Can Activation Steering Generalize Across Languages?
A Study on Syllogistic Reasoning in Language Models

Gabriele Maraia (†) **Leonardo Ranaldi** (⊕,†)
Marco Valentino(•) **Fabio Massimo Zanzotto** (†,‡)
(⊕) ILCC, School of Informatics, University of Edinburgh, United Kingdom
(•) School of Computer Science, University of Sheffield, United Kingdom
(†) Human Centric ART, University of Rome Tor Vergata, (‡) Almawave S.p.A.
{first_name.last_name}@uniroma2.it

Abstract

Large Language Models (LLMs) often struggle with formal logical reasoning, frequently conflating content plausibility with logical validity. This well-known content effect undermines their capacity to act as reliable deductive reasoners, particularly in multilingual contexts where both linguistic variability and world knowledge may deepen biases. Prior work shows that prompting and tuning interventions can alleviate these issues only partially, leaving models vulnerable to semantic interference. While previous studies have explored activation steering and other test-time interventions, this work has focused predominantly on English.

To make reasoning more consistent, robust, and transferable across languages, we investigate the use of activation steering – an inference-time intervention that modulates internal representations towards a cross-lingual reasoning space. Our experiments demonstrate that steering techniques constructed for English-based syllogisms generalise effectively to multilingual datasets, yielding higher formal reasoning accuracy (up to +36%) while minimally affecting language modelling performance. Moreover, steering supports partial transfer to out-of-distribution tasks, highlighting its potential as a scalable mechanism for cross-lingual transferable reasoning. These findings advance the prospect of developing LLMs that can serve as reliable soft reasoners across languages.

1 Introduction

Reasoning with natural language has long been considered a significant challenge for both cognitive science and AI. Within this area, syllogistic reasoning provides a well-established testbed, allowing the disentanglement of formal validity from semantic plausibility (Evans et al., 1983). Recent studies have shown that Large Language Models (LLMs) exhibit systematic biases when confronted with syllogisms: they often mislead plausibility with logical consequence, rarely produce 'nothing follows' for invalid inferences, and mirror well-documented human tendencies, such as the content effect (Eisape et al., 2024). Although recent efforts on reasoning transfer (Ranaldi and Freitas, 2024) and decomposition techniques (Ranaldi et al., 2025c) have addressed these limitations, they still persist. At the same time, a growing body of work highlights that LLMs' internal representations can be steered to mitigate reasoning biases. Activation steering (Rimsky et al., 2024; Turner et al., 2024; Valentino et al., 2025), in particular, offers an inference-time mechanism to disentangle semantic plausibility from formal validity, thereby reducing content effects without requiring retraining. Yet, these approaches have so far been confined to English (Valentino et al., 2025), leaving unexplored whether such interventions generalise beyond a single language. This gap is critical: *if steering is to serve as a principled mechanism for controlling reasoning, it must operate independently of linguistic features*.

To make reasoning more consistent, robust, and transferable across languages, we investigate activation steering in multilingual *syllogistic reasoning*. We demonstrate that steering interventions, learnt and designed for English, actually generalise to multilingual landscapes, yielding higher formal accuracy while minimally affecting fluency. We operationalise steering through two complementary families of methods: *(i)* **direct geometric shifting**, which adds a vector to reposition the model's activations within the representation space, and *(ii)* **learned abstractive transformation**, which employs neural networks to map content-specific activations to their content-agnostic, formal-logic counterparts. Our findings indicate that all proposed families of methods enhance logical consistency, with abstractive transformation strategies emerging as particularly interesting, as their effectiveness appears to depend on the quality of

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 2739–2753
March 24-29, 2026 ©2026 Association for Computational Linguistics

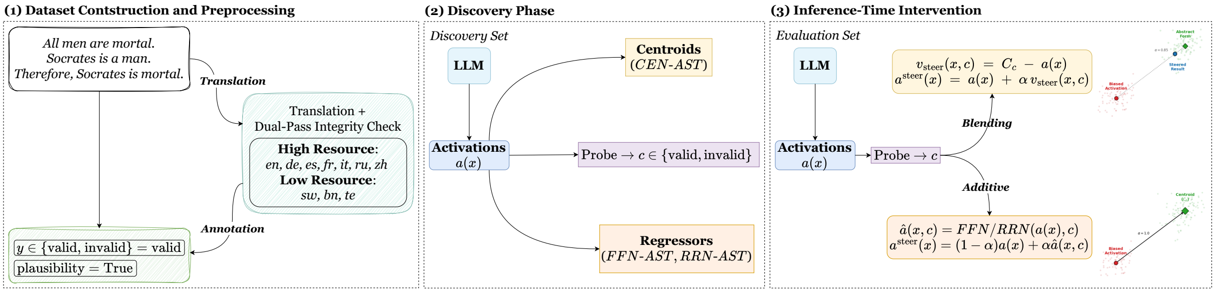
(1) Dataset Contstruction and Preprocessing
All men are mortal.
Socrates is a man.
Therefore, Socrates is mortal.
Translation
Translation +
Dual-Pass Integrity Check
High Resource:
en, de, es, fr, it, ru, zh
Low Resource:
sw, bn, te
Annotation
$y \in \{\text{valid, invalid}\} = \text{valid}$
$\text{plausibility} = \text{True}$
(2) Discovery Phase
Discovery Set
LLM
Activations
$a(x)$
Centroids
(CEN-AST)
Probe $\rightarrow c \in \{\text{valid, invalid}\}$
Regressors
(FFN-AST, RRN-AST)
(3) Inference-Time Intervention
Evaluation Set
LLM
Activations
$a(x)$
Probe $\rightarrow c$
Blending
Additive
$v_{\text{steer}}(x,c) = C_c - a(x)$
$a^{\text{steer}}(x) = a(x) + \alpha v_{\text{steer}}(x,c)$
$\hat{a}(x,c) = FFN/RRN(a(x),c)$
$a^{\text{steer}}(x) = (1-\alpha)a(x) + \alpha\hat{a}(x,c)$
Abstract Form
$\alpha = 0.85$
Steered Result
Biased Activation
Centroid (C_c)
$\alpha = 1.0$
Biased Activation

Figure 1: Overview of the pipeline for investigating the cross-lingual generalization of activation steering methods for syllogistic reasoning. The process consists of three stages: (1) *Dataset Construction*, where English syllogisms are translated into high- and low-resource languages using a round-trip back-translation protocol for quality control; (2) *Training Phase*, where internal model activations $a(x)$ from a discovery set are used to fit a linear probe (validity classifier), class-specific centroids (CEN-AST), and abstractive regressors (FFN/RRN-AST); and (3) *Inference-Time Intervention*, where the probe gates the input activation to apply either additive shifting or blending transformations to produce the steered activation $a^{steer}(x)$.

the model's internal geometry. Collectively, these methods reveal steering as a cross-lingual intervention, showing consistent benefits across structurally diverse languages and partial transfer to out-of-distribution logical tasks.

Our contributions are threefold: *(i)* we introduce the first systematic evaluation of activation steering in multilingual syllogistic reasoning; *(ii)* we provide empirical evidence that steering suppresses content effects while preserving language modelling performance; and *(iii)* we release our code and datasets to support future research.

To the best of our knowledge, our work is the first to apply activation steering to instruct LLMs to perform multilingual syllogistic reasoning, demonstrating that steering interventions for English generalise across languages and remain effective in diverse linguistic settings.

