters. We use greedy decoding (temperature=0) with $\text{max_new_tokens}=512$. The system prompt varies by model family but generally presents the assistant as an expert and skilled assistant in logical reasoning.

Answer extraction. Since CoT responses contain both reasoning and the final answer, we extract the validity prediction using pattern matching. The system searches for explicit markers like "final answer", "therefore", or "the syllogism is" followed by "valid" or "invalid". If no clear marker exists, we examine the last 200 characters for standalone validity keywords.

B Dataset and Translation Details

B.1 Language Selection Rationale

Our evaluation suite comprises ten languages selected to maximise typological and representational diversity across three key axes:

Family and syntax. The languages span four major families: Indo-European (English, French, Spanish, Italian, German, Russian, Bengali), Sino-Tibetan (Chinese), Dravidian (Telugu), and Niger-Congo (Swahili). This selection includes varying syntactic orders, like SVO (English, Chinese, Swahili) and SOV (Bengali, Telugu), forcing the model to locate logical terms in different positions. This tests whether the Abstractors target word order patterns or deeper structural abstractions.

Script and tokenisation. We include languages that challenge the tokenizer differently: shared Latin subwords (Romance/Germanic), distinct scripts (Cyrillic for Russian, Bengali script, Telugu script), and logographic systems (Chinese). If steering relied on token-level patterns, we would expect dramatic performance drops when script systems change. The robust transfer to Chinese and Russian (§4.2) demonstrates that the intervention operates above the token level.

Resource stratification. High-resource languages (French, Spanish, Italian, German, Russian, Chinese) receive substantial exposure during pre-training, resulting in well-structured latent representations. Low-resource languages (Swahili,

Telugu, Bengali) receive limited exposure, yielding noisier representations. This stratification allows us to disentangle steering mechanism quality from base model representation quality: if steering fails on LRLs, is it because the mechanism itself is weak, or because the underlying representations are insufficient? Our results (§4.2) support the latter interpretation.

B.2 Translation and Quality Check

We perform a three-stage quality control to preserve both *form* (logical structure, quantifiers, polarity) and *content* (lexical, named categories):

Structural and lexical guards. We execute template-level invariants on determiners/quantifiers (*all/no/some/some...not*), refuse contradiction, and check preservation for key nouns via bilingual lexicons. We require high semantic similarity between source and back-translation (SBERT cosine $\gtrsim 0.90$) and surface consistency (chrF $\gtrsim 0.70$).

Bilingual review. Native/near-native reviewers verify (i) faithful rendering of quantifiers, (ii) absence of idiomatic changes that could alter plausibility, and (iii) strict isomorphism of logical form. Disagreements trigger a corrective re-prompt with targeted constraints.

Invariance check. We validate that *validity* and *plausibility* labels remain unchanged after translation/back-translation.

Language	SBERT Cosine	chrF	Pass Rate (%)
English (en)	0.94	0.81	100.0
German (de)	0.92	0.76	97.8
Spanish (es)	0.93	0.77	98.2
French (fr)	0.91	0.74	96.5
Italian (it)	0.93	0.79	97.1
Russian (ru)	0.90	0.72	95.4
Chinese (zh)	0.89	0.71	94.8
Swahili (sw)	0.91	0.70	93.2
Bengali (bn)	0.90	0.73	94.6
Telugu (te)	0.88	0.69	92.5
Average	0.91	0.74	95.8

Table 7: Translation quality metrics and pass rates per language.

C Design Rationale and Class-Specific Transformations

C.1 Unified vs. Conditional Architecture

Preliminary analysis revealed that valid and invalid input syllogisms produce distinct activation

patterns, suggesting that optimal transformations should be class-specific. While this observation motivates a conditional approach, such as using a linear probe to gate separate interventions for each validity class (Maraia et al., 2026), we opt for a unified single-model architecture in the main experiments for three reasons:

Error propagation. Conditional systems introduce cascading failure modes: classifier errors compound with transformation errors, creating a failure ceiling determined by the weaker component.

Feasibility. Reliable validity probing, with performance matching that of the class-specific Abstractors, presupposes that the base model already encodes near-perfect internal representations distinguishing valid from invalid reasoning. However, if this were the case, it would be an important finding by itself. Our experiments confirm that validity probes achieve $\approx 95\%$ accuracy on English but degrade substantially in multilingual settings, limiting the practical effectiveness of conditional approaches.

Efficiency. Auxiliary classifiers add computational overhead and increase system complexity. Rather than relying on explicit gating, our unified Abstractor uses contrastive loss to *implicitly* route activations based on their input geometry, automatically mapping valid and invalid inputs to their respective abstract manifolds.

C.2 Oracle Validation: Class-Specific Upper Bound

To validate that the limitations of our approach stem from routing, we implemented class-specific Abstractors. When applied in an "oracle" setting (where we use ground truth labels to select the corresponding valid or invalid Abstractor) we achieved $\sim 100\%$ **accuracy across HRLs for all models**. This confirms that the observed failure modes arise mostly from the difficulty of distinguishing valid from invalid logic paths.

To quantify the theoretical upper bound without an oracle, we trained validity probes on the last-token activations of each steering layer using Support Vector Machines (SVMs) with an RBF kernel, applied to the standardised activations with a stratified 80/20 train-test split.

Since the class-specific Abstractors achieve perfect accuracy when given the ground truth, the bottleneck lies solely in the probe. Therefore, the

results shown in Figure 6 represent the *hard upper bound* for a possible class-specific implementation. As observed, accuracy on non-English languages, in particular on LRLs, is only acceptable when the probe is trained with few-shot examples in the target languages. The English-trained probe exhibits a mild degradation on HRLs and a severe drop on LRLs, undermining the zero-shot generalisation objective of our design.

This analysis confirms our architectural choice: while class-specific transformations represent a theoretical upper bound, the unified approach is better suited for zero-shot generalisation, offering the best practical trade-off between performance, robustness, and system complexity.

D Detailed Baseline Abstract Results

Model	Categories				Global	
	VP	VI	IP	II	Acc	BPA
Qwen-2.5-7B	71.59	71.77	94.94	94.84	83.27	83.15
Qwen-3-14B	80.06	81.26	99.28	98.28	89.71	88.72
Gemma-2-9B	85.07	85.45	82.33	83.14	83.99	83.81
Gemma-3-12B	98.71	98.70	88.57	89.99	93.99	93.32
Mistral-7B	100.00	100.00	70.91	70.25	85.29	85.01
Ministral-3-14B	82.92	82.85	87.98	86.55	85.07	84.49

Table 8: Performance on the abstract subset. VP/II are belief-consistent, while VI/IP are belief-conflict. These results provide a reference performance that steering aims to approximate or exceed.

Model	EN	HRLs	LRLs
Qwen-2.5-7B	1.15	1.05	0.83
Qwen-3-14B	1.09	1.03	0.85
Gemma-2-9B	1.17	1.11	0.90
Gemma-3-12B	1.04	0.99	0.89
Mistral-7B	1.14	0.96	0.65
Ministral-3-14B	1.15	1.07	0.88

Table 9: Abstract alignment (η) across language groups. Values are the ratio of steered content-laden accuracy to abstract-input accuracy at the value of α yielding the highest BPA.

E Comparison with Prior Activation Steering Work

To contextualise our method against concurrent works in activation steering, we compare our results with [Valentino et al. \(2026a\)](#), which uses static, class-conditional contrastive steering. Both studies conduct evaluations on the same English syllogistic reasoning corpus, and we compare the two overlapping models: Gemma-2-9B and Qwen-2.5-7B. Static, class-conditioned steering achieves BPAs of

81.7 and 89.0 on Gemma and on Qwen respectively. Our Abstractors achieve much higher BPAs: 98.0 on Gemma, and 95.4 on Qwen. Our dynamic Abstractors provide significantly better performance without requiring a class-conditioned intervention, and at near-zero additional computational overhead compared to static steering.

