Disentangling Continued Pre-Training: Attention-Driven Routing and Semantic Hub Preservation in Language Adaptation

Khanh-Tung Tran, Vinh-Khanh Tran, Barry O'Sullivan, Hoang D. Nguyen
Research Ireland Centre for Research Training in Artificial Intelligence
Insight Research Ireland Centre for Data Analytics
School of Computer Science and Information Technology, University College Cork, Ireland
{123128577, 125117729}@umail.ucc.ie, {b.osullivan, hn}@cs.ucc.ie

Abstract

Continued Pre-Training (CPT) enables Large Language Models (LLMs) to acquire second-language capabilities, yet the underlying mechanisms remain poorly understood. In this work, we investigate how CPT adapts model representations across diverse language families and scripts, model sizes, and architectures. We find that second-language abilities emerge through a selective adaptation mechanism: task-solving capabilities are preserved in the "semantic hub", while interface layers retarget to shifted token distributions. Layer-swapping experiments demonstrate that semantic understanding can be surgically transferred between base and CPT models with minimal loss (e.g., swapping 50% of model parameters reduces performance by only 0.3%). Furthermore, we establish that attention components route language adaptation: larger parameter changes than feedforward networks, correlate more strongly with language-specific neurons, and their surgical replacement substantially degrades performance. Overall, our work provides a mechanistic understanding of CPT, guiding future work on efficient strategies for language adaptation.

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities across a wide range of NLP tasks (Qin et al., 2025; Matarazzo and Torlone, 2025). However, they remain limited for low-resource languages, reflecting the linguistic imbalance in pre-training corpora (Touvron et al., 2023). Continued Pre-Training (CPT) has proven to be an effective method for language adaptation (Shi et al., 2025). By exposing pre-trained LLMs to additional corpora in a target language, CPT allows LLMs to acquire language-specific knowledge and enhance performance in the target language (Parmar et al., 2024; Yong et al., 2023), even in low-resource settings (Tran et al., 2024b).

Prior work primarily studies the final model, without examining how patterns evolve during CPT. For example, Tang et al. (2024) identify language-specific neurons, while other studies demonstrate the existence of a "semantic hub" in LLMs where semantic concepts are encoded agnostically across languages and modalities (Wendler et al., 2024; Hämmerl et al., 2024; Wu et al., 2025). Recent work has begun to investigate CPT dynamics. Elhady et al. (2025) reveal the critical role of including English data to mitigate catastrophic forgetting. Zhang et al. (2025) study how cultural knowledge transfer during CPT. However, these studies treat models as monolithic systems and overlook the component-level mechanisms, such as how attention and feedforward network (FFN) layers contribute differently to adaptation. Fundamental questions remain unanswered: *which components of the model are most responsible for learning new language-specific knowledge, how existing capabilities are preserved, and what structural changes occur during adaptation?*

In this work, we conduct systematic experiments to provide a mechanistic understanding of how CPT enables language adaptation, looking at the whole training trajectory rather than just inspecting the model after convergence. Specifically, we analyze layer-wise functional specialization, measure parameter dynamics in interface layers, and perform causal interventions through layer-swapping experiments to isolate functional contributions. Our key contributions are:

- Through sentence-retrieval–based layer identification, parameter-drift analysis, and layer-swapping interventions, we show that **CPT enables second-language abilities through a selective adaptation mechanism: interface layers adapt to shifted token distributions, while task-solving capabilities are preserved within the semantic hub.**

Findings of the Association for Computational Linguistics: ACL 2026, pages 24335–24357

- Using weight-change and correlational analyses with language-specific neuron activations and component-level swapping experiments, we identify the **attention component as the routing factor behind language adaptation.**
- We validate our findings across the CPT processes of a broad range of models, languages, and experimental settings. Our experiments span Chinese, Irish, and Basque; model sizes from 0.5B to 13B; and architectures from Qwen2.5 to Llama2. For example, swapping semantic hub layers between base and CPT models incurs only a 0.3% performance drop for L2-13B-GA across 4 different tasks (FLORES+, Belebele, SIB200, and IrishQA) while swapping the same number of layers randomly causes up to a 66.29% degradation ($p = 0.016$).

Our source code, model weights, and results are made publicly available at `https://github.com/ReML-AI/disentangling_cpt`

2 Related Works

2.1 CPT for Language Adaptation

CPT has become a standard method for adapting LLMs to new domains and languages (Fujii et al., 2024; Gururangan et al., 2020). CPT has proven to be effective for low-resource languages through cross-lingual transfer (Tran et al., 2024b).

Recent works begin to study the dynamics and mechanism of CPT for language adaptation. For instance, Zhang et al. (2025) study the dynamics of cultural knowledge transfer between high-resource languages and low-resource languages in LLMs during CPT. Most relevant to our work, Elhady et al. (2025) investigate how emergent abilities arise during CPT and show that including English during CPT is critical for downstream capabilities despite having negligible impact on validation perplexity. However, they did not inspect the inner workings of the model or how specific components contribute to the acquisition of new linguistic abilities. Our work fills this gap by mechanistically analyzing how internal model components evolve and contribute to language adaptation throughout CPT.

