A Declarative–Procedural Perspective on Expert Routing in Bilingual Mixture-of-Experts Language Models

Anonymous ACL submission

Abstract

We investigate whether Mixture-of-Experts (MoE) language models develop linguistically structured expert routing during bilingual language acquisition. Inspired by the Declarative–Procedural framework, we analyze lexical, grammatical, and syntactic processing in a decoder-only English–German MoE Transformer trained under sequential language exposure. We construct a probe-based validation set and extract token-level routing distributions to quantify category-dependent specialisation using mutual information, routing entropy, and Jensen–Shannon divergence. The curriculum-trained model exhibits a peak mutual information of 0.1148 at layer 5, indicating category-dependent differences in routing distributions across linguistic categories. Surprisingly, a no-curriculum baseline trained on mixed English–German data shows stronger aggregate specialisation, reaching a peak mutual information of 0.2599 at the same layer. These results suggest that interpretable linguistic organization emerges within MoE routing patterns even without sequential language exposure. Rather than uniformly increasing specialisation, staged bilingual exposure appears to redistribute specialisation across languages, yielding a more balanced bilingual routing profile. Code and data will be released upon acceptance.

1 Introduction

Mixture-of-Experts (MoE) language models route each token across a subset of expert networks instead of sending every token through the same feed-forward network (Shazeer et al., 2017; Fedus et al., 2022). This makes MoE models useful not only for scaling, but also for studying how neural language models choose experts internally. If different kinds of linguistic tokens are routed to different experts, the routing mechanism may exhibit an organized form of expert distribution inside the model.

In this work, we examine whether bilingual MoE language models develop routing patterns that correspond to meaningful linguistic categories. We focus on three categories motivated by the Declarative–Procedural view of language: lexical knowledge, grammatical processing, and syntactic structure (Ullman, 2001b,a, 2020). In this paper, a probe is a token chosen from a sentence to be examined. Lexical probes are tokens that test word knowledge, such as irregular forms. Grammatical probes are tokens that test rule-based morphology and agreement. Syntactic probes are tokens that participate in clause structure and sentence organization. Instead of examining only the model's final predictions, we directly analyse the expert-routing distributions associated with these probe tokens.

We train a sparse bilingual transformer using a staged English–German curriculum. The training curriculum starts with English and gradually transitions to German, approximating structured second-language exposure and curriculum learning in bilingual settings (Bengio et al., 2009; Platanios et al., 2019; Zhang et al., 2019). We compare this structured curriculum to an unstructured no-curriculum setting in which English and German are introduced without gradual cumulative exposure. This lets us examine whether MoE routing becomes organized by language and whether staged bilingual exposure affects how that organization is distributed across languages.

To measure the development of an organized routing schema, we use a held-out validation set with lexical, grammatical, and syntactic probes in both English and German. We examine the router's probability distribution over experts at each MoE layer for each probe token. We then measure the relationship between expert routing and linguistic category using mutual information, entropy, Jensen–Shannon divergence, and sentence-level permutation testing (Shannon, 1948; Lin, 1991; Cover and Thomas, 2006). This complements recent work that studies behavioral declarative and procedural knowledge in datasets and large language models

(Li et al., 2024) by moving the analysis to the level of internal routing.

Our results show that MoE routing distributions contain measurable information about linguistic category membership: the category of a probe token can be partially inferred from its expert probability distribution because routing and category exhibit non-trivial mutual information. Lexical, grammatical, and syntactic probes are routed differently. The clearest specialisation appears in intermediate MoE layers, whereas later layers route more diffusely and show reduced mutual information. We also find that curriculum affects how specialisation is balanced between languages. The no-curriculum model develops a stronger English routing signature, while the staged L1–L2 curriculum yields a more balanced bilingual organization and supports clearer German expert structure in the middle layers.

In general, this study shows that bilingual MoE routers develop a language-sensitive quantifiable expert-routing structure that reflects linguistically significant differences. These findings suggest that expert routing provides a useful mechanistic window into how bilingual language models allocate computation across different types of language processing.

The central motivation of this paper is to move beyond asking whether a bilingual MoE model performs well on a diagnostic task and instead ask how its internal routing mechanism organizes linguistic computation. Our explicit contribution is a routing-level analysis framework for testing whether expert allocation is sensitive to lexical, grammatical, and syntactic probe categories under staged bilingual exposure. By comparing a forward English–German curriculum with a no-curriculum baseline, we isolate how training order changes the distribution of routing specialisation across languages. This makes the study a mechanistic investigation of expert routing in bilingual MoE models, rather than a general benchmark of language-model accuracy.

2 Related Work

This section places the study in the context of four connected areas: the Declarative–Procedural view of language, the linguistic analysis of transformer models, expert routing in MoE architectures, and curriculum learning for bilingual exposure. Together, these areas lead to the main question of the paper: does expert routing in a bilingual MoE model become sensitive to lexical, grammatical, and syntactic differences? Also, does staged language exposure change that routing structure?

Declarative–Procedural theory. The Declarative–Procedural framework divides language knowledge into two main parts (Ullman, 2001b,a, 2020). Declarative knowledge relates to stored word knowledge, such as vocabulary and irregular forms. Procedural knowledge relates to rule-based grammar and compositional structure. In this paper, we use this framework only as a linguistic lens. We do not claim that MoE routers are the same as human memory systems.

Declarative and procedural knowledge in LLMs. Recent studies have looked at declarative and procedural knowledge in language models mainly through model outputs and task performance (Li et al., 2024). Our work explores this question within the model. Instead of just asking if the model provides the right answer, we investigate whether a sparse bilingual model allocates computation differently for varying types of probe tokens.

Transformer linguistic structure. Previous studies indicate that transformer layers do not all encode the same type of linguistic information (Vaswani et al., 2017; Tenney et al., 2019; Clark et al., 2019; Elhage et al., 2021). Intermediate layers often have clearer syntactic and semantic structures than very early or very late layers. This motivates our layer-wise analysis of MoE routing. We specifically question whether middle routed layers show stronger category-sensitive expert allocation.