F Steering vs. SFT vs. CoT Results

The method-level comparison was conducted for one cross-validation split. In this split, two folds were used to train the SFT adapters and Fold 0 was used as the held-out evaluation fold. CoT requires no task-specific training and was evaluated on the same held-out fold. For a fair comparison, steering was also evaluated on Fold 0. Thus, all values in Table 10 use the same evaluation examples. The three-fold steering averages reported elsewhere are not used to calculate the differences in this table.

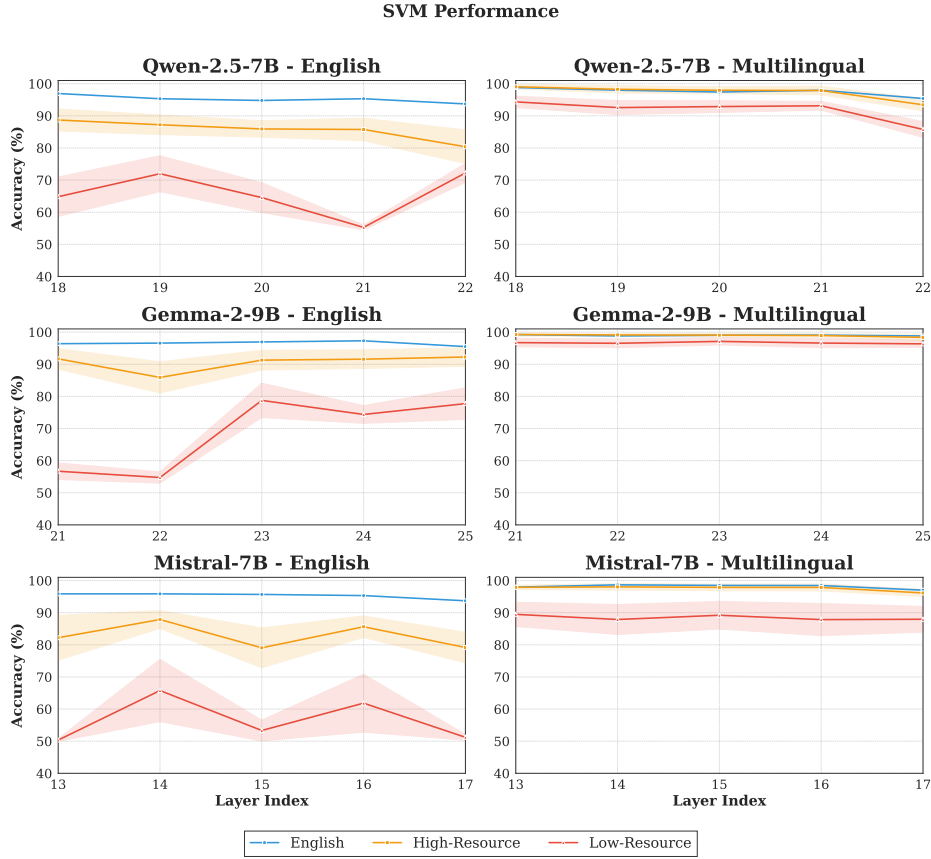


Figure 6: Validity probe (SVM with RBF kernel) accuracy across selected steering layers for Qwen-2.5-7B, Gemma-2-9B and Mistral-7B. The probe trained only on English maintains acceptable accuracy on HRLs but degrades severely on LRLs. Few-shot probes achieve optimal cross-lingual performance but undermine our framework’s zero-shot objective.

Model	Method	English	HRLs	LRLs
Qwen-2.5-7B	Steering	94.99	85.43	65.30
	SFT	98.57 (+3.77%)	84.63 (-0.94%)	70.12 (+7.38%)
	CoT	81.65 (-14.04%)	57.73 (-32.43%)	55.00 (-15.77%)
Gemma-2-9B	Steering	98.25	89.65	76.62
	SFT	98.78 (+0.54%)	93.39 (+4.17%)	76.96 (+0.44%)
	CoT	69.92 (-28.83%)	58.54 (-34.71%)	53.58 (-30.08%)
Mistral-7B	Steering	94.77	77.71	56.49
	SFT	88.09 (-7.05%)	78.17 (+0.59%)	53.68 (-4.98%)
	CoT	34.25 (-63.86%)	34.20 (-55.99%)	49.97 (-11.55%)

Table 10: Comparison of Bias-Penalised Accuracy (BPA) across activation steering, Supervised Fine-Tuning (SFT), and Chain-of-Thought (CoT) prompting. Percentage changes relative to steering are shown in green (improvement) or red (degradation).

G Detailed Language-Specific Results

G.1 Qwen-2.5-7B

Language	Baseline						Steered					
	Categories				Global		Categories				Global	
	VP	VI	IP	II	Acc	BPA	VP	VI	IP	II	Acc	BPA
English (en)	94.5	31.1	90.6	98.0	78.6	50.8	99.4	98.8	93.2	91.7	95.8	95.4
French (fr)	87.9	21.7	80.8	93.5	71.0	43.0	90.7	90.6	91.3	91.6	91.0	90.8
Spanish (es)	79.2	16.2	85.0	97.6	69.6	43.3	80.6	82.8	93.3	95.3	88.0	87.9
Italian (it)	80.0	14.6	82.9	96.3	68.5	41.5	84.9	86.4	88.0	91.9	87.8	86.8
German (de)	67.5	15.8	90.2	95.9	67.4	48.0	82.1	83.9	89.7	92.0	86.9	86.8
Russian (ru)	70.8	10.0	82.5	97.2	65.2	40.6	72.2	71.9	86.3	90.9	80.4	78.4
Chinese (zh)	86.2	15.8	83.3	97.2	70.7	40.9	81.2	78.3	94.6	93.2	86.8	86.2
Bengali (bn)	72.1	27.9	79.9	87.8	67.0	49.5	62.9	59.3	82.2	88.3	73.2	69.7
Swahili (sw)	37.1	23.8	86.8	91.5	59.8	54.4	38.5	29.7	85.8	88.2	60.6	57.2
Telugu (te)	70.0	54.6	73.1	79.7	69.4	61.7	59.9	55.7	80.8	84.4	70.2	67.5

Table 11: Detailed accuracy breakdown in Qwen-2.5-7B at best α (highest BPA) per language. Left: Baseline model performance. Right: Performance with steering. VP/II are belief-consistent; VI/IP are belief-conflict.

G.2 Qwen-3-14B

Language	Baseline						Steered					
	Categories				Global		Categories				Global	
	VP	VI	IP	II	Acc	BPA	VP	VI	IP	II	Acc	BPA
English (en)	83.1	59.7	94.1	98.0	83.7	72.3	99.7	99.1	96.8	96.0	97.9	97.8
French (fr)	89.2	51.7	88.0	97.2	81.6	62.5	96.9	92.2	92.6	92.7	93.6	91.4
Spanish (es)	81.7	42.9	87.6	97.2	77.4	58.7	97.5	90.3	92.9	95.8	94.1	89.4
Italian (it)	84.2	42.9	88.5	97.2	78.2	58.7	96.7	87.9	93.3	95.5	93.4	88.2
German (de)	87.9	46.7	84.2	95.1	78.5	58.0	93.3	84.6	93.7	93.9	91.4	87.3
Russian (ru)	79.6	31.7	84.6	97.6	73.4	51.1	92.4	73.1	89.5	93.6	87.2	76.9
Chinese (zh)	80.8	37.9	87.6	98.0	76.2	55.9	96.0	88.9	91.3	94.3	92.6	88.0
Bengali (bn)	64.6	25.4	91.9	97.2	69.8	54.3	88.6	75.1	87.5	88.9	85.0	78.7
Swahili (sw)	15.4	11.7	94.9	95.1	54.3	53.2	27.9	24.4	89.6	90.9	58.2	56.8
Telugu (te)	56.2	25.0	90.6	97.2	67.3	54.6	86.4	74.7	84.5	88.8	83.6	76.9

Table 12: Detailed accuracy breakdown in Qwen-3-14B at best α (highest BPA) per language. Left: Baseline model performance. Right: Performance with steering. VP/II are belief-consistent; VI/IP are belief-conflict.