2 Methods

This section formalises the task and the activation steering strategies adopted to investigate cross-lingual transfer on syllogistic reasoning. Figure 1 provides an overview of the methodology and the proposed pipeline.

2.1 Problem Formulation and Notation

Let x denote a syllogistic argument presented in natural language. The task is to determine its formal validity, denoted by the label $y \in \{\text{valid}, \text{invalid}\}$. Our intervention targets the hidden activations within the model's residual stream. We denote the activation at layer ℓ and token position t as $a_{\ell,t}(x) \in \mathbb{R}^d$, where d is the hidden dimension of the model.

For a given input x, the base activation for steering, $a(x)$, is defined as the mean-pooled activation vector across all token positions at a target layer $\ell^\star$:

$$a(x) \equiv \frac{1}{|x|}\sum_{t=1}^{|x|} a_{\ell^\star,t}(x). \qquad (1)$$

The target layer $\ell^\star$ is selected from the final third of the model's layers, where prior probing studies (Valentino et al., 2025) indicate that validity information is most prominent. The intervention modifies this base activation to produce a steered activation, $a^{\text{steer}}(x)$, with the magnitude of the intervention controlled by a scalar hyperparameter $\alpha \in \mathbb{R}$.

2.2 Probe Gating: Conditional Intervention

The core assumption underlying our approach is that the model's internal representations for valid and invalid syllogisms are geometrically separable. Our empirical analysis confirms this hypothesis; as visualised in Figure 2, activations from a target layer form distinct clusters corresponding to their logical validity.

This also suggests that valid and invalid activations do not simply align along opposite activations, motivating the need for class-specific transformations (rather than a single transformation with dynamic sign flipping). Capitalising on this structure, we implement a conditional intervention framework. We train a linear probe, specifically a Logistic Regression model, on activations from a held-out discovery set to act as a validity classifier. We refer to this mechanism as **probe gating**. At inference time, for a new input x, the probe predicts its

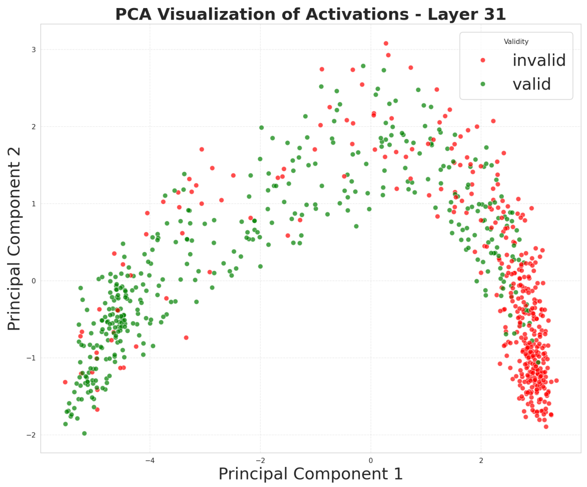

Figure 2: PCA visualisation of activations from Layer 31 of Gemma-2-9B. Each point corresponds to a syllogism and is coloured by its ground-truth validity (green for valid, red for invalid), revealing clear geometric separability.

validity class based on its activation geometry:

$$c = g_{\text{probe}}(a(x)), \tag{2}$$

where $c \in \{\text{valid}, \text{invalid}\}$. This prediction, c, then determines which class-specific steering transformation to apply, enabling targeted intervention.

2.3 Activation Steering Strategies

We investigate two families of steering strategies: additive methods that translate the original activation, and blending methods that interpolate towards a reconstructed, content-abstracted activation.

2.3.1 Additive Strategy

Additive steering modifies the original activation by adding a class-conditioned direction vector $v_{\text{steer}}(x, c)$, scaled by α:

$$a^{\text{steer}}(x) \;=\; a(x) \;+\; \alpha\, v_{\text{steer}}(x, c). \tag{3}$$

CEN-AST (Centroid-based Steering). This strategy computes a dynamic, point-dependent shift towards a pre-computed class centroid. First, we compute class centroids C_c using activations from the discovery set where the model's prediction was correct: $C_c = \frac{1}{N_c} \sum_{j=1}^{N_c} a_j$. The steering vector then directs the current activation $a(x)$ towards the centroid C_c corresponding to the probe-predicted class c:

$$v_{\text{steer}}(x, c) \;=\; C_c \;-\; a(x). \tag{4}$$

2.3.2 Blending Strategies

Blending strategies first generate a target activation, $\hat{a}(x, c)$, that represents a content-agnostic version of the input syllogism. The final steered activation is then computed by linearly interpolating between the original and the target activations:

$$a^{\text{steer}}(x) \;=\; (1 - \alpha)\, a(x) \;+\; \alpha\, \hat{a}(x, c). \tag{5}$$

The key to these strategies lies in how the target activation $\hat{a}$ is obtained. The supervision for this process is provided by a parallel dataset constructed from the English discovery set. For each content-laden syllogism, we create a corresponding *abstract* version by systematically replacing content-bearing terms with placeholder variables (e.g., "all flowers are plants" is mapped to "all X are Y"). This establishes a direct, one-to-one correspondence between a syllogism and its content-free logical form.

We then train class-specific regressors to learn the mapping from the activation of a content-rich syllogism to that of its abstract counterpart. This process forces the learned transformation to isolate the underlying logical structure from the distracting lexical content.

FFN-AST (Two-Headed Feed-Forward Steering). For each class c, we train a separate feed-forward network to predict the target activation $\hat{a}$. To stabilise training and improve geometric fidelity, the network has two heads that predict the direction and magnitude of the target vector separately. The **direction head** outputs a vector whose normalised form $\hat{d}_c$ is optimised using a cosine similarity loss against the true abstract activation's direction. The **magnitude head** outputs a scalar $\hat{m}_c$, optimised with a Mean Squared Error (MSE) loss. The final target activation is reconstructed as $\hat{a}(x, c) = \hat{d}_c\, \hat{m}_c$.

RRN-AST (Rotational-Residual Network Steering). This method refines the blending approach by modelling the transformation from a content-laden to an abstract activation as an explicit geometric operation rather than a black-box function. For each class c, we train a network that learns a structured transformation consisting of a rotation and a residual translation. The process involves four steps:

1. **Projection:** The input activation $a(x)$ is projected into a low-dimensional latent space using PCA, yielding a_ℓ. The PCA is fitted on discovery activations for class c.

2. **Parameter Prediction:** A small network takes a_ℓ as input and predicts the parameters of the transformation: an orthogonal rotation matrix R_c and a residual vector δ_c.
3. **Latent Transformation:** The transformation is applied in the latent space: $\hat{a}_\ell = R_c a_\ell + \delta_c$.
4. **Inverse Projection:** The resulting vector $\hat{a}_\ell$ is projected back into the original high-dimensional activation space to yield the final target activation $\hat{a}(x, c)$.

2.4 Implementation Details

Our strategies target a contiguous block of layers identified through preliminary probing. The steering coefficient α is fixed for the additive CEN-AST strategy and for the blending strategies based on validation performance. Final outputs are obtained via greedy decoding and parsed for validity keywords to compute our metrics.

Full details, including training hyperparameters, are provided in Appendix A.1.

3 Experimental Setup

3.1 Models

To evaluate the effectiveness and generalisability of our steering methods, we experimented on a set of open-source LLM families. We prioritised models with varying architectures and training data to assess the robustness of our techniques. This set includes Qwen-2.5-7B (Qwen et al., 2025), Gemma-2-9B (Gemma Team et al., 2024), Mistral-7B (Jiang et al., 2023), Llama-3.1-8B (AI@Meta, 2024), EuroLLM-9B (Martins et al., 2025) and Velvet-2B (Almawave, 2024).