Mixture-of-Experts routing. MoE models use routers to direct tokens to a subset of experts (Shazeer et al., 2017; Fedus et al., 2022). Earlier work has shown that experts can specialize based on token statistics, domains, routing design, and balancing constraints (Dai et al., 2024; Falke et al., 2026; Sun et al., 2026). Our research stands out by focusing on bilingual linguistic probe categories and exploring how curriculum structure impacts the distribution of routing specialization across English and German.

Curriculum learning and bilingual exposure. Curriculum learning investigates how the order of training examples influences learning (Bengio et al., 2009; Platanios et al., 2019; Zhang et al., 2019). In bilingual training, staged exposure is useful be-

cause it allows us to compare two scenarios: one where the model sees English first and German later, and another where both languages are mixed from the start. This comparison helps us assess whether the order of language exposure affects how routing specialization develops across languages.

3 Routing Analysis Framework

Let E denote the expert-routing random variable, C denote the linguistic category random variable, and $r^{(l)}(t)$ denote the routing probability vector of token t at layer l. Analysis is performed on the routed layers $L = \{1, 3, 5, 7\}$.

The category set is defined as

$$C = \{\textit{lexical}, \textit{grammatical}, \textit{syntactic}\}$$

Probing mechanism. Probe tokens are annotated in each sentence using a parser (details in Methodology and Appendix B). The routing vector corresponding to the probe token is extracted from the router during inference and used to compute the routing statistics described below.

Analysis is performed independently for each routed layer. All routing distributions and statistical measures are therefore computed separately for each layer.

To analyse the relationship between the model's routing behaviour and different categories, we frame the null hypothesis as

$$H_0 : P(E|C) = P(E)$$

meaning that the expert routing distribution is independent of linguistic category.

The alternate hypothesis is

$$H_1 : P(E|C) \neq P(E)$$

meaning that the expert routing distribution depends on linguistic category.

3.0.1 Statistical Testing Framework

The primary statistical tool used is mutual information between expert-routing distributions and linguistic category. Higher mutual information indicates a stronger association between the router's expert probability mass and linguistic category.

The marginal routing probability assigned to expert e is computed as

$$p(e) = \frac{\sum_{\tau \in T} r^{(l)}(\tau)[e]}{\sum_{e'} \sum_{\tau \in T} r^{(l)}(\tau)[e']} \qquad (1)$$

and the joint probability between expert e and category c is computed as

$$p(e, c) = \frac{\sum_{\tau \in T_c} r^{(l)}(\tau)[e]}{\sum_{c'} \sum_{e'} \sum_{\tau \in T_{c'}} r^{(l)}(\tau)[e']} \qquad (2)$$

where T denotes the complete set of probe tokens and T_c denotes the subset belonging to category c.

Mutual information is then computed as

$$MI(E; C) = \sum_e \sum_c p(e, c) \log \left(\frac{p(e, c)}{p(e)p(c)} \right) \qquad (3)$$

where $p(e)$ is the marginal routing probability assigned to expert e and $p(c)$ is the marginal probability of the routed token belonging to category c.

Robustness of the mutual information is analysed with a permutation test where the p-level is set to be less than 0.01 in order to eliminate coincidence of mutual information.

We can now frame our hypothesis based on mutual information:

$$H_0 : MI(E; C) = 0$$

$$H_1 : MI(E; C) > 0$$

3.0.2 Entropy Analysis

Entropy is considered as the second statistical tool to analyse how distributed the routing is across experts. Higher entropy indicates a more diffuse expert-usage distribution across probes across experts, whereas lower entropy indicates concentration of probability mass on a smaller subset of experts.

Shannon entropy is given by

$$H(E) = -\sum_e p(e) \ln p(e) \qquad (4)$$

where $p(e)$ is the marginal distribution of expert e being chosen. When entropy is computed for a specific linguistic category, $p(e)$ is estimated only from the probe tokens belonging to that category. Thus, category-wise entropy measures how concentrated or diffuse expert usage is for lexical, grammatical, or syntactic probes separately.

Other tools include Jensen–Shannon Divergence (JSD), which gives a measure of how routing between different linguistic categories is similar or different, which is given by

$$\begin{aligned} &JSD(P(E|C_i) \,||\, P(E|C_j)) \\ &= \frac{1}{2} D_{KL}(P(E|C_i) \,||\, M) \\ &+ \frac{1}{2} D_{KL}(P(E|C_j) \,||\, M) \end{aligned} \tag{5}$$

where

$$M = \frac{1}{2}\left(P(E|C_i) + P(E|C_j)\right).$$

4 Methodology

4.1 Dataset Construction and Probe Annotation

The dataset is constructed from the FineWeb-Edu corpus (Penedo et al., 2024) for the English part and the German portion of mC4 (Xue et al., 2021) for the German part. The pipeline uses the Stanza module to categorise words in each sentence as lexical, grammatical, and syntactic. The taxonomy for this classification is given in Appendix A.

One probe token is annotated for each sentence-category pair. A sentence may contain probe tokens belonging to multiple linguistic categories and can therefore appear in more than one category-specific dataset. In such cases, the sentence is duplicated across the relevant categories, with each category using its corresponding probe token.

The following table provides the size of the English and German datasets:

Language	Lexical	Grammatical	Syntactic
English	460,250	650,000	650,000
German	138,127	175,000	175,000

Table 1: Dataset size by language and linguistic category.

Note: Lexical probe extraction was terminated before reaching the target dataset size because of computational constraints, resulting in smaller lexical subsets for both languages. This does not affect the reported analyses, which are conducted on a fixed held-out validation set.

4.2 Sequential L1–L2 Curriculum

The curriculum is designed to approximate staged second-language acquisition. Here, the first language is English (EN) and the second language is German (DE). The first 1–8 epochs consist of an EN-only curriculum. From epochs 9–16, the curriculum slowly transitions from EN-only to bilingual (EN + DE). The choice of language was based on the differences in grammatical structure and the linguistic category taxonomy (more details in Appendix A).

$$\lambda_s = 0.20 + 0.05(s - 9), \qquad 9 \le s \le 14 \tag{6}$$

where λ_s denotes the proportion of German samples presented during epoch s.