G.3 Gemma-2-9B

Language	Baseline						Steered					
	Categories				Global		Categories				Global	
	VP	VI	IP	II	Acc	BPA	VP	VI	IP	II	Acc	BPA
English (en)	92.1	45.2	75.4	93.1	76.5	51.8	99.7	98.7	97.5	96.8	98.2	98.0
French (fr)	90.4	45.0	70.1	90.7	74.2	49.7	98.1	96.5	90.6	95.8	95.3	92.1
Spanish (es)	91.7	40.4	70.1	91.9	73.7	46.8	97.8	96.4	93.0	95.7	95.7	93.8
Italian (it)	90.4	35.8	67.5	90.2	71.2	43.6	98.3	91.4	93.6	95.4	94.7	90.5
German (de)	87.1	30.4	67.1	90.2	68.8	41.4	96.1	93.9	92.2	94.4	94.2	92.0
Russian (ru)	83.8	30.0	68.4	93.5	69.1	41.8	90.7	82.6	87.0	90.8	87.8	82.6
Chinese (zh)	85.4	24.2	71.4	93.5	68.8	40.1	92.9	81.1	90.3	93.8	89.5	82.7
Bengali (bn)	70.8	36.2	74.8	92.3	68.7	50.8	73.3	72.6	91.7	92.3	82.5	82.0
Swahili (sw)	57.1	22.9	69.7	87.0	59.3	44.0	54.7	48.9	79.1	82.9	66.4	63.2
Telugu (te)	61.7	21.2	79.5	96.8	64.9	46.2	60.1	58.1	93.7	95.1	76.8	75.4

Table 13: Detailed accuracy breakdown in Gemma-2-9B at best α (highest BPA) per language. Left: Baseline model performance. Right: Performance with steering. VP/II are belief-consistent; VI/IP are belief-conflict.

G.4 Gemma-3-12B

Language	Baseline						Steered					
	Categories				Global		Categories				Global	
	VP	VI	IP	II	Acc	BPA	VP	VI	IP	II	Acc	BPA
English (en)	99.7	87.5	79.7	89.0	89.0	79.4	98.9	97.3	97.7	98.3	98.0	97.0
French (fr)	98.8	85.4	65.0	87.4	84.3	69.2	95.3	91.8	94.4	95.3	94.2	92.2
Spanish (es)	98.3	85.4	59.0	87.8	82.8	65.5	92.1	92.9	95.2	96.8	94.2	93.9
Italian (it)	97.5	84.6	62.8	89.8	83.8	67.1	93.1	92.8	95.0	97.3	94.5	93.3
German (de)	94.6	78.8	65.0	86.2	81.2	66.2	94.2	91.5	93.5	95.5	93.7	91.5
Russian (ru)	96.7	70.0	56.4	88.6	78.1	55.1	91.9	83.2	90.9	94.3	90.1	84.6
Chinese (zh)	96.7	72.5	59.4	89.4	79.7	58.1	95.3	86.0	90.3	94.8	91.6	85.3
Bengali (bn)	97.5	82.9	56.0	86.2	80.8	62.7	89.2	87.8	88.5	93.1	89.7	87.0
Swahili (sw)	74.2	46.7	55.1	80.5	64.3	47.3	59.6	49.3	83.9	88.1	70.2	65.2
Telugu (te)	95.0	71.7	68.4	86.2	80.4	63.9	92.1	87.4	85.3	88.5	88.3	84.8

Table 14: Detailed accuracy breakdown in Gemma-3-12B at best α (highest BPA) per language. Left: Baseline model performance. Right: Performance with steering. VP/II are belief-consistent; VI/IP are belief-conflict.

G.5 Mistral-7B

Language	Baseline						Steered					
	Categories				Global		Categories				Global	
	VP	VI	IP	II	Acc	BPA	VP	VI	IP	II	Acc	BPA
English (en)	99.7	41.1	35.7	84.8	65.4	30.2	99.6	96.8	96.5	96.1	97.2	96.1
French (fr)	96.7	31.4	45.3	84.2	64.7	31.0	94.2	87.4	82.3	87.3	87.8	82.6
Spanish (es)	93.8	25.3	45.5	81.3	61.8	29.6	89.0	84.4	82.8	82.9	84.8	82.8
Italian (it)	95.8	32.1	44.4	80.5	63.5	31.8	89.8	78.8	73.3	75.1	79.3	74.2
German (de)	94.6	27.5	38.9	78.4	60.1	28.1	89.3	85.5	79.2	83.5	84.4	81.0
Russian (ru)	93.8	26.6	34.6	79.2	58.9	26.0	79.8	66.7	61.0	68.2	69.0	62.0
Chinese (zh)	83.8	25.3	50.4	80.9	60.4	33.5	78.5	72.9	69.9	77.2	74.7	69.9
Bengali (bn)	71.2	53.6	54.9	64.7	61.1	52.7	72.9	63.6	61.2	69.5	66.8	61.0
Swahili (sw)	37.2	22.7	60.8	71.0	47.9	41.9	37.2	40.3	69.6	68.0	54.0	52.7
Telugu (te)	33.0	29.2	67.5	72.2	50.2	48.1	34.5	35.1	54.9	58.3	45.9	45.2

Table 15: Detailed accuracy breakdown in Mistral-7B at best α (highest BPA) per language. Left: Baseline model performance. Right: Performance with steering. VP/II are belief-consistent; VI/IP are belief-conflict.

G.6 Ministral-3-14B

Language	Baseline						Steered					
	Categories				Global		Categories				Global	
	VP	VI	IP	II	Acc	BPA	VP	VI	IP	II	Acc	BPA
English (en)	88.4	61.4	78.7	92.3	80.2	63.9	98.8	98.0	97.2	96.6	97.7	97.6
French (fr)	85.4	48.8	70.1	91.5	74.1	52.6	94.2	93.3	90.6	96.6	93.7	90.5
Spanish (es)	78.3	38.3	75.2	94.7	71.8	50.4	91.5	89.2	92.3	96.3	92.4	89.4
Italian (it)	79.2	41.2	70.1	90.2	70.3	49.9	93.6	91.5	90.9	92.7	92.2	90.4
German (de)	72.1	36.2	77.3	92.7	69.7	51.9	94.0	89.9	93.0	95.5	93.1	90.0
Russian (ru)	70.4	30.0	76.5	95.1	68.1	48.0	86.1	73.2	87.3	92.3	84.8	77.2
Chinese (zh)	72.1	25.4	70.9	95.9	66.2	42.5	85.6	71.7	93.3	96.6	86.8	79.3
Bengali (bn)	70.8	27.9	80.8	94.7	68.7	49.1	78.1	64.4	89.6	95.5	81.9	73.9
Swahili (sw)	29.2	15.8	85.5	89.0	54.9	50.3	29.7	23.9	95.9	95.4	61.2	59.6
Telugu (te)	71.2	38.3	73.9	90.2	68.5	51.7	72.1	57.4	95.3	95.7	80.1	74.1

Table 16: Detailed accuracy breakdown in Ministral-3-14B at best α (highest BPA) per language. Left: Baseline model performance. Right: Performance with steering. VP/II are belief-consistent; VI/IP are belief-conflict.