3.2 Dataset

We conduct all experiments on a controlled corpus of syllogistic arguments. Building on Valentino et al. (2025) and Bertolazzi et al. (2024) as a seed dataset, we extended it to construct a collection of syllogistic arguments spanning 24 logical schemes. Each scheme was instantiated through structured natural-language templates derived from first-order logic forms, populated with taxonomic relations from WordNet. Each instance comprises two premises and a conclusion, together with two ground-truth Boolean labels indicating the argument's logical **validity** and the **plausibility** of its conclusion. For example:

> **Premise 1:** all flowers are plants.
> **Premise 2:** no blossoms are plants.
> **Conclusion:** some blossoms are not flowers.
> **Validity:** `true`
> **Plausibility:** `true`

Languages Starting from the original version in English (en), we construct multilingual versions in German (de), Spanish (es), French (fr), Italian (it), Russian (ru), Chinese (zh), Swahili (sw), Bengali (bn), and Telugu (te). We translate into the target language using GPT-4 and then apply round-trip back-translation to English and assess consistency.

Plausibility Annotation Plausibility labels indicate whether a conclusion accords with common sense or world knowledge independently of formal validity. We follow a lightweight yet reliable protocol: each English item inherits the gold plausibility label from the source corpus; for translated items, plausibility is *constrained to be label-invariant* under the QC pipeline above. During bilingual review, annotators confirm that lexical choices in the target language do not inadvertently alter real-world likelihood When uncertainty arises, the item is revised or removed. This procedure preserves cross-lingual comparability while avoiding misleading content.

Evaluation split To ensure robust and unbiased assessment, we split the data into a 70% **training set** and a 30% **evaluation set**. The training set is used to fit the steering components (e.g., centroid vectors) and to train the probe, FFN, and RRN variants; the evaluation set is held out strictly for cross-validation testing.

3.3 Metrics

We compare our methods to baselines using the following metrics:

Accuracy. The proportion of correct responses over all items, reported as the mean of the four category accuracies VP (valid & plausible), VI (valid & implausible), IP (invalid & plausible) and II (invalid & implausible).

Content Effect (CE). The degree to which performance is driven by plausibility rather than formal validity, decomposed as:

- **Cross-Plausibility CE** (Cross-CE): the absolute difference between accuracy on plausible vs. implausible items:

$$\text{Cross-CE} = \left|\text{acc}(\text{VP} \cup \text{IP}) - \text{acc}(\text{VI} \cup \text{II})\right| . \tag{6}$$

- **Intra-Plausibility CE** (Intra-CE): the mean of absolute within-plausibility gaps:

$$\text{Intra-CE}_{\text{P}} = \left|\text{acc}_{\text{VP}} - \text{acc}_{\text{IP}}\right|, \quad (7)$$

$$\text{Intra-CE}_{\text{I}} = \left|\text{acc}_{\text{II}} - \text{acc}_{\text{VI}}\right|, \quad (8)$$

$$\text{Intra-CE} = \frac{\text{Intra-CE}_{\text{P}} + \text{Intra-CE}_{\text{I}}}{2}. \quad (9)$$

We report **Overall CE** as the mean of Cross- and Intra-CE.

ACE Ratio Our principal summary of the trade-off between accuracy and content invariance: the ratio of **Global Accuracy** to **Overall CE**.

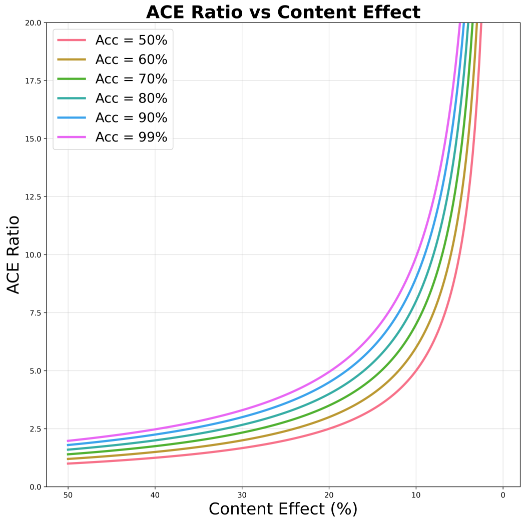

Figure 3: Relationship between ACE Ratio and Content Effect for different Accuracy levels (60% to 99%) on synthetic data. The ACE Ratio increases exponentially toward a vertical asymptote as the Content Effect approaches 0%. This hyperbolic nature of the ACE Ratio makes its interpretation non-linear.

We propose the ACE Ratio as our primary metric because it rewards logical performance improvement while penalising reliance on content-based heuristics.

Steering Efficiency (η) : Our proposed secondary metric is designed to diagnose failures and decouple the intrinsic efficacy of strategies from the accuracy of probe gating, which can mask the strategies' true potential. η is computed as the conditional accuracy limited to the subset of examples where the linear probe's prediction was correct, providing an upper bound on the strategy's accuracy, assuming a perfectly accurate validity signal.

ACE vs. Efficiency The ACE Ratio provides a precise tool for comparing the scalability of the presented strategies, accounting for both steering and probe-gating performance. However, because suboptimal results could be due solely to probe performance, Steering Efficiency is the key metric for evaluating how well English-trained assets generalise to other languages. We make this distinction because our simple English-trained probe gating serves as a stress-test for the multilingual internal representations of the analysed LLMs. As acknowledged in Section 6, a more robust multilingual gating probe could be readily implemented.

4 Results

4.1 Activation Steering Improves Logical Consistency

Across all tested models, activation steering delivers substantial improvements in logical reasoning, both in English and, critically, across nine other languages. This confirms that interventions trained on English data can generalise effectively as a robust cross-lingual control mechanism. The summary of these results in Table 1 highlights three key trends: *(i)* steering provides a universal boost over baselines and improves the ACE Ratio, *(ii)* the interventions show effective cross-lingual generalisation, and, most critically, *(iii)* the results reveal model dependent efficacy.

Models like Qwen and Gemma respond well to all strategies, while the more complex blending strategies fail on Mistral and Llama. This motivates our deeper diagnostic analysis in the following section, where we investigate the geometric properties underlying these successes and failures.

The detailed results for each language are reported in AppendixB.

4.1.1 Case Studies: Qwen and Gemma

As shown in Table 1, Qwen and Gemma respond exceptionally well to all steering interventions. We attribute this susceptibility to their well-structured internal geometry, specifically the high linear separability of validity representations demonstrated by our probing diagnostics (Section 2.2). Although their baselines indicate significant vulnerability to content bias, all strategies deliver substantial gains. The straightforward additive CEN-AST boosts Qwen's ACE Ratio by over 700%, while the more complex blending strategies (FFN-AST/RRN-AST) are also highly effective, confirming their ability to push the model toward an abstract, content-free reasoning space.

4.2 Cross-Lingual Generalisation

A central hypothesis of this work is that activation steering can function as a cross-lingual control mechanism. Our findings reveal a more nuanced reality: the ability to generalise steering interventions is model-dependent. In contrast to the robust cross-lingual generalisation observed in Qwen-2.5-7B and Gemma-2-8B, models such as Mistral-7B and Llama-3.1-8B fail to generalise effectively, particularly with the more complex blending strategies.