Within each epoch, the amount of lexical, grammatical, and syntactic probe tokens is equal and follows a repetition + new exposure system. This system can be given by

$$\begin{aligned} Cur_s = &\, 0.6 \cdot Cur_{s-1} \\ &+ 0.4 \cdot (New \text{ or } Cur_{s'} \text{ where } s' < s - 1) \end{aligned} \tag{7}$$

for $2 \le s \le 16$.

$$Cur_1 = 1 \cdot New \tag{8}$$

where Cur_s is the curriculum of epoch s and New is the set of samples that are not part of any previous curriculum.

4.3 Diagnostic Held-out Validation

The validation set consists of 12,000 samples (6,000 per language and 2,000 per category per language). The validation set is a minimal-pairs dataset consisting of a pair of sentences, S^+ and S^-, where S^+ is a linguistically valid and acceptable sentence and S^- is a linguistically invalid sentence.

The S^+ set was generated synthetically using OpenAI's GPT-5.1 model in order to obtain controlled, diverse, and task-specific examples. The generation was guided by rigorous prompting (prompts provided in Appendix E), with valid probe tokens injected into the prompts. Automatic post-generation validation is applied to ensure that generated sentences follow the correct category subtype, are not duplicates, and contain the required metadata. The S^- sentence is then deterministically obtained by modifying the probe token to invalidate the sentence in a linguistically meaningful way (rules provided in Appendix E).

4.4 Experimental Setup

A custom 8-layer MoE language model is used, with a feed-forward dimension of 2048, an embedding size of 512, and 8 attention heads per layer. The model uses the multilingual mBERT tokenizer with a vocabulary size of 119,547. The total parameter count of the model is 86.4 million.

A relatively small model size is chosen to facilitate controlled experimentation and clearer analysis of expert routing behaviour. The MoE layers are placed in alternate transformer layers (1, 3, 5, and 7). The primary training setup uses the proposed sequential L1–L2 curriculum, with an additional No-Curriculum setting used as an ablation baseline.

The no-curriculum baseline uses a fixed 80:20 English-German mixture at every epoch. This keeps the overall bilingual composition broadly similar to the forward curriculum, although the aggregate exposure is not exactly identical: the forward curriculum yields 81.56% English and 18.44% German across the full schedule, whereas the no-curriculum condition remains fixed at 80% English and 20% German.

Both conditions receive the same total number of training samples. The difference between them is the order in which the languages are introduced. In the forward curriculum, the model is trained on English first and German is introduced gradually. In the no-curriculum condition, English and German are mixed from the start using the same overall English-German proportions.

The primary runs use top-k routing with $k = 3$ and a router load-balancing coefficient of 0.01. Additional hyperparameters are provided in Appendix C.

5 Results and Analysis

All the results, mathematical analyses, and metrics reported below are obtained through inference on the forward-curriculum model and the no-curriculum ablation model over the held-out validation set. Unless otherwise stated, all MI values reported in Sections 5.1–5.5 are computed on the pooled bilingual validation set combining English and German probes. Language-specific analyses are introduced separately in Section 5.6.

5.1 Behavioural Performance

Table 2 reports validation performance for the forward-curriculum and no-curriculum conditions. The two setups achieve broadly comparable behavioural outcomes on the held-out diagnostic validation set. The Accuracy is computed as token-level next-token accuracy on the grammatical sentence strings only, averaged over all non-padding tokens; it is not a pairwise sentence-choice accuracy between S^+ and S^-. The forward-curriculum model attains lower perplexity (111.19) than the no-curriculum model (117.01), whereas the no-curriculum condition achieves marginally higher accuracy (29.65% versus 28.99%).

Because the two training conditions achieve similar behavioural performance, the routing patterns reported in the following sections are less likely to reflect simple differences in overall model competence.

The relatively low absolute scores reflect the challenging evaluation setup. The purpose of the diagnostic set is to expose routing behaviour under controlled linguistic contrasts rather than to serve as a benchmark of language-model capability. These scores are not the main outcome of the paper. They just indicate that both models can handle the diagnostic examples. The primary analysis focuses on the router. We examine whether different token types are sent to different experts during validation.

Condition	Perplexity	Accuracy (%)
Forward Curriculum	111.19	28.99
No Curriculum	117.01	29.65

Table 2: Validation performance on the held-out diagnostic probe set.

5.2 Emergence of Category-Dependent Routing

The primary analysis investigates whether routing specialisation occurs across different linguistic categories. Mutual Information (MI) is computed from the inference routing logs to quantify this phenomenon. The MI values vary across layers, indicating that the degree of specialisation differs throughout the network. Layer 5 shows the highest mutual information value (0.1148), suggesting the strongest category-dependent routing behaviour. In contrast, the final routed layer (layer 7) shows an MI value substantially lower than the other routed layers.

To determine whether the obtained MI values (routing-category associations) are statistically significant, a permutation test was conducted using 1,000 sentence-level permutations. Across all

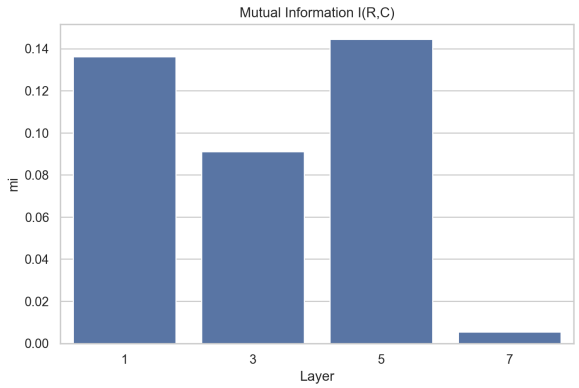

Figure 1: Layer-wise mutual information between expert-routing distributions and linguistic categories.

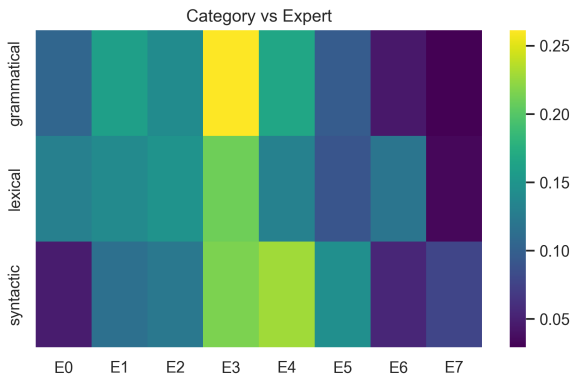

Figure 2: Category-conditioned expert allocation patterns.

routed layers, the observed MI values were significantly higher than the corresponding layer-wise null distributions, with statistical significance of $p < 0.001$ for all layers. This result makes a shuffled-label explanation unlikely and supports the presence of category-dependent routing behaviour.