H Steering Strength Ablation Study

H.1 Qwen-2.5-7B

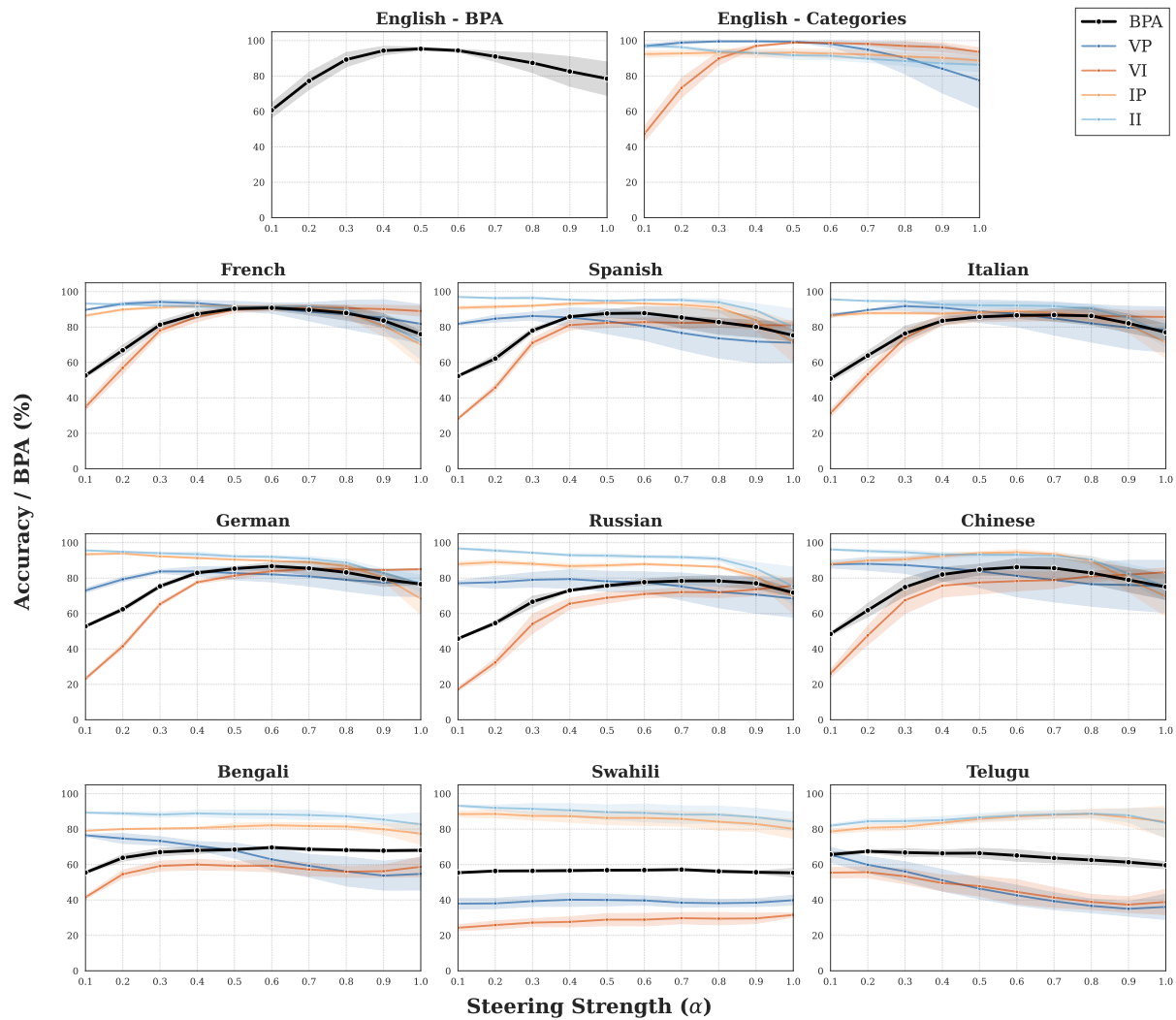


Figure 7: Steering strength (α) ablation study for Qwen-2.5-7B, showing per-category accuracy and BPA (black) across all languages. VP/II (blue) are belief-consistent; VI/IP (orange) are belief-conflict.

H.2 Qwen-3-14B

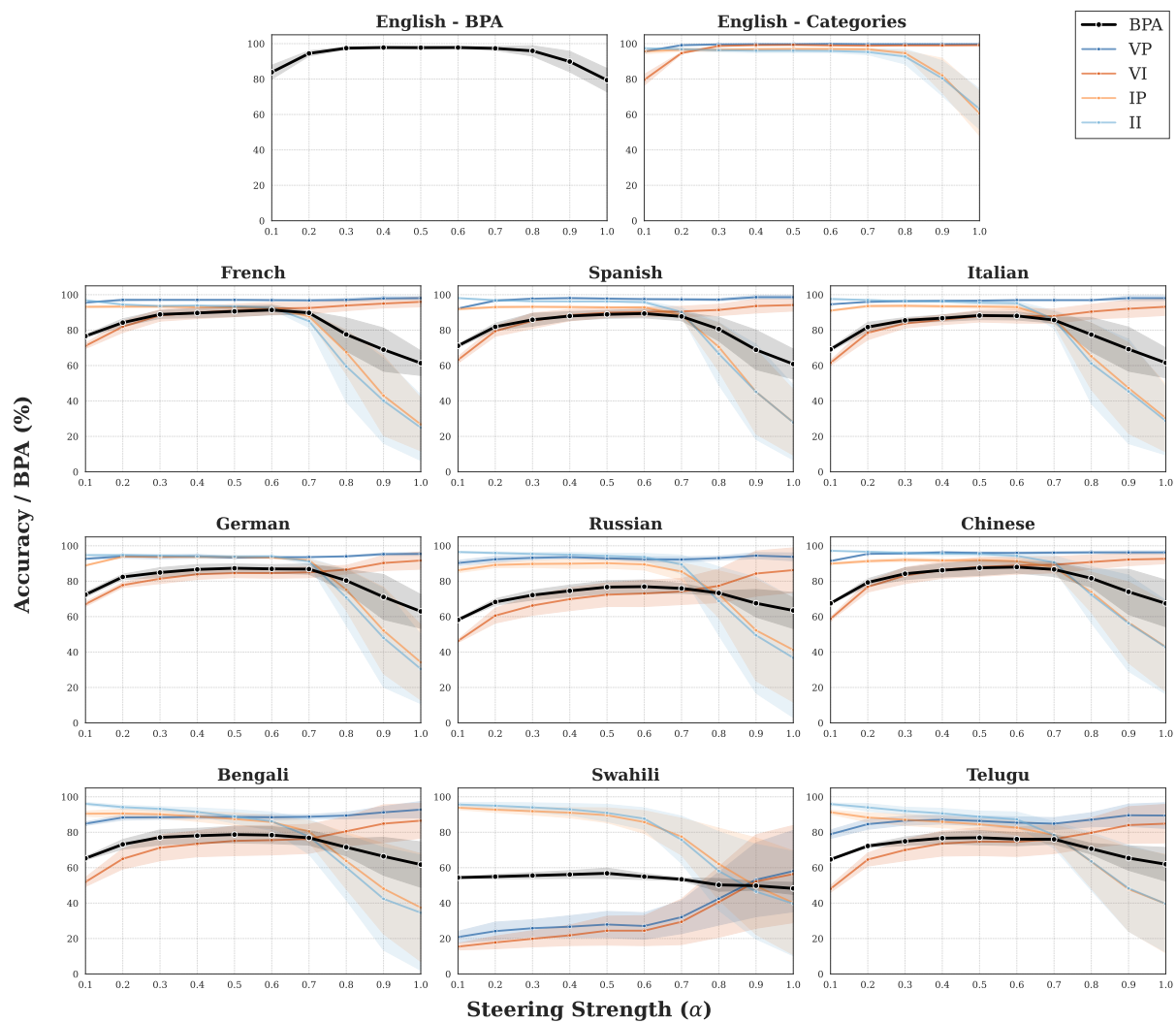


Figure 8: Steering strength (α) ablation study for Qwen-3-14B, showing per-category accuracy and BPA (black) across all languages. VP/II (blue) are belief-consistent; VI/IP (orange) are belief-conflict.