For Mistral-7B, FFN-AST, and RRN-AST, which perform reasonably well in English, performance collapses in multilingual settings, resulting in an average degradation in the ACE Ratio.

Interestingly, Llama-3.1-8B struggles to learn the in-domain English transformation in the first place.

These failures are not uniform but arise from distinct, diagnosable issues in the models' internal representations, which we investigate using the Steering Efficiency (η) metric.

4.3 Diagnosing Failures

Our analysis reveals that the poor performance of blending strategies in Mistral and Llama stems from two fundamentally different issues: one related to generalisation, and the other to learnability.

Case Study 1: Mistral At first glance, Mistral's poor multilingual results might suggest that its internal representations of logic are not generalisable across languages. However, our analysis provides a more precise explanation. The simple additive CEN-AST strategy achieves near-perfect Steering Efficiency ($\eta \approx 100\%$) across all languages, demonstrating that Mistral indeed possesses a geometrically consistent, cross-lingual path for logical validity. The ACE Ratio degradation reported in Table 1 is a consequence of the poor scalability of the English-trained probe.

The failure of the non-linear blending strategies FFN-AST and RRN-AST, as evidenced by poor Steering Efficiency in multilingual settings, stems from the complexity of their learned transformations. To test this hypothesis, we implemented a linear abstractive strategy based on ridge regression. As in Table 2, Ridge-AST not only performs well in English but also maintains a relatively high Steering Efficiency across all languages.

Model	Strategy	Metric	en	multi	$\Delta(\%)$
Qwen-7B	Baseline	Acc	78.06	65.27	-
		ACE	2.38	2.40	-
	CEN-AST	Acc	**97.30**	83.43	+27.82%
		ACE	**59.58**	15.88	+561.67%
	FFN-AST	Acc	96.87	**83.81**	+28.41%
		ACE	35.84	14.53	+505.42%
	RRN-AST	Acc	97.16	83.65	+28.16%
		ACE	42.46	**16.64**	+593.33%
Gemma-9B	Baseline	Acc	72.70	67.35	-
		ACE	2.67	2.87	-
	CEN-AST	Acc	**89.63**	**82.30**	+22.19%
		ACE	**13.80**	**9.06**	+215.68%
	FFN-AST	Acc	81.99	76.53	+13.63%
		ACE	4.90	4.43	+54.36%
	RRN-AST	Acc	79.63	75.24	+11.71%
		ACE	4.95	5.04	+75.61%
Mistral-7B	Baseline	Acc	63.95	60.12	-
		ACE	2.36	**5.45**	-
	CEN-AST	Acc	**92.45**	**71.62**	+19.13%
		ACE	**13.40**	3.89	-28.62%
	FFN-AST	Acc	88.89	69.36	+15.37%
		ACE	10.56	3.12	-42.75%
	RRN-AST	Acc	84.08	55.67	-7.40%
		ACE	7.89	1.48	-72.84%
Llama-8B	Baseline	Acc	58.87	57.62	-
		ACE	1.83	1.72	-
	CEN-AST	Acc	**87.41**	**77.00**	+33.63%
		ACE	**13.72**	**13.53**	+686.63%
	FFN-AST	Acc	51.24	51.20	-11.14%
		ACE	1.06	1.06	-38.37%
	RRN-AST	Acc	61.89	56.68	-1.63%
		ACE	0.77	0.61	-64.53%

Table 1: Accuracy and ACE Ratio on the multilingual syllogism benchmark. "multi" is the average across non-English languages. "$\Delta(\%)$" shows the percentage change relative to the baseline on the "multi" average.

This confirms that the relationship between content-laden and abstract activations in Mistral is predominantly linear. The non-linear FFN/RRN strategies, while powerful, overfit on subtle, language-specific non-linearities, causing their generalisation to fail.

Case Study 2: Llama Llama presents an even more fundamental challenge. While it mirrors Mistral in achieving near-perfect CEN-AST efficiency, it fails catastrophically across all abstractive strategies, both linear and non-linear, even in English ($\eta \approx 50\%$). Crucially, the linear Ridge-AST also fails to find a consistent mapping, achieving $\approx 55\%$ efficiency in English that degrades further in multilingual settings. This suggests that the problem is not one of overfitting to non-linearities, but a more fundamental issue of learnability: the geometric relationship between a concrete syllogism and its abstract logical form in Llama is noisy and cannot be represented in a stable, learnable way. The hypothesis is also supported by the baseline model's poor performance on the abstract dataset.

Model	Strategy	η (en %)	η (multi %)
Mistral-7B	FFN/RRN	91.7	48.1
	Ridge-AST	81.77	76.41
Llama-3.1-8B	FFN/RRN	49.2	45.3
	Ridge-AST	54.56	48.34

Table 2: Diagnostic Steering Efficiency (η) for Abstractive Strategies. Efficiency is averaged across non-English languages for the multilingual setting.

4.4 Probe Gating and Steering Efficiency

To provide a clearer picture of the cross-lingual generalisation bottleneck, we first report the accuracy of the probe gating mechanism in Table 3. The probe is a linear classifier trained exclusively on English syllogism activations that learns a decision boundary in the activation space to separate activations from valid and invalid syllogisms. Its performance on non-English languages, therefore, measures how consistently the activations from different languages fall on the correct side of this English-derived boundary.

Model	Probe Acc (en %)	Probe Acc (multi %)
Qwen-2.5-7B	97.34	87.08
Gemma-2-9B	97.37	91.67
Mistral-7B	93.20	80.62
Llama-3.1-8B	91.69	82.66

Table 3: Probe gating accuracy for each model. The probe was trained only on English data. "multi" is the average accuracy across the nine non-English languages.

The results confirm that the probe's ability to generalise is a critical factor. While most models achieve high probe accuracy on English, there is a noticeable degradation in the multilingual setting. Gemma-2 shows the most robust cross-lingual alignment, with only a modest drop in accuracy. In contrast, models such as Mistral experience a significant decline, which directly contributes to the weaker multilingual performance reported earlier. This highlights the importance of the efficiency metric (η), which decouples the intrinsic effectiveness of a steering strategy from the performance of the probe gating. Since an incorrect probe prediction forces the application of the wrong steering vector, η provides a clearer view of a strategy's upper-bound performance by measuring accuracy only on the subset of instances where the probe was correct.

Model	Strategy	η (en %)	η (multi %)
Qwen-2.5-7B	CEN-AST	**100.00**	98.26
	FFN-AST	98.06	**98.52**
	RRN-AST	98.42	98.40
Gemma-2-9B	CEN-AST	**100.00**	**97.25**
	FFN-AST	89.07	90.10
	RRN-AST	86.54	95.38
Mistral-7B	CEN-AST	**100.00**	**96.91**
	FFN-AST	91.68	48.14
	RRN-AST	90.91	37.43
Llama-3.1-8B	CEN-AST	**100.00**	**100.00**
	FFN-AST	47.22	45.56
	RRN-AST	36.25	29.81

Table 4: Steering Efficiency (η) for models and strategies. η (multi) is the multilingual average.

The efficiency results in Table 4 reinforce our earlier diagnoses. For both Mistral and Llama, CEN-AST achieves (near) perfect efficiency across all languages, confirming that a simple geometric shift is highly effective when the correct validity class is known. The catastrophic failure of FFN-AST and RRN-AST on these same models, even in multilingual settings where the probe is correct, demonstrates that the issue lies with the learnability and generalisability of the abstractive transformations themselves, not just the gating mechanism.