2.2 Multilingual Mechanism of LLMs

Recent studies investigate how multilingual LLMs process and handle different languages. For instance, Wendler et al. (2024) show compelling evidence that Llama-based models use English as a pivot language and rely on a universal concept space for cross-lingual transfer. In contrast, Schut et al. (2025) argue that concepts have language-centric representations determined by the training-dominant language. In parallel, Zhao et al. (2024); Wu et al. (2025) demonstrate evidence for the existence of a semantic hub universal across languages or modalities inside LLMs. While these works offer valuable insights into the internal mechanisms of handling multilinguality in LLMs, they primarily analyze the final checkpoint of the models after convergence. In contrast to prior research, our work focuses on how CPT changes and shapes LLMs for language adaptation by inspecting each checkpoint throughout the learning trajectory, examining how functional structures emerge during adaptation rather than characterizing their properties after convergence.

3 Experiment Setup

To investigate the effects of CPT on second-language acquisition, we conduct experiments across a diverse set of languages and model architectures. The base language for all experiments is English, while the target languages include *Irish*, an extremely low-resource language; *Basque*, another low-resource language; and *Chinese*, a high-resource language that does not use Latin scripts.

3.1 Base Model Selection

Base model	Description
Qwen2.5-0.5B	Trained from scratch with 5B English tokens
Llama2-7B	Meta's Llama2 7B model
Llama2-13B	Meta's Llama2 13B model

Table 1: Base models description.

We consider three different base models in our experiments, described in Table 1. As a small model, we use `Qwen2.5-0.5B` (Qwen et al., 2025), which we train from scratch on 5 billion English tokens from the Wikipedia dataset in order to analyze the dynamics of second-language acquisition in a controlled setting. For larger-scale models, we use `Llama2-7B` and `Llama2-13B`, both pre-trained by Meta on a large English-dominant corpus (Touvron et al., 2023), to study whether the observed effects of CPT scale with model size.

Base Model	CPT Settings	CPT Models
Qwen2.5-0.5B	1.5B Chinese tokens	Q2.5-ZH
	300M English tokens + 1.5B Chinese tokens	Q2.5-EN-ZH
	1.5B Irish tokens	Q2.5-GA
	300M English tokens + 1.5B Irish tokens	Q2.5-EN-GA
	1.5B Basque tokens	Q2.5-EU
	300M English tokens + 1.5B Basque tokens	Q2.5-EN-EU
Llama2-7B	4.7B Basque tokens	L2-7B-EU
Llama2-13B	1.5B Irish tokens (UCCIX checkpoints)	L2-13B-GA

Table 2: CPT settings with abbreviations.

3.2 CPT Settings

Following established practices in the field (Zhang et al., 2025; Elhady et al., 2025), we construct language-specific corpora from the Wikipedia dataset (Foundation). To ensure comparability across languages within the same experimental setting, we standardize the token count for each language with respect to the base model under study. For Qwen2.5-0.5B, we use approximately 1.5B tokens per target language, we also consider curriculum learning by including 300M English tokens for each target language experiment. For Llama2-7B, we perform CPT with 4.7B Basque tokens from Wikipedia, following the work of Elhady et al. (2025). Finally, for Llama2-13B, we build on existing work by reusing checkpoints released by UCCIX (Tran et al., 2024b), a Llama2-based model that has been adapted to Irish through CPT. Overall, the CPT settings are described in Table 2.

We use NVIDIA HGX/H100 and DGX/A100 cluster for our experiments. We configure both model families (Qwen2.5-based and Llama2-based) with a learning rate of 1e-4, implementing cosine scheduling with a 10% warmup ratio. Maximum sequence lengths differ between models: 2048 for Qwen2.5 and 4096 for Llama2.

4 Interface Layer Language Adaptation

Prior studies suggest that not all layers in LLMs serve identical functions during inference (Zhao et al., 2024). Recent work by Wu et al. (2025) proposes the semantic hub hypothesis, which posits that a model's initial layers process data in its specific language or modality, while intermediate layers convert tokens into language-agnostic representations, functioning as the semantic hub. However, no prior work has investigated how the semantic hub behaves or emerges during CPT, or what mechanism preserves its language-agnostic properties while the model acquires new linguistic abilities.

We hypothesize that CPT enables second-language abilities through a selective adaptation mechanism: certain layers specialize in language-specific processing and undergo substantial changes to align with new token distributions (denoted *interface* layers), while other layers preserve cross-lingual semantic and task-solving capabilities (denoted *semantic hub* layers).

4.1 Identifying Layer Specialization

First, to empirically identify interface and semantic hub layers, we leverage the sentence retrieval task (Tran et al., 2024a; Yong et al., 2023), measuring the semantic alignment capability across models' layers. Layers with high retrieval accuracy indicate language-agnostic semantic representations, identifying them as semantic hub layers where meaning is encoded independently from language-specific attributes, while layers with low sentence retrieval accuracy encode language-specific features such as morphology and syntax, effectively serving as interface layers.

We use FLORES+ (NLLB Team et al., 2024) to evaluate sentence retrieval accuracy at each layer. For each CPT model θ^* with $\mathcal{L}$ layers, we select *interface* layers and *semantic hub* layers based on binary *k-mean* clustering ($k = 2$) (MacQueen, 1967) on retrieval accuracy of the final checkpoint:

$$\{C_1, C_2\} = \text{k-means}(\{\mathcal{A}_l(\theta^*_{\text{final}}, ret)\}_{l \in \mathcal{L}}, k = 2)$$

The cluster with higher average accuracy is designated as the semantic hub layers while the other contains the interface layers. The interface layers are summarized in Table 4 for each CPT model.