Layer	Observed MI	Null Mean	Null Std	p-value
1	0.0520	0.0004	0.0002	0.000999
3	0.0384	0.0003	0.0001	0.000999
5	0.1148	0.0003	0.0001	0.000999
7	0.0207	0.0001	0.0000	0.000999

Table 3: Permutation test results for routing-category mutual information. All routed layers exhibit statistically significant routing-category associations. All reported p-values correspond to the minimum attainable empirical value under 1000 permutations, indicating that no permuted sample exceeded the observed MI.

5.3 Expert Allocation Patterns

The secondary analysis supporting the expert specialisation claim is performed through category-wise and layer-wise expert allocation patterns. This analysis also explains how experts have been utilized across different linguistic categories. Although the probability mass is distributed across different experts for each category, certain experts consistently receive higher routing probabilities for specific categories. This indicates the presence of preferential expert allocation rather than uniform routing.

We can also observe that there is no collapse to a single expert for any particular linguistic category. Instead, each category exhibits a distribution of expert preferences, where some experts are selected more frequently than others. For example, grammatical probes show stronger preferences towards Experts E1 and E3, while syntactic probes exhibit higher utilisation of Experts E3–E5. This shows that category-dependent specialisation emerges through differences in routing preference rather than strict expert exclusivity.

Figure 3 presents the layer-wise expert usage distribution. It is interesting to note that the different routed layers exhibit different expert usage distributions, with different subsets of experts dominating at different stages of the processing. This observation is consistent with the layer-wise MI analysis, suggesting that the routing distribution differs across different parts of the model.

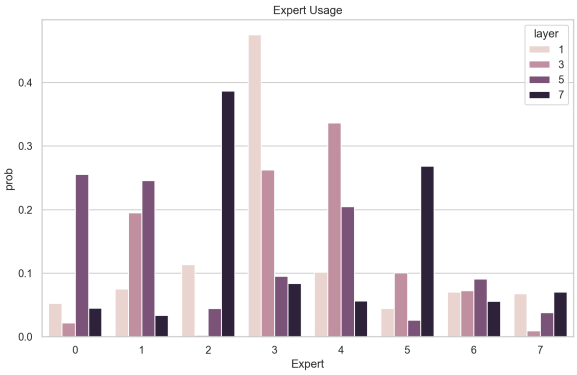

Figure 3: Layer-wise expert utilisation across routed layers.

While the heatmap and expert allocation reveal how the experts are allocated layer-wise as well as category-wise, they do not indicate how diverse or concentrated the routing distribution is. To examine this aspect, we analyse routing entropy across the routed layers.

5.4 Routing Entropy Analysis

The entropy values vary across different layers and among the different categories as well. Grammatical probes exhibit the lowest entropy consistently in all the layers (with the strongest effect in layer 3), which suggests that the expert utilisation distribution is more concentrated with a smaller subset of experts consistently utilized for this category.

In contrast, the syntactic probes exhibit the highest entropy values across different layers (or equal to lexical in layer 1). This indicates that expert usage is spread across more experts for this particular category across all layers. Together, these observations suggest that different linguistic categories rely on distinct routing strategies, with grammatical processing exhibiting more concentrated expert utilisation and syntactic processing exhibiting more distributed expert usage.

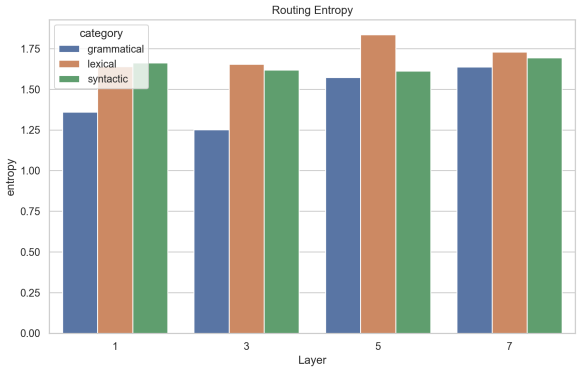

Figure 4: Routing entropy across routed layers and linguistic categories. Lower entropy indicates more concentrated expert utilisation.

5.5 Category Separation Analysis

While entropy explains how concentrated or distributed the expert utilisation is, it does not explain how the routing distribution differs between different linguistic categories. To examine this aspect, we compute pairwise Jensen–Shannon Divergence (JSD) between category-wise routing distributions.

Left Category	Right Category	JSD
Grammatical	Lexical	0.0788
Grammatical	Syntactic	0.1147
Lexical	Syntactic	0.1218

Table 4: Pairwise Jensen–Shannon Divergence between category-conditioned routing distributions.

Table 4 shows the pairwise JSD values between grammatical, lexical and syntactic routing distributions. All category pairs exhibit non-zero divergence, suggesting that the router assigns different expert utilisation distributions to different linguistic categories. We can observe that the highest divergence in the routing distribution is between lexical and syntactic processing, whereas the lowest divergence is between grammatical and lexical processing. These observations are consistent with the mutual information and expert allocation analyses, providing additional evidence for category-dependent routing.

5.6 Effect of Curriculum on Specialisation

In the no-curriculum setting, the mutual information peaks at 0.2599 in layer 5 and reaches 0.1945 in layer 1. Similar to the forward curriculum setting, the strongest routing-category association is observed in the intermediate routed layers. These observations indicate that routing specialisation emerges under both training conditions, suggesting that curriculum learning is not strictly necessary for category-dependent routing to develop.

Layer	No-Curr MI
1	0.1945
3	0.0674
5	0.2599
7	0.0739

Table 5: Pooled bilingual routing-category MI under the no-curriculum condition.

The following table shows language-wise MI values computed separately for English and German.