4.5 Robustness to Linguistic Variation

To examine whether our methods capture rigid prompt-derived patterns instead of the underlying logical structure, we constructed a sub-test set of paraphrased syllogisms that preserve the logical form but are rephrased. For instance, an item is transformed as follows:

Premise 1: All celestial bodies are stars.
→ *Every single celestial body is a star.*

Premise 2: Some planets are celestial bodies.
→ *There are some planets which are celestial bodies.*

Conclusion: No planets are stars.
→ *Planets are in no way stars.*

Validity: `false`, **Plausibility:** `true`

This set allows us to test the generalisation of each strategy beyond surface wording and determine whether the models truly capture logical structures.

	Original		Paraphrased	
Strategy	**Acc (%)**	**ACE**	**Acc (%)**	**ACE**
Qwen-2.5-7B				
Baseline	78.06	2.38	68.12	2.17
CEN-AST	**97.30**	**59.58**	**88.65**	15.12
FFN-AST	96.87	35.84	88.12	**34.50**
RRN-AST	97.16	42.46	88.44	17.98
Gemma-2-9B				
Baseline	72.70	2.67	67.63	2.57
CEN-AST	**89.63**	**13.80**	**83.63**	**12.52**
FFN-AST	81.99	4.90	77.52	4.92
RRN-AST	79.63	4.95	75.13	5.10
Mistral-7B				
Baseline	63.95	2.36	63.44	2.40
CEN-AST	**92.45**	**13.40**	**86.88**	**13.42**
FFN-AST	88.89	10.56	85.94	11.09
RRN-AST	84.08	7.89	79.78	6.58
Llama-3.1-8B				
Baseline	58.87	1.83	58.87	1.89
CEN-AST	**87.41**	**13.72**	**83.12**	**5.22**
FFN-AST	51.24	1.06	51.35	1.07
RRN-AST	61.89	0.77	59.90	0.76

Table 5: Accuracy (%) and ACE Ratio on the English original and paraphrased test sets.

Results on the paraphrased test set (Table 5) reveal that all steering strategies maintain their effectiveness, with ACE Ratios remaining remarkably stable, suggesting that the interventions have captured a true structural understanding of the reasoning task. The few outliers (59.58 → 15.12 and 42.46 → 17.98) are a consequence of the non-linearity and of the ACE Ratio (e.g., $\Delta CE = \frac{97.30}{59.58} - \frac{88.65}{15.12} = -4.23$, an existing but marginal deviation).

4.6 Side Effects on Fluency

A critical concern is whether activation steering compromises a model's language-modelling capabilities. The risk is that English-trained assets may force the model to "reason in English", degrading its fluency in other languages. To assess this potential side effect, we conducted a stress test by measuring perplexity (PPL) on a held-out set of general-domain Wikipedia articles.

This experimental design simulates a real-world scenario where the syllogism-specific steering mechanism is active while the model processes unrelated text. The activations from Wikipedia articles are fed to the pipeline (steered towards the pre-calculated centroids or used to predict an abstract activation for blending strategies) twice: once towards validity and once towards invalidity.

Model	**Strategy**	**en**	**multi**	Δ(%)
Qwen-7B	Baseline	6.04	7.82	-
	CEN-AST	**6.19**	**7.97**	+1.9%
	FFN-AST	7.03	8.74	+11.8%
	RRN-AST	6.74	8.34	+6.7%
Gemma-9B	Baseline	9.94	19.31	-
	CEN-AST	**10.15**	**19.67**	+1.9%
	FFN-AST	11.29	22.13	+14.6%
	RRN-AST	10.51	20.02	+3.7%
Mistral-7B	Baseline	4.79	5.82	-
	CEN-AST	**4.99**	**5.96**	+2.4%
	FFN-AST	6.07	7.45	+28.0%
	RRN-AST	5.10	6.31	+8.4%
Llama-8B	Baseline	6.35	7.59	-
	CEN-AST	**6.56**	**7.73**	+1.8%
	FFN-AST	9.39	10.34	+36.2%
	RRN-AST	7.78	9.02	+18.8%

Table 6: Perplexity (PPL) on out-of-domain Wikipedia text. "multi" is the average PPL across non-English languages. "Δ(%)" shows the percentage degradation relative to the baseline.

The results, summarised in Table 6 and detailed in Table 8, are reassuring. The gains in logical consistency come at a modest cost to fluency.

The simplest and often most effective additive strategy, CEN-AST, has minimal impact. In contrast, the blending strategies, FFN-AST and RRN-AST, exhibit a more pronounced, yet still modest, degradation.

Collectively, these results confirm that activation steering acts as an adjustment to the model's reasoning process rather than an alteration of its core language abilities. The benefits outweigh the minimal side effects on language modelling.

5 Related Work

Syllogistic reasoning in LLMs Syllogistic reasoning has long served as a benchmark for testing the separation of logical reasoning from semantic content (Bertolazzi et al., 2024; Ozeki et al., 2024; Wysocka et al., 2025; Kim et al., 2025). Recent studies demonstrate that LLMs frequently fail to make this distinction: they conflate plausibility with validity, reproduce the content effect, and rarely produce no output when inferences are invalid (Bertolazzi et al., 2024; Kim et al., 2025; Valentino et al., 2025). Bertolazzi et al. (2024) show that chain-of-thought prompting and in-context learning provide limited gains, while supervised fine-tuning mitigates some biases but does not eliminate heuristic behaviour. These findings align with earlier work framing transformers as "soft reasoners" (Clark et al., 2020), capable of emulating reasoning patterns but prone to systematic, human-like errors.

Steering via internal activations An emerging line of work investigates how internal representations can be modified at inference time to guide model behaviour, identifying activation subspaces correlated with a target behaviour and scaling them to suppress or enhance its influence (Stoehr et al., 2024; Soo et al., 2025; Valentino et al., 2025). This approach has been applied to mitigate hallucinations, reduce harmful content, and improve factuality, without retraining or additional supervision. Extensions include contrastive steering and layer-wise modulation to stabilise reasoning trajectories. However, such methods have been studied exclusively in monolingual settings so far, leaving their applicability to multilingual reasoning unexplored.

Multilingual reasoning Research on multilingual reasoning has highlighted challenges in ensuring consistency and robustness across languages. Recent efforts include the construction of multilingual and multimodal data and systems for logical and mathematical reasoning (Ranaldi and Pucci, 2025; Ranaldi et al., 2025a), as well as strategies aimed at disentangling reasoning from language-specific features (Ranaldi et al., 2025c; Deng et al., 2025). Parallel work investigates cross-lingual consistency mechanisms and multilingual prompting to improve robustness (Qin et al., 2023; Ranaldi et al., 2024; Pucci and Ranaldi, 2025; Ranaldi et al., 2025b). Despite these advances, none of these studies consider internal activations as a mechanism for multilingual landscapes.

Our contribution We extend activation steering beyond English and provide the first systematic study of its effectiveness in multilingual syllogistic reasoning. We show that steering interventions, initially devised for English, generalise across multiple languages, suppressing content effects while maintaining fluency and prompt-variation stability. This positions steering, along with transparent-by-design architectures for LLMs (Zanzotto et al., 2025), as a scalable, cross-lingual intervention for controlled reasoning.

6 Conclusion

Activation steering is an effective method for enhancing the formal validity of syllogistic reasoning in LLMs. A key finding is the cross-lingual abstraction of the interventions. The additive strategy, CEN-AST, proved to be a robust baseline. Blending strategies are sound, but their efficacy depends on the geometric properties of a model's representation space. This positions activation steering as a promising technique for developing more reliable and controllable soft reasoners.