CPT Models	$r(\Delta W_{\text{int}}, SIB200(\theta^*))$	$r(\Delta W_{\text{sem}}, SIB200(\theta^*))$
Q2.5-ZH	0.9771	**0.9811**
Q2.5-EN-ZH	**0.9247**	0.9208
Q2.5-GA	**0.8351**	0.8258
Q2.5-EN-GA	**0.9991**	0.9984
Q2.5-EU	0.9997	**0.9999**
Q2.5-EN-EU	0.9900	**0.9917**
L2-7B-EU	**0.8610**	0.8313
L2-13B-GA	**0.8824**	0.8394
Average	**0.9336**	0.9235

Table 3: Pearson correlation between parameter changes in interface and semantic hub layers with SIB200 performance across checkpoints, showing stronger alignment for interface layers.

CPT Models	Interface Layers	# Total Layers
Q2.5-ZH	0-3	24
Q2.5-EN-ZH	0-3,23	24
Q2.5-GA	0-14,23	24
Q2.5-EN-GA	0-14,23	24
Q2.5-EU	0-14,23	24
Q2.5-EN-EU	0-14,23	24
L2-7B-EU	0-15	32
L2-13B-GA	0-13	40

Table 4: Selected interface layers for each CPT models, the rest are semantic hub layers.

4.2 Interface Layers Parameter Drift Correlate with Downstream Performance

For each CPT model θ^*, we perform benchmarking on the SIB200 (Adelani et al., 2024) multiple-choice task for its corresponding target language L across checkpoints θ_c^*. A 10-shot demonstration is used for each evaluation. The result of each run is denoted as $SIB200(\theta_c^*)$.

We define the weight change between a CPT checkpoint θ_c^* and the base model θ as the L1-norm$_{mean}$ difference across layers:

$$\Delta W(c) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \left\| W_l(\theta_c^*) - W_l(\theta) \right\|_{1,\text{mean}}$$

We compute the correlation r between the downstream accuracy and the weight change between components of the CPT model θ^* and base model θ for each checkpoint c:

- $\Delta W_{int}(c)$: the L1-norm$_{mean}$ weight difference between the interface layer components of θ^* and θ for checkpoint c
- $\Delta W_{sem}(c)$: the L1-norm$_{mean}$ weight difference between the semantic hub layer components of θ^* and θ for checkpoint c

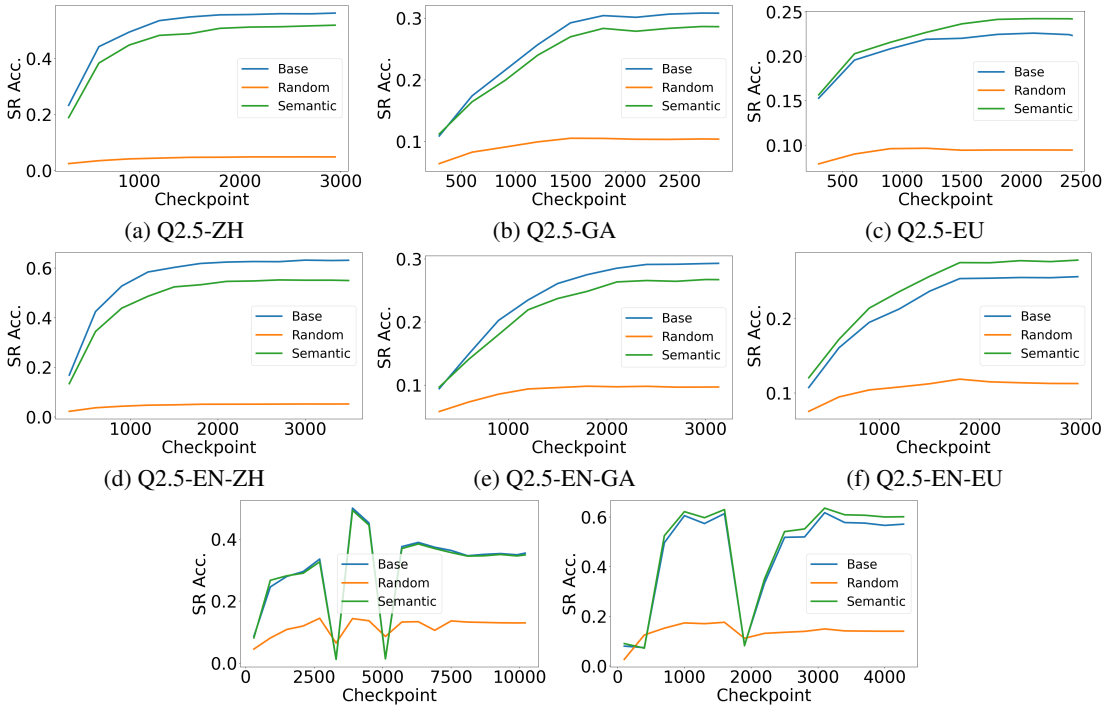
The results, reported in Table 3, show that across most models, parameter changes in interface layers consistently exhibit higher correlation with downstream performance than those in semantic hub layers (average of 0.9336 compared to 0.9235). There are a few exceptions, such as Q2.5-ZH and Q2.5-EU, however, the differences are very small (0.004 and 0.0002, respectively). Across models, the correlation difference is positive (95% CI: $(0.0016, 0.02477)$, $p = 0.069$) and consistent with prior work reporting correlation gaps of similar magnitude (Huang et al., 2025). This pattern suggests that the adaptation of interface layers during CPT is more closely aligned with improvements on downstream tasks than changes in semantic hub layers. In other words, CPT appears to primarily drive performance gains by modifying the interface layers, rather than by altering core semantic representations inside the semantic hub layers.