Layer	Fwd EN	Fwd DE	No-Curr EN	No-Curr DE
1	0.0508	0.1361	0.3810	0.1357
3	0.0366	0.0911	0.1364	0.0804
5	0.1025	0.1444	0.6457	0.1116
7	0.0571	0.0053	0.1141	0.0649

Table 6: Language-wise routing-category mutual information for the forward curriculum and no-curriculum settings.

To verify that the observed category-sensitive routing is not reducible to language identity alone, we additionally compute conditional mutual information $I(R; C \mid L)$, where L denotes language identity. This analysis measures whether routing still contains information about lexical, grammatical, and syntactic category membership after conditioning on whether the probe is English or German.

As reported in Appendix Table 11, conditional mutual information remains non-zero across routed layers in both training conditions, indicating that the routing-category relationship is not explained solely by English-German separation.

Table 6 shows the language-wise mutual information for each routed layer under both the forward curriculum and no-curriculum settings. The language-wise MI values are drastically different for English. In particular, at layer 5, the MI value increases from 0.1025 in the forward curriculum setting to 0.6457 in the no-curriculum setting, representing an increase of approximately 6.3 times. A similar trend can be observed across all routed layers.

In contrast, the difference between the category-dependent routing effect of the two setups is much less visible in German, where MI remains around 0.1 at layers 1 and 5 for both setups (0.1444 in the forward curriculum setting and 0.1116 in the no-curriculum setting at layer 5 for German language). A similar effect is observed in the other layers, where the differences remain comparatively small.

From the above observations, curriculum does not appear to uniformly increase routing specialisation. Instead, the effect of curriculum differs across languages. While the no-curriculum condition produces substantially stronger specialisation in English, the forward curriculum yields a more balanced EN–DE specialisation profile. In contrast, the no-curriculum specialisation pattern is strongly dominated by English routing behaviour.

6 Discussion

Routing distributions provide significant insights into linguistic category membership, even with held-out validation data. Specialisation shows up in both the forward-curriculum and no-curriculum conditions. This indicates that curriculum learning is not necessary for category-sensitive routing to develop. The key question is how curriculum changes how this specialisation is distributed across different languages and layers.

The strongest relationship between routing and category occurs in the intermediate routed layers, particularly in layer 5. This aligns with previous research that suggests intermediate transformer layers often feature clearer linguistic structures. Conversely, the final routed layer exhibits lower mutual information and higher entropy. This implies that later routing becomes more evenly spread across experts instead of being distinctly separated by category.

The comparison between the two training conditions is crucial. The no-curriculum model shows stronger routing-category dependence in English, which boosts the overall bilingual mutual information. However, this stronger pooled signal mainly originates from English and does not distribute evenly between English and German. The forward curriculum creates a more balanced bilingual routing profile and maintains the German category structure more robustly in the middle layers. Therefore, sequential L1–L2 exposure does not simply boost specialisation everywhere; it alters how specialisation spreads across languages.

This is important because pooled statistics can be misleading. When combining English and German into one score, the no-curriculum model seems stronger. Yet, when analyzing the two languages separately, the staged curriculum appears to lessen English's dominance and foster a more balanced bilingual routing structure. This provides a clearer understanding of the effects of the curriculum than merely assessing which model has higher total mutual information.

Interpreting behavioral accuracy values requires considering the study design. The model is intentionally small, and the validation set is challenging since it uses controlled minimal-pair contrasts. The aim is not to achieve high benchmark accuracy but to compare internal routing patterns across controlled linguistic probe families. In this context, the behavioral results primarily indicate that both models are sufficiently functional for meaningful routing analysis.

7 Conclusion

Bilingual MoE routing develops measurable category-dependent organization across lexical, grammatical, and syntactic probe families. The strongest routing specialisation is observed in intermediate routed layers and remains robust on held-out validation data. Curriculum learning does not uniformly increase specialisation; instead, it redistributes it across languages. While the no-curriculum condition produces stronger English routing-category dependence, staged L1–L2 exposure yields a more balanced bilingual specialisation profile and stronger German middle-layer routing structure.

Limitations

This study uses a single EN–DE language pair and a single primary sparse architecture. While the observed routing patterns are consistent across multiple analyses, including mutual information, entropy, Jensen–Shannon divergence, permutation testing, and conditional mutual information controls, the extent to which the findings generalize to other language families, larger-scale MoE architectures, or different curriculum schedules remains an open question.

The diagnostic validation set is generated through a controlled LLM-assisted pipeline and subsequently validated using parser-based checks and deterministic probe construction rules. Although these procedures improve consistency and coverage, the evaluation set does not undergo manual human validation. Future work could incorporate expert human review and naturally occurring linguistic examples to further verify that the observed routing patterns generalize beyond synthetic diagnostic probes. The validation set is dominated by single-piece probes under the mBERT tokenizer, but full subword-aggregation robustness was not recomputed for the final archived routing logs used in this draft

The present study uses distinct source corpora for English and German. Although language-conditioned analyses remain significant, future work should examine matched-domain multilingual corpora.

A reverse curriculum is not included because the forward curriculum is explicitly designed around a fixed cumulative English-German exposure ratio. Simply reversing the schedule would alter not only the temporal order of language presentation but also the total exposure received by each language across training. Consequently, a naive DE→EN reversal would confound sequencing effects with differences in cumulative language exposure. A fair reverse-curriculum comparison would therefore require a separately constructed schedule that preserves overall language proportions while reversing the order of introduction.

Several potential confounds were explicitly examined. Routing-category associations remained stable across alternative subword aggregation strategies and remained non-zero after conditioning on language identity and coarse lexical frequency controls. Nevertheless, the present study does not fully disentangle all possible interactions between linguistic category, token identity, lexical frequency, morphological complexity, and part-of-speech information. As a result, the reported routing effects should be interpreted as category-sensitive routing behaviour rather than evidence of perfectly isolated category-specific mechanisms.

A stricter frequency-balanced lexical control was explored but was not included in the final analysis. The curriculum-generated lexical distributions were highly skewed, making it difficult to construct well-matched irregular and regular subsets while maintaining sufficient probe coverage and sample diversity.

We do not yet report extensive architectural sweeps, multiple language pairs, or broad-scale hyperparameter sensitivity analyses. In addition, expert intervention experiments (e.g., expert masking or routing interventions) are not included, limiting our ability to make strong causal claims about the functional role of individual experts.