Limitations

Dependence on Pre-Existing Reasoning Abilities of LLMs The proposed methods are corrective interventions; hence, success relies on the base model possessing an initial, geometrically robust representation of logical validity. Our methods cannot create this structure from scratch. Consequently, these techniques are best understood as amplifiers of a latent reasoning ability.

Cross-Lingual Generalisation Gap of the Probe The entire framework, particularly its probe gating mechanism, is based on a validity probe trained on English activations. While our results show a reasonable level of generalisation, their accuracy on non-English languages is not perfect and represents a bottleneck. The final accuracy of our method is therefore capped by the cross-lingual transferability of this gating mechanism. This is the main reason we proposed a Steering Efficiency metric.

Model Scale and Representational Coherence Because of hardware limitations, our experiments were conducted on models in the 7-9B parameter range. The observed failures of the blending strategies on specific architectures may be a symptom of this relatively small scale. Verifying this remains a compelling avenue for future research.

Ethics Statement

Our research focuses on the abstract reasoning capabilities of LLMs, using a controlled dataset of syllogisms that contains no identifiable or sensitive information. We believe that improving the logical consistency of LLMs and mitigating cognitive biases, such as the content effect, is a crucial and ethically positive step towards developing more reliable and trustworthy systems.

Acknowledgements

The authors thank Almawave S.p.A. for the scientific guidance and technical resources that enabled this study on multilingual syllogistic reasoning in LLMs.

References

AI@Meta. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Almawave. 2024. Velvet-2b: A 2b-parameter open large language model. `https://huggingface.co/Almawave/Velvet-2B`. Accessed: 2025-09-06.

Leonardo Bertolazzi, Albert Gatt, and Raffaella Bernardi. 2024. A systematic analysis of large language models as soft reasoners: The case of syllogistic inferences. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13882–13905, Miami, Florida, USA. Association for Computational Linguistics.

Peter Clark, Oyvind Tafjord, and Kyle Richardson. 2020. Transformers as soft reasoners over language. *Preprint*, arXiv:2002.05867.

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 amp; Cognition*, 11(3):295–306.

Gemma Team and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

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.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report. *Preprint*, arXiv:2506.04079.

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.

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, 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. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Leonardo Ranaldi and Andre Freitas. 2024. 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 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, Federico Ranaldi, and Giulia Pucci. 2025a. R2-MultiOmnia: Leading multilingual multimodal reasoning via self-training. In *Proceedings* *of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8220–8234, Vienna, Austria. Association for Computational Linguistics.

Leonardo Ranaldi, Federico Ranaldi, Fabio Massimo Zanzotto, Barry Haddow, and Alexandra Birch. 2025b. Improving multilingual retrieval-augmented language models through dialectic reasoning argumentations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9064–9085, Suzhou, China. Association for Computational Linguistics.

Leonardo Ranaldi, Marco Valentino, and Andre Freitas. 2025c. 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 Rimsky, 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.

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 Stoehr, 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. 2025. Mitigating content effects on reasoning in language models through fine-grained activation steering. *ArXiv*, abs/2505.12189.

Magdalena Wysocka, Danilo Carvalho, Oskar Wysocki, Marco Valentino, and Andre Freitas. 2025. SylloBio-NLI: Evaluating large language models on biomedical syllogistic reasoning. 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 7235–7258, Albuquerque, New Mexico. Association for Computational Linguistics.

Fabio Massimo Zanzotto, Elena Sofia Ruzzetti, Giancarlo A. Xompero, 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

A.1 Experimental Hyperparameters

This section provides the specific hyperparameter values used for the steering interventions described in the main paper.

A.1.1 Hardware Resources

All experiments, including model activation extraction, probe training, and steering validation runs, were conducted on a local **NVIDIA GeForce RTX 3080 Ti (12 GB VRAM)**. For the largest models and the most extensive multilingual evaluation runs, cloud resources, including an **NVIDIA A40 GPU (48 GB VRAM)**, were employed. All models were processed using `bfloat16` precision, although steering effectiveness remained robust under 4-bit and 8-bit quantisation.

A.1.2 Target Layers

Our interventions were applied simultaneously across a contiguous block of layers for each model for a more stable and distributed effect. The specific layer sets, identified via preliminary probing analysis to identify where logical validity information is most distinctly represented, are as follows:

- **Qwen 2.5 7B:** Layers 17, 18, and 19.
- **Gemma 2 9B:** Layers 25 through 30.
- **Mistral 7B:** Layers 11 through 17.
- **Llama 3.1 8B:** Layers 12 through 16.
- **Velvet 2B:** Layers 14 through 16.
- **EuroLLM 9B:** Layers 21 through 25.

These layer blocks typically reside in the second or third quarter of the network.

The steering vector for each layer is computed independently, and the intervention is applied at each specified layer during the forward pass.

A.1.3 Steering Coefficient (α)

The steering strength was fixed based on performance on a validation set. The values for each strategy family were:

- **CEN-AST (Additive):** $\alpha = 1.0$
- **FFN-AST & RRN-AST (Blending):** $\alpha = 0.85$

For CEN-AST, $\alpha = 1.0$ is a solid choice since it results in a shift towards the target centroid. For FFN-AST and RRN-AST, a lower value of $\alpha = 0.85$ is a safer choice since $\alpha = 1.0$ would result in a complete replacement of the last token activation, potentially losing positional information.

A.2 Probes Training

To dynamically gate the steering direction at inference time (i.e., selecting between valid and invalid target representations), we trained lightweight diagnostic probes for each fold. We employed a **Logistic Regression** classifier with an L_2 penalty (regularisation strength $C = 1.0$) and the LBFGS solver, configured with a maximum of 100 iterations.

The input features for the probe consist of the model's internal activations extracted from the target steering layers. For a given input sequence, we compute the mean-pooled activation over the last token for each layer in the steering block. These layer-wise vectors are then concatenated into a single feature vector, $v_{in} \in \mathbb{R}^{L \times d}$, where L is the number of steering layers and d is the model's hidden dimension.

A.3 Regressors Implementation Details

This section details the architectures and training objectives for the mapping networks used to predict abstract representations from content-laden inputs. All regressors were trained using the Adam optimizer with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}, employing a `ReduceLROnPlateau` scheduler (factor 0.8, patience 15).

A.3.1 FFN-AST (Two-Headed MLP)

The Feed-Forward Network (FFN) Abstractor is designed to decouple the geometric orientation of the representation from its norm. The architecture comprises:

- **Shared Backbone:** A two-layer MLP with a hidden dimension of 512. Each layer is followed by Layer Normalisation, a ReLU activation, and Dropout ($p = 0.1$) to prevent overfitting on the limited training pairs.
- **Direction Head:** A projection layer that outputs a normalised unit vector, trained to maximise cosine similarity with the target abstract activation.
- **Magnitude Head:** A scalar regression layer culminating in a Softplus activation, trained via Mean Squared Error (MSE) to match the norm of the target activation.

The total loss is a weighted sum of the cosine distance loss, magnitude MSE, and a transformation norm penalty to encourage minimal necessary intervention.

A.3.2 RRN-AST (Rotational-Residual Network)

The Rotational-Residual Network (RRN) operates on a lower-dimensional manifold to learn a robust geometric transformation.