4.3 Layer Swapping Interventions Confirm Selective Adaptation

We next establish causal evidence for our hypothesis through layer swapping experiments, where semantic hub layers in the CPT checkpoints are surgically replaced with the semantic hub layers from the base models, ceteris paribus.

Formally, for each CPT model θ^* and its checkpoints, we examine the cause and effect in temporal precedence by swapping the *semantic hub* layers from the base model θ into θ^*, obtaining a new model for each checkpoint c, denoted as θ_c^{swap}. We evaluate downstream performance on 4 tasks: sentence retrieval, SIB200 (Adelani et al., 2024), Belebele (Bandarkar et al., 2024), and IrishQA (Tran et al., 2024b). As a baseline, we swap an equal number of randomly selected layers from the base model into the CPT model using 5 different seeds and report the average performance. For Belebele, we report results only for L2-7B-EU and L2-13B-GA, and for IrishQA only for L2-13B-GA as the smaller models (0.5B) perform close to random (Zheng et al., 2024; Cobbina and Zhou, 2025), making the effect of swapping uninformative. Detailed results are reported in the Appendix.

The sentence retrieval results, shown in Figure 1, demonstrate a preservation of semantic functionality: the models maintain strong performance, often approaching or matching the original CPT model with only a 0.7% drop across checkpoints for L2-

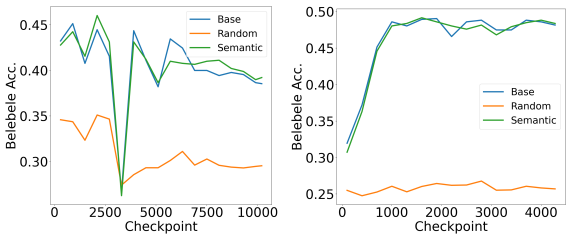


(a) Q2.5-ZH (b) Q2.5-GA (c) Q2.5-EU

(d) Q2.5-EN-ZH (e) Q2.5-EN-GA (f) Q2.5-EN-EU

(g) L2-7B-EU (h) L2-13B-GA

Figure 1: Effects of layer-swapping on sentence retrieval. Swapping semantic hub layers preserves performance, even when up to 50% of parameters are replaced, whereas random layer swapping consistently degrades accuracy.



(a) L2-7B-EU (b) L2-13B-GA

Figure 2: Effects of layer-swapping on Belebele. Same experimental setup as Figure 1.

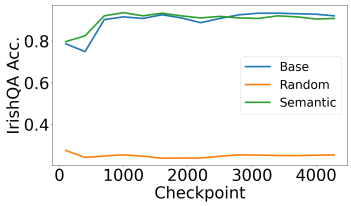
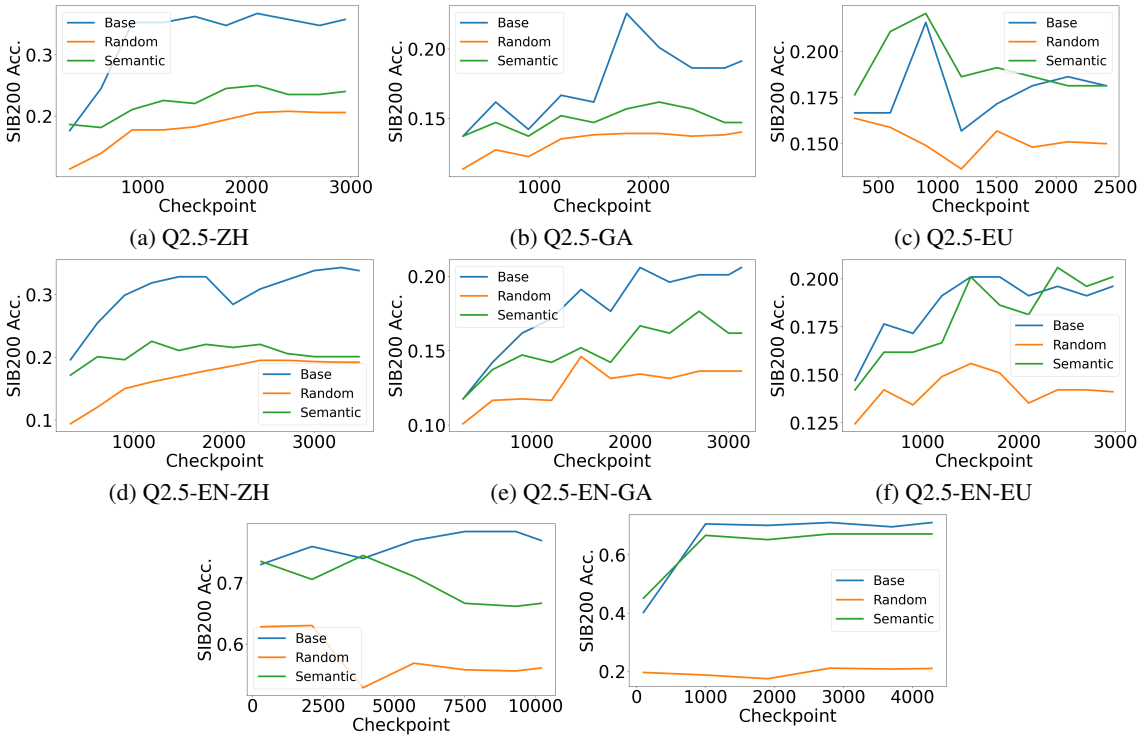


Figure 3: Effects of layer-swapping on IrishQA for L2-13B-GA. Same experimental setup as Figure 1.