Finally, the work focuses specifically on sparse MoE architectures because expert-routing distributions constitute the primary object of analysis. Dense transformers do not expose an explicit routing mechanism, making direct comparisons of routing specialisation impossible. While dense baselines remain useful for behavioural benchmarking, they cannot provide the routing-level signals studied in this work.

While the present study focuses on identifying category-dependent routing behaviour, future work could perform expert masking or routing interventions to determine whether the identified expert preferences play a causal role in linguistic processing.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning*, pages 41–48.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory*, 2 edition. Wiley.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng,

Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1280–1297.

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

Tobias Falke, Nicolas Anastassacos, Samson Tan, Chankrisna Richy Meas, Chandana Satya Prakash, Nitesh Sekhar, M. Saiful Bari, Krishna Kompella, and Gamaleldin F. Elsayed. 2026. Moe routing testbed: Studying expert specialization and routing behavior at small scale.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Zhuoqun Li, Hongyu Lin, Yaojie Lu, Hao Xiang, Xianpei Han, and Le Sun. 2024. Meta-cognitive analysis: Evaluating declarative and procedural knowledge in datasets and large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11222–11228.

Jianhua Lin. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabás Póczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1162–1172.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *Proceedings of the International Conference on Learning Representations*.

Hanchi Sun, Yixin Liu, Yonghui Wu, and Lichao Sun. 2026. Expert threshold routing for autoregressive language modeling with dynamic computation allocation and load balancing.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601.

Michael T. Ullman. 2001a. The neural basis of lexicon and grammar in first and second language: The declarative/procedural model. *Bilingualism: Language and Cognition*, 4(2):105–122.

Michael T. Ullman. 2001b. A neurocognitive perspective on language: The declarative/procedural model. *Nature Reviews Neuroscience*, 2(10):717–726.

Michael T. Ullman. 2020. The declarative/procedural model. In Bill VanPatten, Jessica Williams, Gregory D. Keating, and Stefanie Wulff, editors, *Theories in Second Language Acquisition: An Introduction*, 3 edition, pages 128–161. Routledge.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, pages 5998–6008.

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.

Xuan Zhang, Pamela Shapiro, Gaurav Kumar, Paul McNamee, Marine Carpuat, and Kevin Duh. 2019. Curriculum learning for domain adaptation in neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pages 1903–1915.

A Probe Token Taxonomy

English lexical probes

A token τ is classified as English lexical if either condition holds:

1. **Irregular past-tense verb.** $\text{UPOS}(\tau) = \texttt{VERB}$, $\texttt{Tense=Past} \in \text{FEATS}(\tau)$, $\texttt{VerbForm=Fin} \in \text{FEATS}(\tau)$, and $\text{LEMMA}(\tau) \in \mathcal{V}_{\text{irr}}^{\text{EN}}$.

2. **Irregular plural noun.** $\text{UPOS}(\tau) = \texttt{NOUN}$ and $\text{TEXT}(\tau) \in \mathcal{N}_{\text{irr}}^{\text{EN}}$.

German lexical probes

German lexical probes are operationalized as item-specific or lexically exceptional forms:

1. **Strong past-tense verbs.** $\text{UPOS}(\tau) = \texttt{VERB}$, $\texttt{Tense=Past} \in \text{FEATS}(\tau)$, $\texttt{VerbForm=Fin} \in \text{FEATS}(\tau)$, and $\text{LEMMA}(\tau) \in \mathcal{V}_{\text{irr}}^{\text{DE}}$.

2. **Suppletive or highly irregular finite paradigm forms.** $\text{TEXT}(\tau) \in \mathcal{S}_{\text{DE}}$ and $\text{UPOS}(\tau) \in \{\texttt{VERB}, \texttt{AUX}\}$.

3. **Irregular plural nouns.** $\text{UPOS}(\tau) = \texttt{NOUN}$, $\texttt{Number=Plur} \in \text{FEATS}(\tau)$, and $\text{TEXT}(\tau) \in \mathcal{N}_{\text{irr}}^{\text{DE}}$.

4. **Gender-marked determiners.** $\text{UPOS}(\tau) = \texttt{DET}$ with nominative singular definite or indefinite gender-marked forms. These probes are grouped with lexical probes because they depend on lexically specified noun-gender distinctions, even though they also involve morphosyntactic marking.

English grammatical probes

1. **Regular past-tense verb.** $\text{UPOS}(\tau) = \texttt{VERB}$, $\texttt{Tense=Past} \in \text{FEATS}(\tau)$, $\texttt{VerbForm=Fin} \in \text{FEATS}(\tau)$, and $\text{LEMMA}(\tau) \notin \mathcal{V}_{\text{irr}}^{\text{EN}}$.

2. **Auxiliary agreement.** $\text{UPOS}(\tau) = \texttt{AUX}$ with number-marked auxiliary agreement.

German grammatical probes

1. **Weak past-tense verb.** $\text{UPOS}(\tau) = \texttt{VERB}$, $\texttt{Tense=Past} \in \text{FEATS}(\tau)$, $\texttt{VerbForm=Fin} \in \text{FEATS}(\tau)$, $\text{LEMMA}(\tau) \notin \mathcal{V}_{\text{irr}}^{\text{DE}}$, and the form is not part of the suppletive set.

2. **Auxiliary agreement.** $\text{UPOS}(\tau) = \texttt{AUX}$ with number-marked auxiliary agreement.

Syntactic probes

1. $\text{UPOS}(\tau) = \texttt{SCONJ}$, or

2. $\text{DEPREL}(\tau) \in \{\texttt{advcl}, \texttt{ccomp}, \texttt{xcomp}, \texttt{acl}, \texttt{csubj}\}$.

These labels should be interpreted as operational probe families rather than perfectly discrete theoretical classes. Their purpose is to test whether MoE routing is sensitive to broad linguistically motivated contrasts under a consistent parser-based annotation scheme.

B Full Curriculum Schedule

Category sampling

Within every epoch, lexical, grammatical, and syntactic sentences are sampled uniformly and shuffled randomly.