H.3 Gemma-2-9B

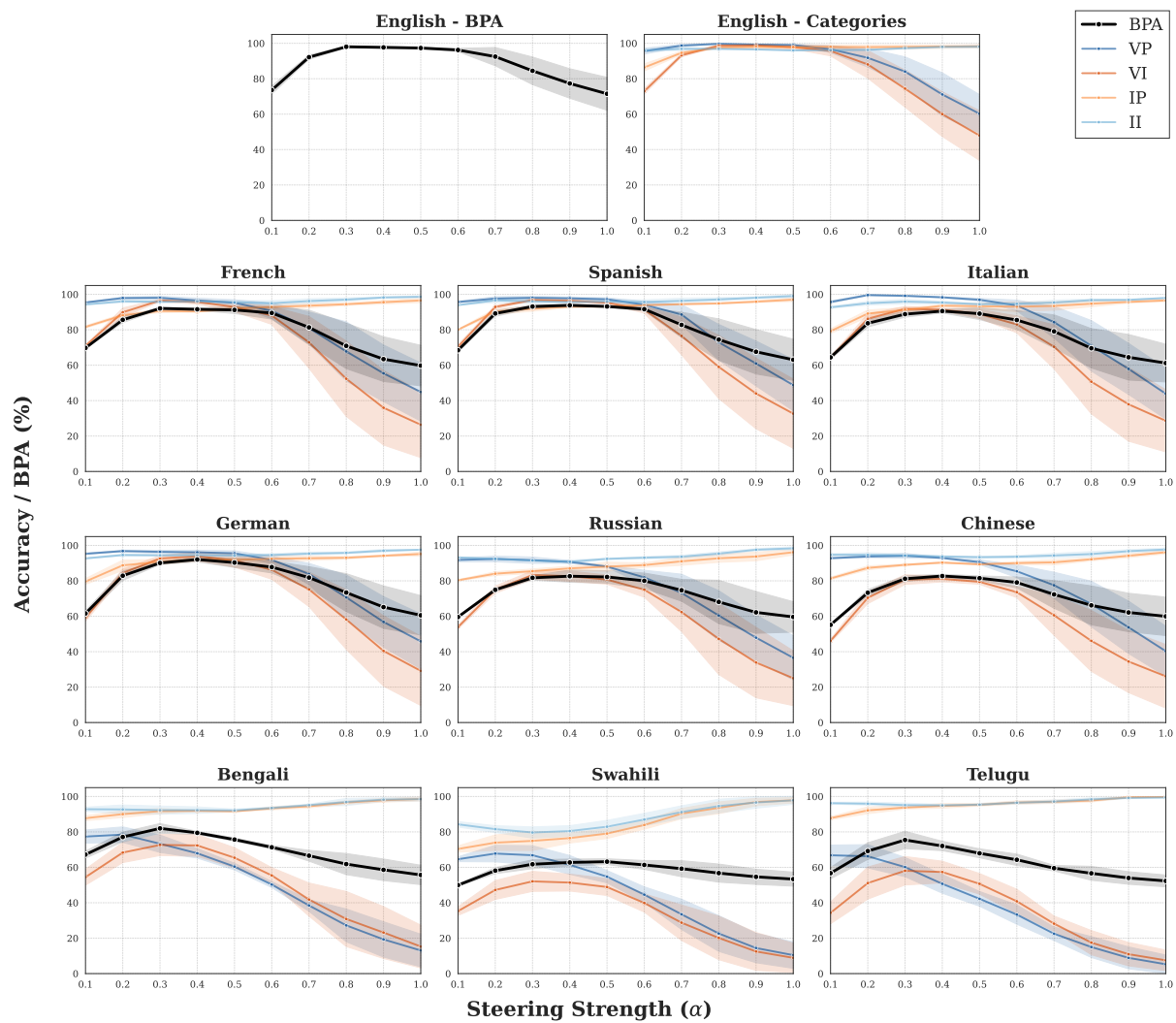


Figure 9: Steering strength (α) ablation study for Gemma-2-9B, showing per-category accuracy and BPA (black) across all languages. VP/II (blue) are belief-consistent; VI/IP (orange) are belief-conflict.