7B-EU and 16.9% for Q2.5-ZH while swapping 50% of the parameters. In contrast, swapping the same number of layers picked randomly results in a severe performance decline, with Q2.5-ZH experiencing a dramatic 94.3% drop across checkpoints. This finding is consistent across the CPT process, languages, and model architectures.

Results on other downstream tasks in Figure 2, 3, and 4 continue to strongly support our claim. For example, on Belebele, semantic hub swapping maintains performance with an average 0.6% performance drop for L2-13B-GA, and 0.37% performance increase for L2-7B-EU, while random layer swapping cause severe degradation averaging 43.46% and 23.28% decrease, respectively. Across all 4 tasks, semantic hub layer swapping incurs only a 0.3% performance drop for L2-13B-GA compared to a 66.29% degradation from random swapping ($p = 0.016$). By adhering to the properties of causality, these experiments reveal two key insights. First, semantic understanding is inherent inside the semantic hub layers and can be surgically moved between the base model and the CPT model while retaining their function. Second, CPT works by adapting the interface layers to retarget the new language representation into the semantic representation space, rather than forcing the semantic hub to relearn the semantic representation space.

We note that some models, such as L2-7B-EU, exhibit sharp performance drops at certain checkpoints. We find these are caused by transient training instability. For example, at checkpoint 3300 of L2-7B-EU the gradient norm reaches 2.7 compared to 0.2 near convergence, while the learning rate remains high at 9×10^{-5}. Affected steps are identified as local gradient norm maxima within



(a) Q2.5-ZH (b) Q2.5-GA (c) Q2.5-EU

(d) Q2.5-EN-ZH (e) Q2.5-EN-GA (f) Q2.5-EN-EU

(g) L2-7B-EU (h) L2-13B-GA

Figure 4: Effects of layer-swapping on SIB200. Same experimental setup as Figure 1.

a window of 50 neighboring steps, accounting for only 0.45% of total training steps. Performance recovers in subsequent checkpoints, consistent with known instability patterns in LLM training (Chowdhery et al., 2023; Guo et al., 2025), and does not affect our overall conclusions.

5 Attention-Driven Routing of Latent Language-Specific Neurons

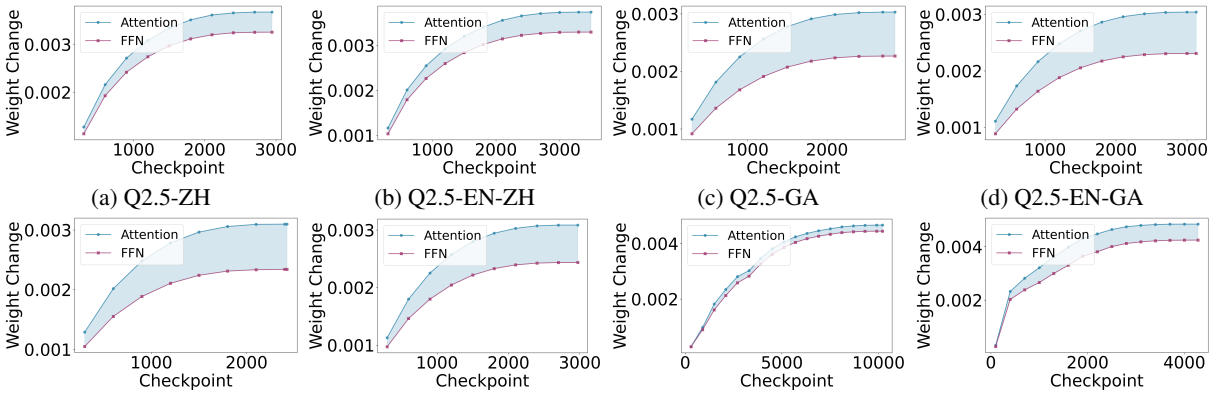
Having established that CPT adapts interface layers while preserving semantic hub functionality, we now investigate which transformer components drive the adaptation. Feed-forward networks (FFNs) contain neurons that perform distinct computational roles (Dai et al., 2022; Voita et al., 2024). More recently, Zhao et al. (2024) identify language-specific neurons, neurons that selectively activate for particular languages, predominantly located within FFNs. We hypothesize that CPT enables language adaptation primarily by adapting attention to route these language-specific neurons. To test this hypothesis, we provide three lines of evidence: (1) Attention weights overall change more than FFN weights during CPT, (2) Attention-pattern changes correlate with language-specific neuron activations, (3) Attention component surgery lowers performance, while swapping FFN experiences minimal performance degradation.

5.1 Attention Components Exhibit Greater Weight Changes than FFN Components

We begin by quantifying the magnitude of changes during CPT. Formally, for each CPT model θ^* we compute the L1-norm$_{mean}$ of weight differences from the base model θ for Attention and FFN components separately for all layers $l \in \mathcal{L}$. We denote these component-specific weight changes as ΔW_{attn} for changes in attention parameters and ΔW_{ffn} for changes in feed-forward parameters

Figure 5 shows that attention weights change more substantially than FFNs across all CPT models and checkpoints. This pattern holds regardless of target language, model size, or training curriculum. This trend becomes more pronounced as training progresses, with the final checkpoint exhibiting a relative difference of 33.75% between attention weight changes and FFN weight changes for Q2.5-GA and 12.87% for Q2.5-EU.