Forward curriculum

Epochs	λ_{DE}	Description
1–8	0.00	EN only
9	0.20	L2 introduction
10	0.25	
11	0.30	
12	0.35	
13	0.40	
14	0.45	
15–16	0.50	Stable bilingual

Table 7: Forward curriculum L2 exposure schedule.

C Hyperparameter Details

Hyperparameter	Value
Layers	8
Embedding size	512
Attention heads	8
Feed-forward size	2048
MoE layers	1, 3, 5, 7
Experts N	8
Top-k	3
Load-balance α	0.01
Vocabulary	mBERT WordPiece, $\sim$119k
Optimizer	AdamW
Peak LR	3×10^{-4}
Warmup steps	500
Batch size	8 sequences
Gradient clip	1.0
Epochs	16

Computing Infrastructure and Budget

All training, validation, and routing-analysis experiments were conducted on NVIDIA T4 GPUs and NVIDIA RTX A6000 GPUs. The primary bilingual MoE model used in the study contains 86.4M parameters. Each training condition was run for 16 epochs with 300,000 sentence instances per epoch, corresponding to 4.8M sentence instances per condition. Validation, routing extraction, and downstream statistical analyses were performed from saved checkpoints and held-out routing logs on the same hardware class. Depending on hardware availability, experiments were conducted on

either NVIDIA T4 or NVIDIA RTX A6000 GPUs. Training time therefore varied across runs, ranging from approximately 14–28 GPU-hours per condition

D Additional Routing Analysis Details

Permutation testing

Category labels are shuffled at the sentence level for 1,000 permutations per layer. Empirical p-values are computed as the fraction of permutations with $I(R;C) \geq I_{\text{observed}}$. This tests whether the observed routing-category dependence is larger than would be expected under category-independent routing.

E OpenAI Generation Details, Validation Parsing, and Prompt Examples

The held-out validation sentences were generated with the OpenAI Responses API using the Python OpenAI client. In the generation script, the default model was `GPT-5.1`, configurable through the `OPENAI_MODEL` environment variable. Requests were made with `max_output_tokens=420`, `reasoning.effort="none"`, `store=false`, request timeout `40s`, and `max_retries=0`. The API call also included a fixed instruction string: *"Follow the output format exactly. Return only the requested lines and nothing else."*

Generated candidates were then automatically parsed and validated with Stanza using the processors `tokenize`, `pos`, `lemma`, and `depparse`. For each generated S^+ sentence, the pipeline checked that the marked probe token occurred exactly once, that the output matched the required subtype inventory, and that the parsed probe token satisfied the subtype-specific constraints used in dataset construction (e.g., finite past-tense verb, plural noun, auxiliary, or subordinating conjunction, depending on the subtype). Additional filters removed malformed outputs, duplicates, and subtype violations before acceptance into the held-out set.

For the final validation pair construction, the ungrammatical sentence S^- was obtained deterministically from the validated S^+ sentence by editing only the probe token or deleting the probe in the syntactic conjunction cases. The transformation rule depended on subtype. For example, English irregular past forms were replaced with regularized forms, English regular past-tense verbs were replaced with their lemma/base form, auxiliary agreement probes were swapped to mismatching number forms, German strong verbs were weak-regularized, German suppletive forms were replaced with present/agreement-incompatible alternatives, and subordinating conjunction probes were removed to create a syntactically degraded variant. This yielded minimal-pair contrasts in which S^+ remained parser-validated and S^- differed by a controlled probe-level manipulation.

Below we provide two illustrative prompt examples adapted directly from the generation templates used in the pipeline.

Prompt Example 1: English irregular verb probe.

> Write exactly 10 EN sentences.
>
> Subtype: irregular_verb
> Preferred lemmas: become, draw, awake, speak, throw, choose, write, drive, sing
> Avoid recent probes: none
>
> Rules: - Output exactly 10 lines, no more and no fewer. - One sentence per line. - 6–12 words per sentence. - Prefer a different listed lemma on each line. - Mark target as [PROBE: word]. - Prefer the correct irregular simple-past form of one listed lemma. - Probe must be the main verb. - Use simple past only. - No auxiliaries or participles with the probe. - Use the real irregular past form, not the base form. - Examples: become → became, awake → awoke, draw → drew. - Prefer the listed lemmas strongly. - If one listed lemma feels awkward, use another listed lemma. - No numbering, bullets, quotes, parentheses, or extra text. - Bare sentence text only.

Prompt Example 2: German subordinating conjunction probe.

> Write exactly 10 DE sentences.
>
> Subtype: sconj
> Preferred subordinating conjunctions: weil, obwohl, bevor, nachdem, falls, sobald, damit, dass
> Avoid recent probes: none
>
> Rules: - Output exactly 10 lines, no more and no fewer. - One sentence per line. - 6–12 words per sentence. - Prefer a different listed conjunction on each line. - Mark target as [PROBE: word]. - Prefer one listed subordinating conjunction exactly as shown. - Use the probe as a true subordinating conjunction with UPOS=SCONJ. - Build a clear subordinate-clause construction. - Keep the sentence fully grammatical as written. - Avoid coordinators, adverbs, or discourse markers. - No numbering, bullets, quotes, parentheses, or extra text. - Bare sentence text only.

Parsing and Validation Tools

Parser-based probe annotation and validation were performed with Stanza. We used the Stanza pipelines for English and German with the processors `tokenize`, `pos`, `lemma`, and `depparse`. These pipeline outputs were used to identify probe tokens,

verify subtype constraints, and validate generated held-out examples.

F Dataset Diversity Statistics

To address concerns regarding dataset health and template collapse, we provide diversity statistics for the diagnostic set. All reported probe sets maintain 100% sentence-pair uniqueness.

Probe Set	Records	Tokens	Uniqueness
EN Irregular Verb	1500	46	100%
EN Irregular Plural	1500	30	100%
DE Strong Verb	500	35	100%
EN SCONJ	3000	21	100%
EN Aux Agreement	1500	17	100%

Table 8: Illustrative diagnostic probe diversity statistics.

G Ablation Controls

The appendix currently includes four additional controls beyond the main forward and no-curriculum MoE comparisons: random routing, frozen routing, top-k, and load-balancing ($\alpha = 0.005$) ablations under the forward curriculum. These controls are intended to distinguish curriculum effects from generic sparse-routing effects and to test how sensitive the observed specialisation is to router flexibility and routing sparsity.