H.4 Gemma-3-12B

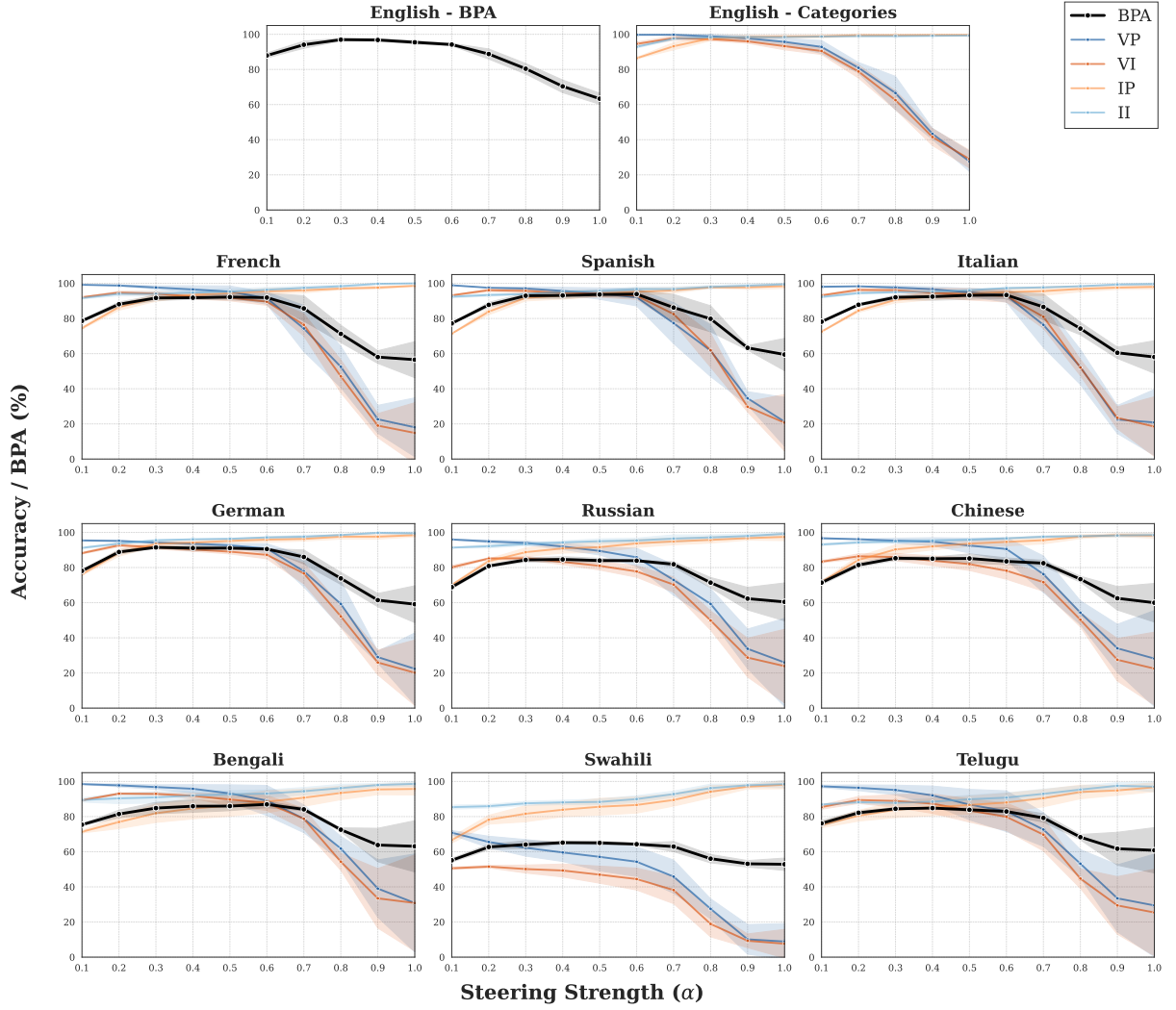


Figure 10: Steering strength (α) ablation study for Gemma-3-12B, showing per-category accuracy and BPA (black) across all languages. VP/II (blue) are belief-consistent; VI/IP (orange) are belief-conflict.

H.5 Mistral-7B

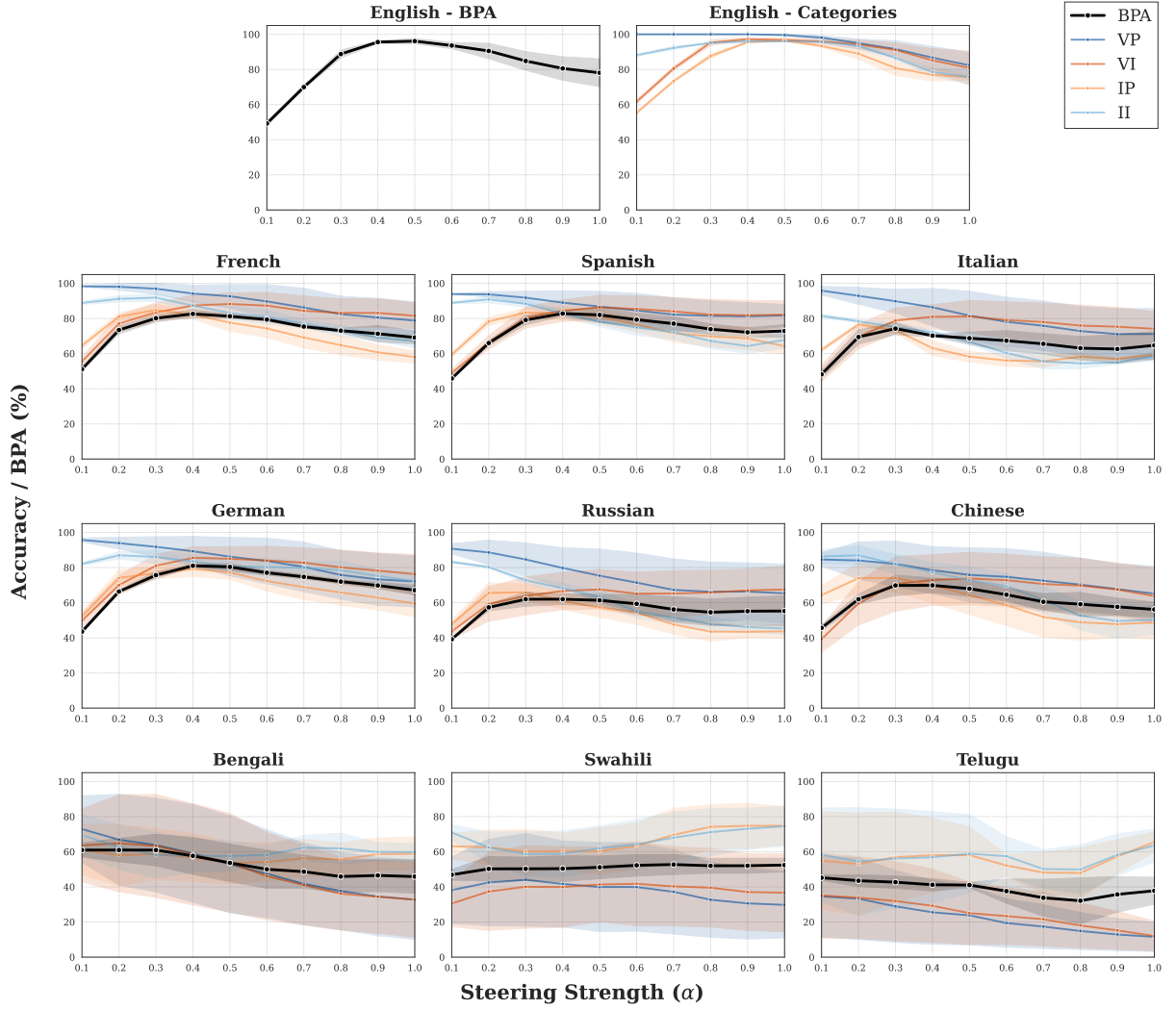


Figure 11: Steering strength (α) ablation study for Mistral-7B, showing per-category accuracy and BPA (black) across all languages. VP/II (blue) are belief-consistent; VI/IP (orange) are belief-conflict.