These observations indicate that attention parameters are more responsive to CPT than FFN parameters. While this does not by itself establish a causal role in language adaptation, it suggests that attention may be a key component for model reconfiguration during CPT for language adaptation.



(a) Q2.5-ZH (b) Q2.5-EN-ZH (c) Q2.5-GA (d) Q2.5-EN-GA

(e) Q2.5-EU (f) Q2.5-EN-EU (g) L2-7B-EU (h) L2-13B-GA

Figure 5: Attention components exhibit larger weight changes than FFN components, indicating greater responsiveness to CPT.

To investigate this hypothesis, we next analyze the relationship between attention parameter changes and language-specific neuron activations, which allows us to assess whether the observed weight dynamics are functionally linked to the emergence of second-language capabilities.

5.2 Neuron Activations Correlate with Attention Adaptation

To investigate whether attention changes route language-specific neurons' activations inside FFN components, we leverage Language Activation Probability Entropy (LAPE) (Tang et al., 2024). We first apply LAPE to detect the language-specific neurons in the final checkpoint for each CPT model. We utilize mOscar (Futeral et al., 2024) to avoid overlap with our Wikipedia training data. Following the original setup, we sample 100 million tokens for each language, with the exception of Irish, where we use approximately 70 million tokens due to data limitations.

Once detected at the final checkpoint, we track the activations backward through the base model and all training checkpoints using mOscar. This allows us to observe how the activation of pre-existing neurons evolves during CPT. For each checkpoint c of CPT Model θ^* we compute three quantities:

Attention weight change in interface layers $\Delta W_{\text{attn}}(c)$

Language-specific neurons weight change $\Delta W_{\text{LSN}}(c)$

Language-specific neurons activation change $\Delta a_{\text{LSN}}(c) = \frac{1}{|LSN|} \sum_{i \in LSN} \left(a^i(h;\theta_c^*) - a^i(h;\theta)\right)$

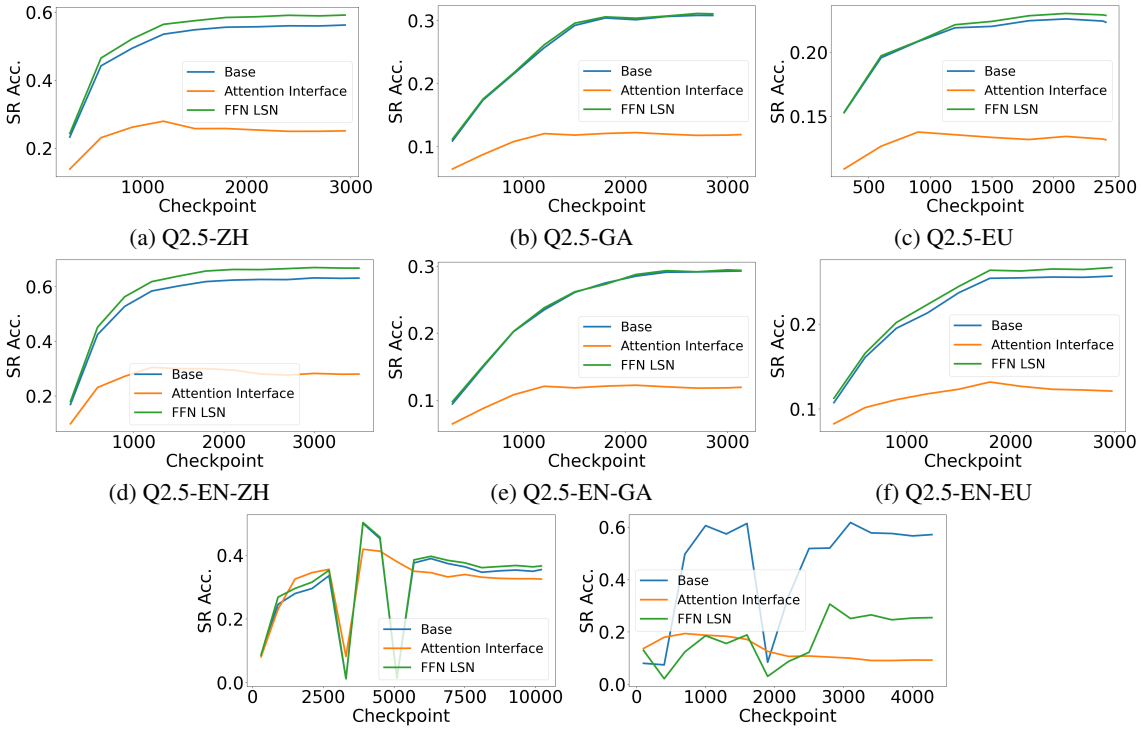
where $a^i(h;\theta)$ is the activation of a language-specific neuron i, computed as a non-linear transformation of input projected through the corresponding column of the gate matrix: $a^i(h;\theta) = \text{act\_fn}\Big(h \cdot W^{gate}[:,i]\Big)$, with W^{gate} denoting the gate projection matrix, and act_fn$(\cdot)$ is the gated activation function (e.g., SiLU), h is the layer input and $W_{\text{gate}}[:,i]$ is the i-th column of the gate projection matrix.