Ablation	**Combined**	**EN**	**DE**
Random routing	0.0018	0.0029	0.0057
Frozen routing	0.0274	0.0647	0.0140
Load balancing ($\alpha = 0.005$)	0.1482	0.3288	0.1042
Forward top-$k = 2$	0.1100	0.1767	0.1349
Forward top-$k = 4$	0.2382	0.4107	0.1144

Table 9: Routing-control and sparsity ablations evaluated on the held-out validation set (MI@L5). Random routing collapses routing-category dependence toward near-null values, indicating that specialisation does not arise from architectural sparsity alone. Freezing router parameters preserves weaker but non-zero specialisation. Lower load balancing increases routing-category dependence, while the routing-sparsity ablations show that top-$k = 2$ yields a similar but slightly lower pooled MI than the main model, whereas top-$k = 4$ produces substantially stronger category-conditioned routing, especially for English probes.

The random-routing condition replaces learned router assignments with uniformly sampled expert selection while preserving the underlying expert parameters. As expected, routing-category mutual information collapses to near-zero values, indicating that the observed specialisation patterns require adaptive routing rather than arising from architectural sparsity alone.

The frozen-routing condition initializes routing normally but prevents subsequent router optimization during training. Although specialisation is substantially weaker than in the fully trainable model, non-trivial routing-category dependence remains, suggesting that expert differentiation can emerge through changes in token representations interacting with fixed routing boundaries.

Load-balancing strength additionally exerts a strong influence on specialisation structure. Reducing the auxiliary load-balancing coefficient to $\alpha = 0.005$ markedly increases routing-category dependence, indicating that weaker balancing constraints permit stronger expert partitioning. However, this increase coincides with increasingly skewed expert utilisation, suggesting a trade-off between specialisation strength and balanced expert participation.

Routing sparsity also affects specialisation, but not in the initially expected direction. Reducing expert selection from top-$k = 3$ to top-$k = 2$ leaves pooled routing-category dependence at a similar level, with a slight decrease overall. In contrast, increasing routing breadth to top-$k = 4$ yields a substantially stronger routing-category association, especially for English probes. In the present runs, broader routing therefore coincides with stronger category-conditioned separation rather than weaker specialisation.

G.1 Frequency-Binned Lexical Analysis

To examine whether lexical routing specialisation is driven primarily by token frequency, English lexical probes were grouped into frequency bins using curriculum-weighted lemma frequencies. Mutual information was then recomputed using only lexical irregular-versus-regular contrasts within each frequency bin.

Frequency Bin	Forward MI@L5	No-Curr MI@L5
Frequent	0.0149	0.0055
Medium	0.0206	0.0976
Rare	0.2261	0.2494

Table 10: English lexical irregular-versus-regular routing mutual information at layer 5 after binning probe lemmas by curriculum-weighted frequency.

Mutual information remains non-zero across all frequency bins in both training conditions. Notably, the strongest routing-category association is observed for rare lexical items rather than the

most frequent items. This suggests that the lexical routing effect cannot be explained solely by the highest-frequency lexical forms.

Layer	Forward $I(R;C \mid L)$	No-Curr $I(R;C \mid L)$
1	0.0935	0.2584
3	0.0639	0.1084
5	0.1235	0.3787
7	0.0312	0.0895

Table 11: Conditional mutual information between routing and linguistic category given language, $I(R;C \mid L)$, across routed layers. In both the forward-curriculum and no-curriculum conditions, conditional MI remains clearly non-zero and permutation-significant at all routed layers ($p = 0.000999$), indicating that category-sensitive routing is not reducible to language separation alone.

We additionally examined the evolution of routing specialisation during training using epoch-wise routing logs. These analyses are provided here as supplementary evidence.

G.2 Routing Specialisation Dynamics During Training

To analyse routing specialisation throughout training, mutual information and entropy were computed from the routing logs at checkpoints from each training epoch. Figure 5 shows that category-dependent routing emerges rapidly during the early stages of training and stabilises over the following epochs.

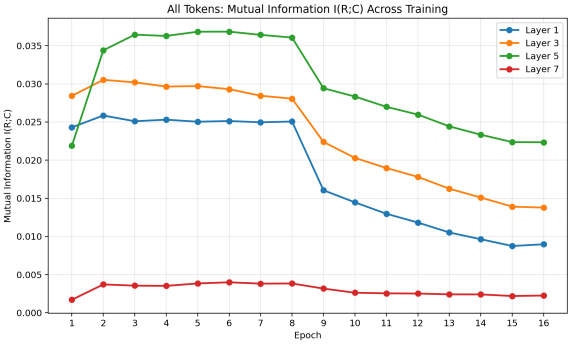

Figure 5: Evolution of routing-category mutual information across training epochs for the routed layers.

The relative ordering of the routed layers remains largely consistent throughout training, with layer 5 exhibiting the highest mutual information across all epochs. The figure also shows a decline in category-dependent routing after epoch 8. Following this transition, the mutual information decreases across all routed layers before stabilising again during the later stages of training.

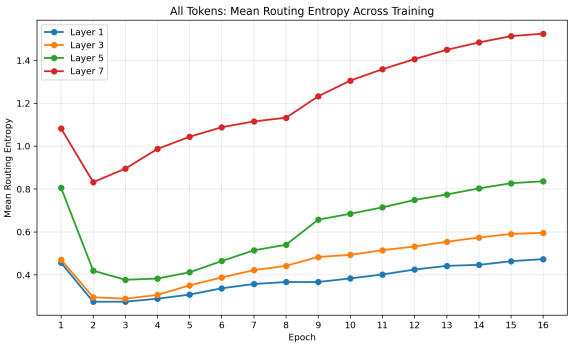

Figure 6: Evolution of routing entropy across training epochs for the routed layers.

Figure 6 presents the routing entropy across the routed layers throughout training. A complementary trend is observed between mutual information and entropy. As mutual information decreases after epoch 8, routing entropy increases across all routed layers during the same period. This indicates that expert utilisation becomes progressively more distributed while category-dependent routing behaviour remains present.