H.6 Ministral-3-14B

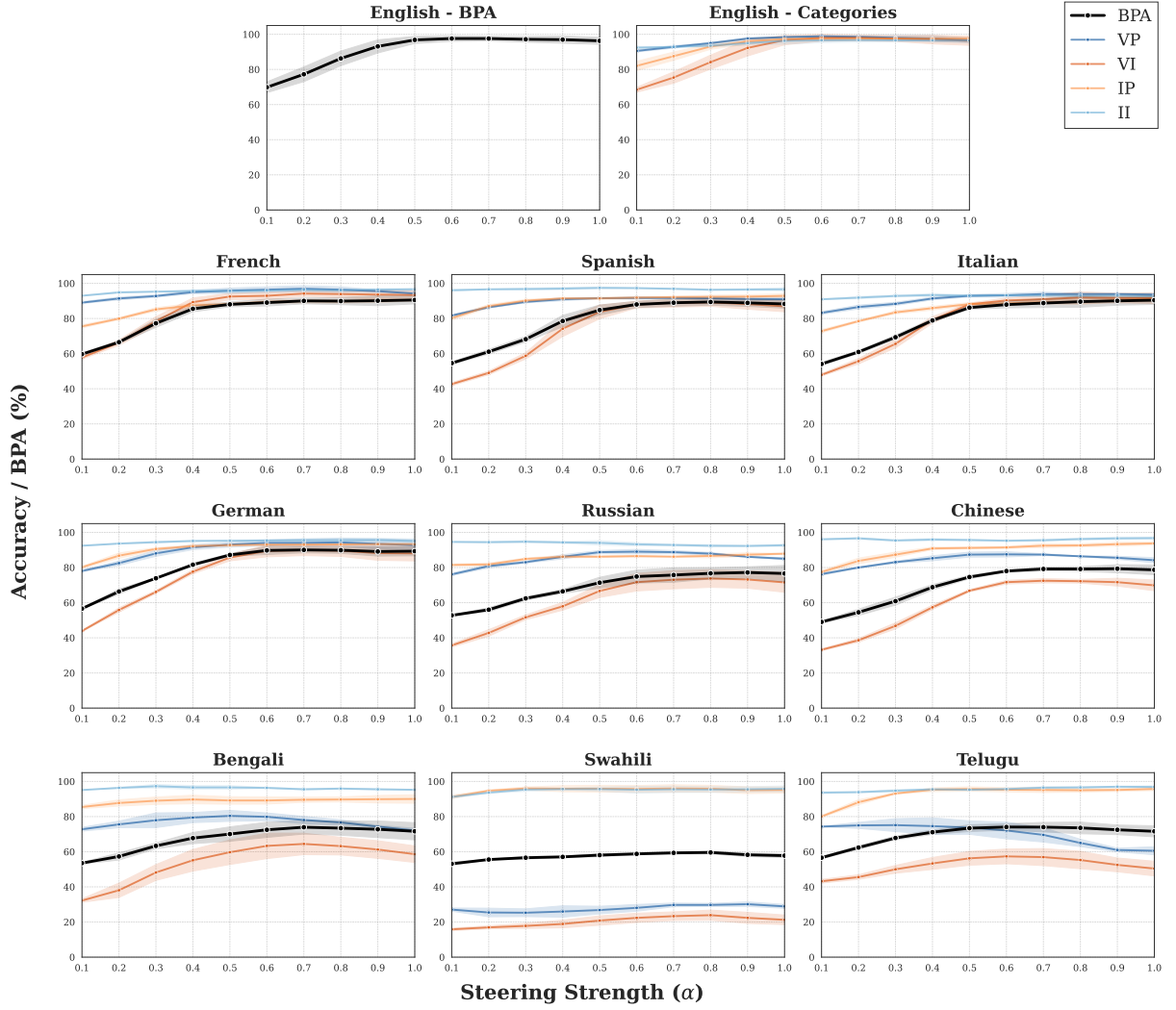


Figure 12: Steering strength (α) ablation study for Ministral-3-14B, showing per-category accuracy and BPA (black) across all languages. VP/II (blue) are belief-consistent; VI/IP (orange) are belief-conflict.

I Detailed PPL Results

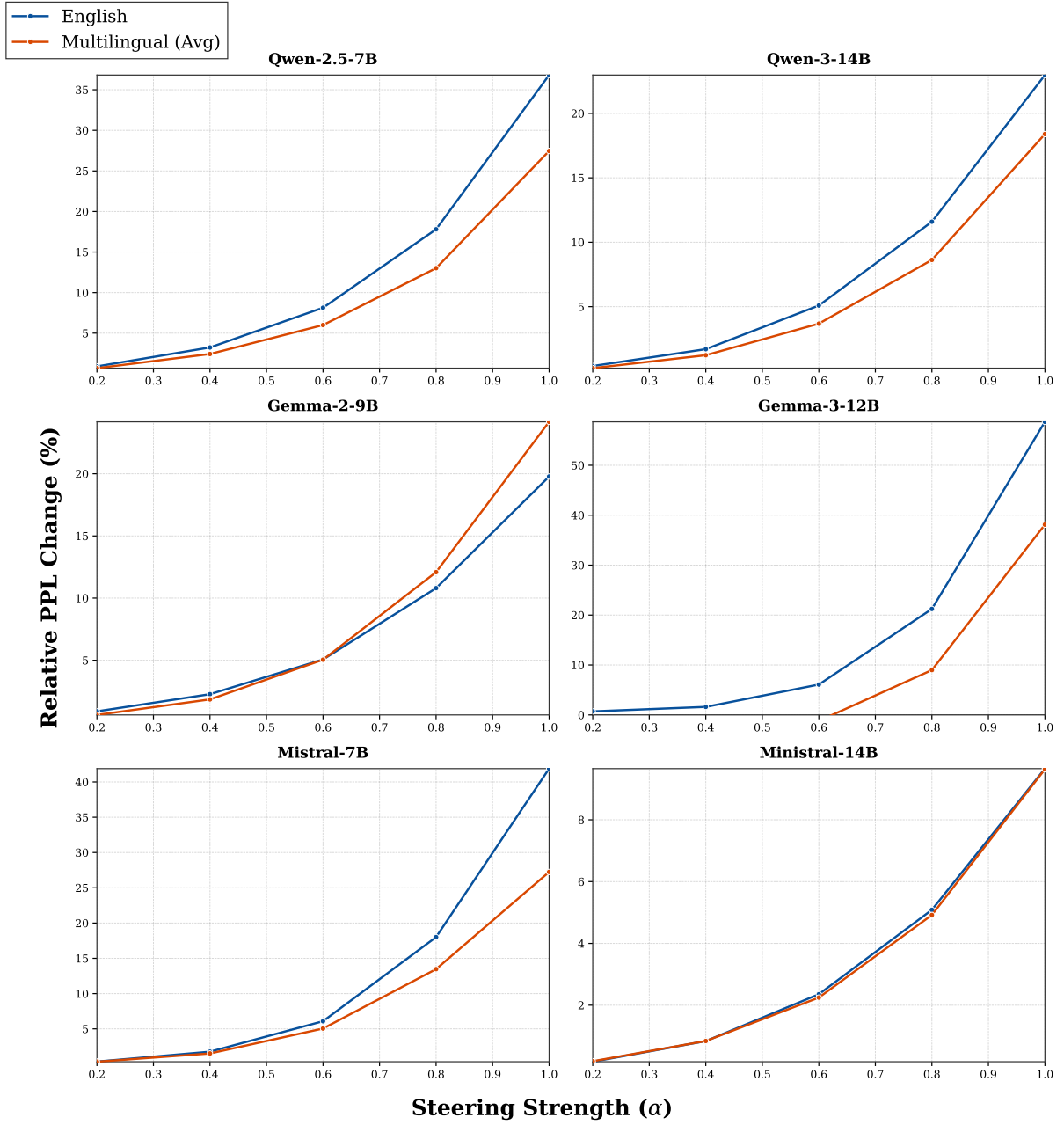


Figure 13: Perplexity degradation as a function of steering strength across all models and languages. At optimal steering configurations ($\alpha \approx 0.4 - 0.6$), fluency remains largely preserved. Significant degradation emerges only at aggressive steering strength ($\alpha \geq 0.8$).