CPT Models	$r(\Delta W_{\text{attn}}, \Delta a_{\text{LSN}})$	$r(\Delta W_{\text{LSN}}, \Delta a_{\text{LSN}})$
Q2.5-ZH	**0.5683**	0.5628
Q2.5-EN-ZH	**0.8184**	0.7980
Q2.5-GA	0.9265	**0.9284**
Q2.5-EN-GA	**0.9028**	0.8977
Q2.5-EU	**0.8078**	0.7952
Q2.5-EN-EU	0.4883	**0.4905**
L2-7B-EU	**0.4398**	0.4095
L2-13B-GA	**0.5724**	0.4504
Average	**0.6905**	0.6767

Table 5: Correlation strength of attention weight changes in interface layers and LSN weight changes with LSN activation changes. In 6 out of 8 models, attention weight changes show stronger correlation, supporting the attention-driven routing hypothesis.

We compute two separate correlations strengths across checkpoints: between attention weight changes and LSN activation changes, and between LSN weight changes and LSN activation changes.

Table 5 shows that attention weight changes correlate more strongly with activation than the neurons' own weight changes in 6 out of 8 models, particularly pronounced in larger models (L2-7B-EU and L2-13B-GA) with higher correlation at 0.03 for L2-7B-EU and 0.122 for L2-13B-GA. The two exceptions (Q2.5-GA and Q2.5-EN-EU) show



(a) Q2.5-ZH (b) Q2.5-GA (c) Q2.5-EU

(d) Q2.5-EN-ZH (e) Q2.5-EN-GA (f) Q2.5-EN-EU

(g) L2-7B-EU (h) L2-13B-GA

Figure 6: Swapping interface-layer attention components substantially degrades performance, despite retaining language-specific FFN neurons, underscoring the importance of attention for CPT.

nearly identical correlations for both attention and neuron weights, with differences of less than 0.002. Across models, the mean correlation difference is positive, (95% CI: $(0.007, 0.069)$, $p = 0.071$), consistent with prior work reporting correlation gaps on the order of 0.01 (Huang et al., 2025). The consistently higher or equal correlation for attention highlights its primary role in enabling language-specific neuron activations during CPT.

5.3 Component Swapping Confirms Attention Dominance

To provide causal evidence that attention mechanisms drive language-specific neuron activation, we perform component swapping experiments. For each CPT model θ^* and checkpoint c, we create a hybrid model by selectively replacing components with their counterparts from base model θ:

- Interface layer attention-swap models $\theta_c^{swap-attn}$: Attention weights in interface layers are replaced with those from the base model while retaining FFN components, isolating the contribution of adapted attention mechanisms to language acquisition.

- FFN language-specific neurons swapped models $\theta_c^{swap-ffn}$: FFN weights in layers containing substantial number of language-specific neurons are replaced with those from the base model while preserving the attention components, isolating the contribution of language-specific FFN components.

The sentence retrieval results in Figure 6 demonstrate that swapping interface-layer attention components from the base model substantially degrades performance (e.g. 51.8% for Q2.5-ZH and 38.1% for Q2.5-EU), despite retaining the language-specific neurons in FFN components, indicating that attention weights have a crucial role in language adaptation during CPT.

Conversely, $\theta_c^{swap-ffn}$ shows minimal change on sentence retrieval accuracy. For Qwen2.5-based models and L2-7B-EU, swapping FFNs can improve performance, with a 1.0% increase for Q2.5-GA and a 5.3% increase for Q2.5-ZH across all checkpoints. Although L2-13B-GA experiences degradation after swapping, it still outperforms the attention-swapped counterparts by 48.7% across checkpoints.

The asymmetry between the two swapping conditions supports our hypothesis that attention serves as the primary mechanism for language acquisition.

6 Conclusion

We demonstrate that CPT facilitates second language acquisition in LLMs through a layer-wise selective adaptive process. By conducting correlational analysis and layer-wise intervention experiments, we find that *interface* layers play an important role in language understanding and adaptation.

Our analysis reveals that attention mechanisms are the primary router of this adaptation process, through three complementary lines of evidence: parameter change analysis, correlation studies with language-specific neuron activations, and component-level surgical interventions. The substantial parameter drift, coupled with attention's strong correlation with language-specific neuron activations, indicates that CPT reconfigures how information is routed rather than altering what knowledge is stored in FFN components.

These findings have important implications for efficient language adaptation in LLMs. Our results suggest that targeted adaptation strategies focusing on the attention components in the interface layers could be cost-effective while achieving comparable performance. We leave the empirical validation of these strategies to future work.

Limitations

Our study is limited to the Llama2 and Qwen2.5 model families, specifically focusing on small to medium-sized models (0.5B, 7B to 13B parameters). Future work could extend this analysis to larger models to evaluate whether the observed patterns in layer-wise adaptation and attention-driven language acquisition persist at scale. Additionally, while we demonstrate our findings across language families and scripts, we focus on three representative languages: Irish, Chinese, and Basque. Examining more languages to ensure generalization could be a promising future direction.