- **Dimensionality Reduction:** Inputs are first projected into a latent subspace of dimension $d_{latent} = 128$ using Principal Component Analysis (PCA) fitted on the training fold's content-laden activations.
- **Transformation Logic:** A shared feature extractor (hidden dimension 256) predicts two components: a rotation matrix $R \in \mathbb{R}^{d_{latent} \times d_{latent}}$ and a residual translation vector.
- **Orthogonalisation:** To ensure the rotation matrix preserves the geometric structure of the latent space, we apply Gram-Schmidt orthogonalisation (via QR decomposition) to the predicted matrices during the forward pass.

The training objective combines reconstruction MSE with regularisation terms enforcing orthogonality on R and sparsity on the residual vectors.

A.3.3 Linear Abstractor (Ridge Baseline)

For the linear baseline comparisons, we utilised Ridge Regression (Linear Least Squares with L_2 regularisation). Independent regressors were trained for each steering layer to map content-laden activation vectors directly to their abstract counterparts, with the regularisation strength set to $\alpha = 1.0$ and intercept fitting enabled.

B Detailed Experimental Results

Qwen-2.5-7B	**en**	**bn**	**de**	**es**	**fr**	**it**	**ru**	**sw**	**te**	**zh**	**Avg.**
Baseline Acc	78.06	63.92	64.42	64.83	67.00	64.33	62.42	62.92	71.17	66.42	66.45
Baseline ACE	2.38	2.39	1.81	1.83	1.97	1.87	1.78	1.81	6.11	2.02	2.40
Additive											
CEN-AST Acc	97.30	78.02	89.58	91.15	90.07	88.75	83.79	67.78	72.92	88.79	84.82 (+27.62%)
CEN-AST ACE	59.58	6.77	29.34	23.07	19.61	13.65	15.12	7.29	10.13	17.96	20.25 (+743.75%)
Blending											
FFN-AST Acc	96.87	78.33	88.17	90.08	87.92	88.83	81.42	75.83	76.75	86.92	85.11 (+28.08%)
FFN-AST ACE	35.84	10.01	22.71	24.99	12.67	17.29	10.53	9.52	6.21	17.29	16.71 (+596.25%)
RRN-AST Acc	97.16	77.64	88.85	90.56	89.76	90.69	83.51	70.28	73.51	88.09	85.00 (+27.92%)
RRN-AST ACE	42.46	14.36	21.54	24.11	19.83	19.42	15.29	15.00	4.71	15.47	19.22 (+700.83%)
Gemma-2-9B	**en**	**bn**	**de**	**es**	**fr**	**it**	**ru**	**sw**	**te**	**zh**	**Avg.**
Baseline Acc	72.70	66.50	67.30	71.79	69.28	69.69	67.65	63.94	64.25	65.79	67.63
Baseline ACE	2.67	4.29	2.45	2.99	2.78	2.88	2.75	3.11	2.25	2.37	2.85
Additive											
CEN-AST Acc	89.63	72.70	89.58	91.73	90.56	91.34	83.85	66.99	72.70	81.25	83.63 (+23.66%)
CEN-AST ACE	13.80	10.21	10.35	12.77	10.21	12.20	7.58	2.93	10.21	4.07	9.88 (+246.67%)
Blending											
FFN-AST Acc	81.99	76.75	79.58	80.50	82.00	80.42	75.83	72.42	65.50	75.75	77.52 (+14.62%)
FFN-AST ACE	4.90	5.24	4.26	4.75	6.59	5.12	4.05	4.03	2.26	3.60	4.16 (+46.06%)
RRN-AST Acc	79.63	75.67	77.33	79.42	79.67	78.67	75.42	70.17	66.42	74.42	75.13 (+11.08%)
RRN-AST ACE	4.95	7.56	4.45	5.59	6.47	5.71	4.43	4.26	2.55	4.31	5.02 (+76.14%)
Mistral 7B	**en**	**bn**	**de**	**es**	**fr**	**it**	**ru**	**sw**	**te**	**zh**	**Avg.**
Baseline Acc	63.95	61.47	61.29	59.99	63.46	60.49	58.75	56.22	58.58	58.86	60.31
Baseline ACE	2.36	2.53	2.39	2.32	2.29	2.86	2.33	9.89	19.30	2.55	4.91
Additive											
CEN-AST Acc	92.45	57.23	81.45	85.16	85.74	74.22	73.83	63.09	49.61	74.22	75.81 (+25.71%)
CEN-AST ACE	13.40	1.29	4.54	8.51	9.20	3.67	2.64	1.72	0.99	2.48	7.54 (+53.56%)
Blending											
FFN-AST Acc	88.89	57.25	80.25	73.50	81.50	71.25	70.25	66.00	50.00	74.25	72.85 (+20.79%)
FFN-AST ACE	10.56	1.30	3.65	5.07	6.79	3.39	2.55	1.94	1.00	2.43	4.52 (-7.94%)
RRN-AST Acc	84.08	54.95	65.63	50.77	60.79	48.77	45.31	70.55	57.07	47.16	61.35 (+1.72%)
RRN-AST ACE	7.89	1.07	2.25	1.81	1.81	1.06	1.06	2.06	1.07	1.17	2.45 (-50.09%)
Llama-3.1-8B	**en**	**bn**	**de**	**es**	**fr**	**it**	**ru**	**sw**	**te**	**zh**	**Avg.**
Baseline Acc	58.87	58.54	57.42	54.10	58.20	60.55	59.57	52.54	63.30	54.88	58.87
Baseline ACE	1.83	0.23	2.22	1.69	1.66	2.32	3.10	2.30	0.31	1.67	1.73
Additive											
CEN-AST Acc	87.41	69.92	85.94	84.96	88.48	83.40	82.42	65.62	73.63	81.64	80.34 (+36.47%)
CEN-AST ACE	13.72	5.77	10.10	12.26	13.27	4.72	17.03	6.80	45.04	6.79	13.55 (+683.24%)
Blending											
FFN-AST Acc	51.24	51.24	51.08	50.92	50.50	50.58	51.42	51.33	51.24	51.50	51.24 (-11.31%)
FFN-AST ACE	1.06	1.06	1.05	1.03	1.02	1.04	1.10	1.08	1.06	1.09	1.06 (-38.73%)
RRN-AST Acc	61.89	57.14	56.60	53.31	51.25	54.68	48.85	50.55	85.71	52.02	57.10 (-9.37%)
RRN-AST ACE	0.77	0.30	0.73	0.84	0.91	0.73	0.64	0.68	0.05	0.57	0.67 (-61.27%)
Velvet-2B	**en**	**bn**	**de**	**es**	**fr**	**it**	**ru**	**sw**	**te**	**zh**	**Avg.**
Baseline Acc	68.33	53.50	48.48	52.08	57.35	62.60	52.88	50.31	52.00	53.08	55.06
Baseline ACE	3.86	2.10	2.65	2.50	3.01	2.59	1.67	1.94	1.80	1.34	2.35
Additive											
CEN-AST Acc	81.66	63.15	65.23	63.60	75.28	71.97	59.67	50.20	61.40	62.90	65.50 (+18.96%)
CEN-AST ACE	24.29	4.75	11.89	3.51	5.75	2.88	1.67	2.37	4.10	9.25	7.05 (+200.00%)
Blending											
FFN-AST Acc	76.34	56.20	56.00	57.50	66.00	58.25	54.62	51.88	54.65	52.12	58.36 (+5.99%)
FFN-AST ACE	9.45	2.30	1.79	2.34	7.39	1.43	1.61	2.28	1.95	1.12	3.17 (+34.89%)
RRN-AST Acc	79.73	56.00	52.60	60.19	66.58	68.60	50.71	48.68	54.45	47.87	58.54 (+6.32%)
RRN-AST ACE	8.50	3.35	3.41	7.64	1.03	7.30	2.17	2.01	2.90	2.66	4.10 (+74.47%)
EuroLLM-9B	**en**	**bn**	**de**	**es**	**fr**	**it**	**ru**	**sw**	**te**	**zh**	**Avg.**
Baseline Acc	67.57	58.00	57.34	60.20	62.90	62.27	59.88	51.99	57.50	57.35	59.50
Baseline ACE	3.26	2.20	2.30	2.55	2.89	2.46	1.69	1.53	2.10	2.41	2.34
Additive											
CEN-AST Acc	79.62	62.40	62.53	61.97	60.93	56.36	63.04	60.48	61.85	70.74	63.99 (+7.55%)
CEN-AST ACE	7.98	3.45	2.51	2.98	2.54	1.73	2.71	2.40	3.30	6.67	3.63 (+55.13%)
Blending											
FFN-AST Acc	68.46	58.60	58.87	62.60	60.30	62.20	62.38	50.61	58.10	58.99	60.11 (+1.03%)
FFN-AST ACE	9.28	5.50	3.55	4.99	2.58	2.68	12.78	7.35	5.25	4.03	5.80 (+147.86%)
RRN-AST Acc	43.97	48.20	46.59	52.06	49.51	53.50	52.96	51.62	47.80	48.12	49.43 (-16.92%)
RRN-AST ACE	2.31	5.10	2.25	12.25	7.29	4.63	10.50	1.43	4.85	3.08	5.37 (+129.49%)