J Detailed OOD Task Results

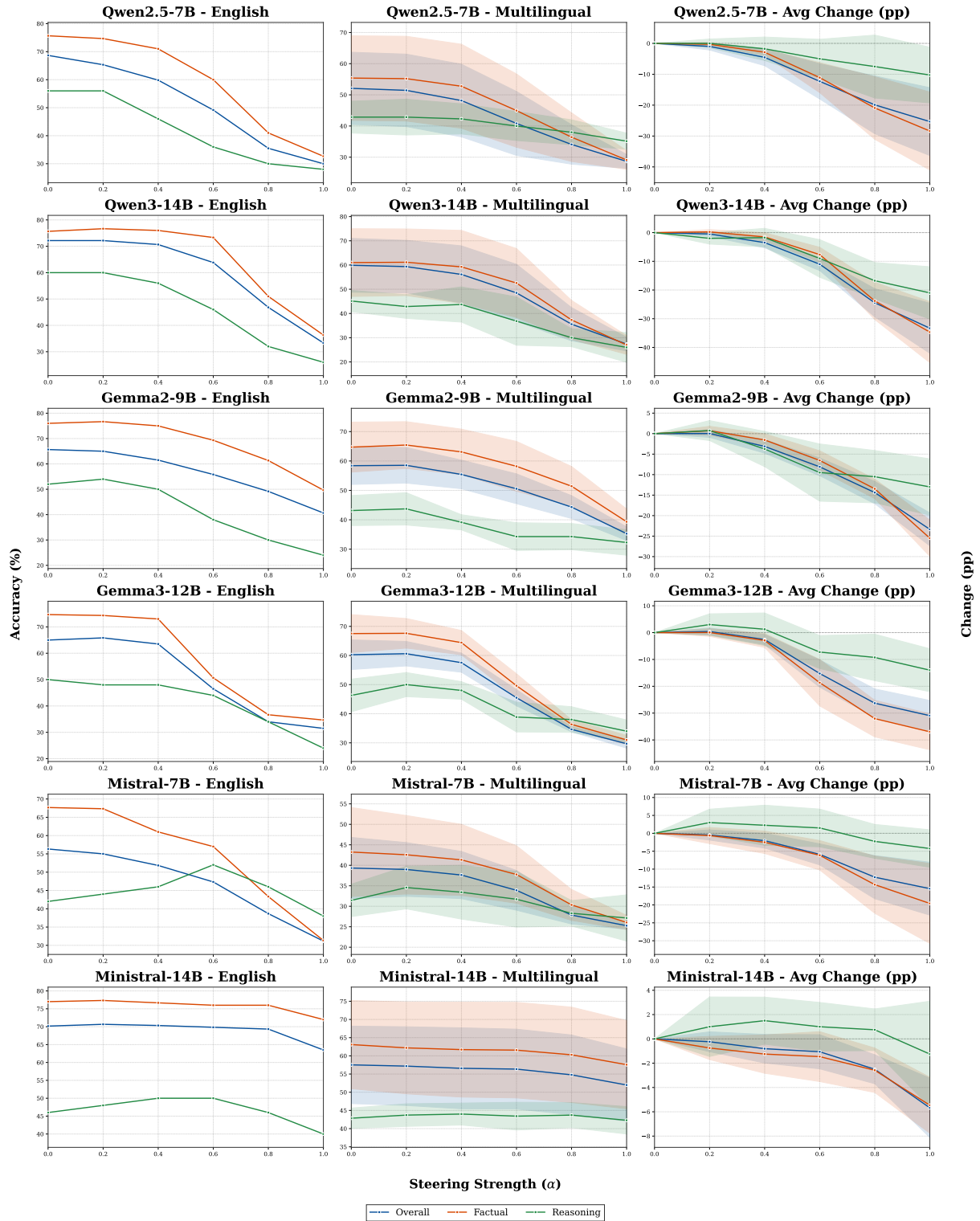
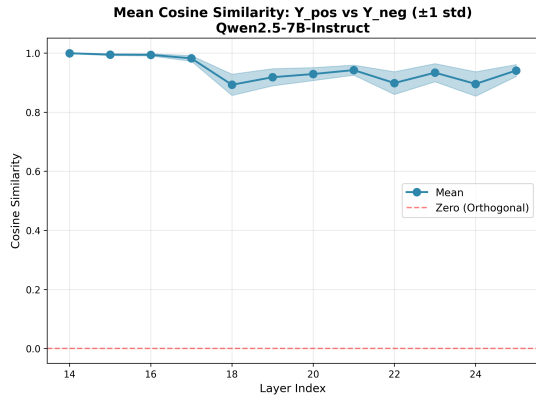
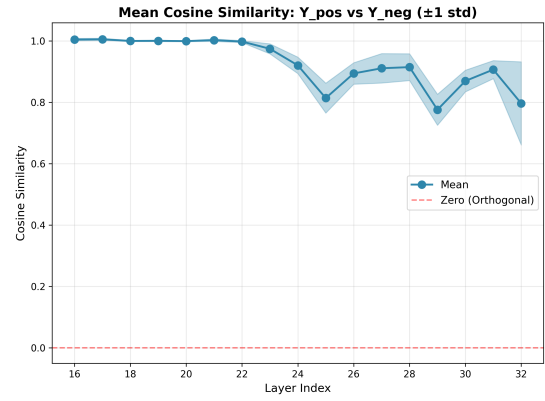


Figure 14: Performance impact on MMMLU subtasks across models. Factual knowledge tasks (orange) degrade sharply as steering suppresses semantic content, while Formal Logic tasks (green) often show more resilience. This divergence confirms the specificity of the intervention mechanism.

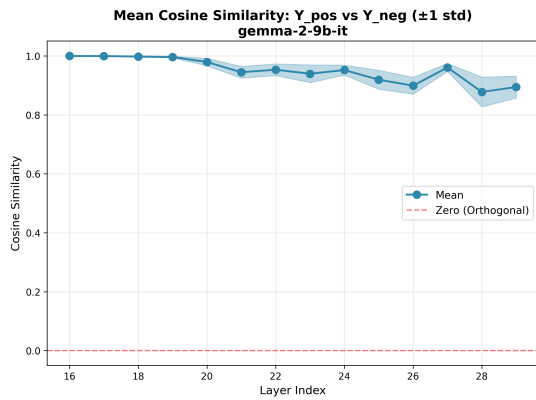
K Positive-Negative Cosine Similarity Analysis



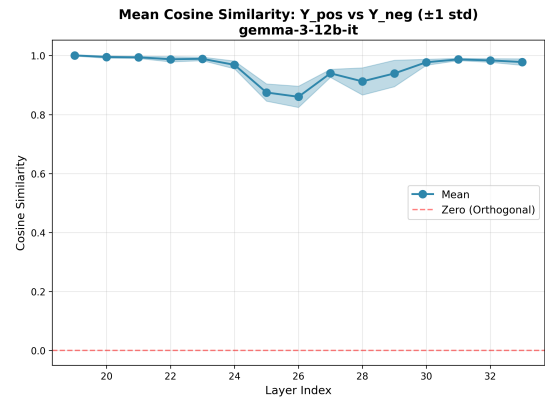
(a) Qwen2.5-7B



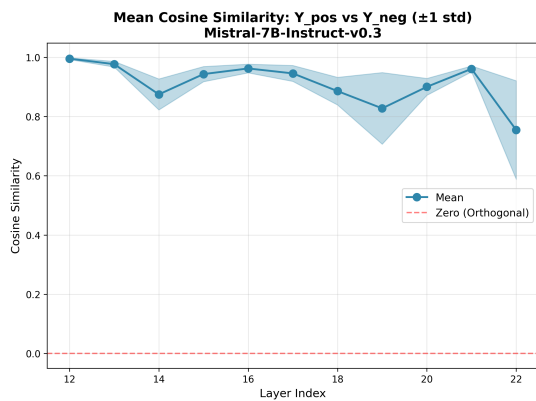
(b) Qwen3-14B



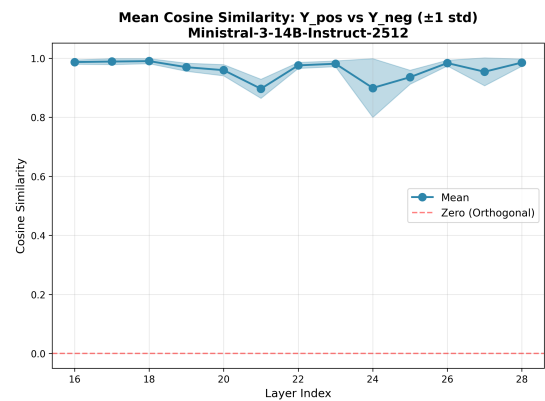
(c) Gemma-2-9B



(d) Gemma-3-12B



(e) Mistral-7B



(f) Ministral-3-14B

Figure 15: Layer-wise cosine similarity between positive and negative training target pairs across all models.