Table 7: Overall Accuracy and ACE Ratio on the multilingual syllogism benchmark.

C Detailed Side Effects on Fluency

Qwen-2.5-7B	**en**	**de**	**es**	**fr**	**it**	**ru**	**zh**	**Avg.**
Baseline	6.04	5.12	7.54	6.99	7.11	4.84	15.34	7.57
Additive								
CEN-AST	6.19	5.22	7.68	7.14	7.22	4.97	15.61	7.72 (+1.98%)
Blending								
FFN-AST	7.03	5.54	8.35	7.68	7.75	5.22	17.90	8.50 (+12.29%)
RRN-AST	6.74	5.34	8.00	7.37	7.46	5.04	16.84	8.11 (+7.13%)
Gemma-2-9B	**en**	**de**	**es**	**fr**	**it**	**ru**	**zh**	**Avg.**
Baseline	9.94	9.24	14.30	13.88	14.18	11.63	52.69	18.00
Additive								
CEN-AST	10.15	9.45	14.65	14.24	14.54	11.95	53.76	18.39 (+2.17%)
Blending								
FFN-AST	11.29	10.54	17.62	16.30	16.72	13.96	62.52	21.28 (+18.22%)
RRN-AST	10.51	9.87	16.65	14.77	15.17	12.60	56.26	19.40 (+7.78%)
Mistral-7B	**en**	**de**	**es**	**fr**	**it**	**ru**	**zh**	**Avg.**
Baseline	4.79	3.92	5.95	5.34	5.51	4.11	10.19	5.69
Additive								
CEN-AST	4.99	4.04	6.08	5.44	5.69	4.22	10.39	5.84 (+2.64%)
Blending								
FFN-AST	6.07	4.86	7.67	6.82	7.14	5.26	12.99	7.26 (+27.59%)
RRN-AST	5.10	4.24	6.46	5.95	5.90	4.40	10.91	6.14 (+7.91%)
Llama-3.1-8B	**en**	**de**	**es**	**fr**	**it**	**ru**	**zh**	**Avg.**
Baseline	6.35	5.24	6.89	6.70	6.35	5.46	15.38	7.48
Additive								
CEN-AST	6.56	5.41	7.01	6.83	6.54	5.69	15.75	7.68 (+2.67%)
Blending								
FFN-AST	9.39	7.85	9.42	8.79	8.25	7.30	20.43	10.20 (+36.36%)
RRN-AST	7.78	6.58	8.18	7.87	7.25	6.32	18.23	8.89 (+18.85%)
Velvet-2B	**en**	**de**	**es**	**fr**	**it**	**ru**	**zh**	**Avg.**
Baseline	18.35	16.90	127.57	34.90	15.51	4.60	12.20	32.86
Additive								
CEN-AST	18.72	17.24	130.12	35.60	15.82	4.69	12.44	33.52 (+2.01%)
Blending								
FFN-AST	19.67	18.74	143.64	39.22	17.60	5.17	14.51	36.94 (+12.42%)
RRN-AST	15.96	18.00	147.11	40.15	15.65	6.91	21.51	37.90 (+4.78%)
EuroLLM-9B	**en**	**de**	**es**	**fr**	**it**	**ru**	**zh**	**Avg.**
Baseline	4.66	3.26	6.08	5.52	5.61	4.78	9.48	5.63
Additive								
CEN-AST	4.78	3.34	6.23	5.66	5.75	4.90	9.72	5.77 (+2.49%)
Blending								
FFN-AST	5.17	3.54	6.70	6.03	6.14	5.17	10.47	6.17 (+9.59%)
RRN-AST	5.03	3.52	6.57	5.96	6.06	5.16	10.24	6.08 (+7.99%)

Table 8: Perplexity (PPL) on the multilingual syllogism benchmark. Steered PPL values are reported as the mean perplexity when steering towards both validity and invalidity (which are always very similar).

D Detailed Steering Efficiency

Model	Strategy	η (en %)	η (multi %)
Qwen-2.5-7B	CEN-AST	100	98.26
	FFN-AST	98.06	98.52
	RRN-AST	98.42	98.40
Gemma-2-9B	CEN-AST	100	97.25
	FFN-AST	89.07	90.10
	RRN-AST	86.54	95.38
Mistral 7B	CEN-AST	100	96.91
	FFN-AST	91.68	48.14
	RRN-AST	90.91	37.43
	Ridge-AST	81.77	76.41
Llama-3.1-8B	CEN-AST	100	100
	FFN-AST	47.22	45.56
	RRN-AST	36.25	29.81
	Ridge-AST	54.56	48.34
EuroLLM-9B	CEN-AST	84.39	81.94
	FFN-AST	70.55	66.18
	RRN-AST	38.46	43.18
	Ridge-AST	44.12	66.66
Velvet-2B	CEN-AST	100.00	99.58
	FFN-AST	89.63	70.46
	RRN-AST	80.01	80.50

Table 9: Summary of Steering Efficiency (η) Metrics. η (en) is the average efficiency across English (EN) examples and their paraphrased versions, and η (multi) is the average across the other languages.

E Models Versions

Model	Version
Qwen2.5-7B	Qwen/Qwen2.5-7B-Instruct
Gemma-2-9B	google/gemma-2-9b-it
Mistral-7B	mistralai/Mistral-7B-Instruct-v0.3
Llama-3.1-8B	meta-llama/Llama-3.1-8B-Instruct
EuroLLM-9B	utter-project/EuroLLM-9B-Instruct
Velvet-2B	Almawave/Velvet-2B

Table 10: Models proposed in this work, which can be found on huggingface.co. We used the standard configurations recommended in model cards of HuggingFace repositories *(access to the following models was verified on 5.10.2025).