Acknowledgements

We would like to acknowledge CloudCIX Limited for the generous collaborative support of computing resources on their NVIDIA HGX/H100 and DGX/A100 GPU cluster. This publication has emanated from research supported in part by grants from Research Ireland under Grant 18/CRT/6223 and 12-RC-2289-P22, the latter being co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 226–245, St. Julian's, Malta. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

Kwesi Adu Cobbina and Tianyi Zhou. 2025. Where to show demos in your prompt: A positional bias of in-context learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29560–29593, Suzhou, China. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Ahmed Elhady, Eneko Agirre, and Mikel Artetxe. 2025. Emergent abilities of large language models under continued pre-training for language adaptation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32174–32186, Vienna, Austria. Association for Computational Linguistics.

Wikimedia Foundation. Wikimedia downloads.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota,

Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Matthieu Futeral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. 2024. moscar: A large-scale multilingual and multimodal document-level corpus. *arXiv preprint arXiv:2406.08707*.

Yiduo Guo, Jie Fu, Huishuai Zhang, and Dongyan Zhao. 2025. Efficient domain continual pretraining by mitigating the stability gap. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32850–32870, Vienna, Austria. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Katharina Hämmerl, Jindřich Libovický, and Alexander Fraser. 2024. Understanding cross-lingual Alignment—A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10922–10943, Bangkok, Thailand. Association for Computational Linguistics.

Chongxuan Huang, Yongshi Ye, Biao Fu, Qifeng Su, and Xiaodong Shi. 2025. From neurons to semantics: Evaluating cross-linguistic alignment capabilities of large language models via neurons alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28956–28974, Vienna, Austria. Association for Computational Linguistics.

J. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability - Vol. 1*, pages 281–297. University of California Press, Berkeley, CA, USA.

Andrea Matarazzo and Riccardo Torlone. 2025. A survey on large language models with some insights on their capabilities and limitations. *Preprint*, arXiv:2501.04040.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Reuse, don't retrain: A recipe for continued pretraining of language models. *Preprint*, arXiv:2407.07263.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S. Yu. 2025. Large language models meet nlp: A survey. *Preprint*, arXiv:2405.12819.

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.

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

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. 2025. Continual learning of large language models: A comprehensive survey. *ACM Comput. Surv.* Just Accepted.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

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

Khanh-Tung Tran, Barry O'Sullivan, and Hoang Nguyen. 2024a. Irish-based large language model with extreme low-resource settings in machine translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 193–202, Bangkok, Thailand. Association for Computational Linguistics.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2024b. *UCCIX: Irish-eXcellence Large Language Model*. IOS Press.

Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2024. Neurons in large language models: Dead, n-gram, positional. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1288–1301, Bangkok, Thailand. Association for Computational Linguistics.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

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

Zheng Xin Yong, Hailey Schoelkopf, Niklas Muennighoff, Alham Fikri Aji, David Ifeoluwa Adelani, Khalid Almubarak, M Saiful Bari, Lintang Sutawika, Jungo Kasai, Ahmed Baruwa, Genta Winata, Stella Biderman, Edward Raff, Dragomir Radev, and Vassilina Nikoulina. 2023. BLOOM+1: Adding language support to BLOOM for zero-shot prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11682–11703, Toronto, Canada. Association for Computational Linguistics.

Chen Zhang, Zhiyuan Liao, and Yansong Feng. 2025. Cross-lingual transfer of cultural knowledge: An asymmetric phenomenon. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 147–157, Vienna, Austria. Association for Computational Linguistics.

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

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

A Sentence Retrieval Layer Swap Results

The results for sentence retrieval accuracy in Tables 6 and 7 along with the SIB200, Belebele and IrishQA results in Tables 8, 9, 10 and 11, provide evidence supporting our hypothesis that CPT adapts to the second language by retargeting the interface layers.

B Attention vs. FFN Weight Change Results

The detailed weight change results in Tables 12 and 13 show that attention layers consistently exhibit larger parameter shifts than FFN layers across training, supporting the attention-driven routing hypothesis in CPT.

C Sentence Retrieval Component Swap Results

The detailed sentence retrieval accuracy for component swapping results in Tables 14 and 15 further support our hypothesis, providing causal evidence for the role of attention in CPT adaptation.

D Qualitative Analysis of Layer Swapping

To provide finer-grained evidence for our layer swapping experiments, we present additional qualitative analysis for L2-13B-GA at the final checkpoint 4280.

D.1 Confusion Matrix Analysis on SIB200

The confusion matrices in Table 16 compare the base, semantic hub swap, and random swap settings on 204 samples from the SIB200 benchmark. Under the semantic hub swap setting, the confusion matrix remains largely intact, with a dominant diagonal structure comparable to the base model. In contrast, under the random swap setting, predictions collapse toward two classes (e and g), accounting for 91.1% of all predictions (126 and 60, respectively, out of 204 samples).

D.2 Sample Analysis on IrishQA

We inspect two IrishQA samples, comparing log-likelihoods across base, semantic hub swap, and random swap settings. The questions, contexts, and choices are shown in Table 17, and the log-likelihood results are reported in Table 18.

In both samples, the semantic hub swap setting achieves a log-likelihood gap between the chosen

prediction and the second-highest prediction of 5.8 for Sample 1 and 0.6 for Sample 2, indicating clear model confidence. In contrast, the random swap setting chose the wrong answer with low confidence in both samples: a gap of 0.193 versus the second-highest prediction for Sample 1 and 0.15 for Sample 2. These gaps are 30 times lower and 18 times lower than that of the base setting for Sample 1 and 2 respectively. This result tells the same story: semantic understanding can be surgically transferred between base and CPT models.

CPT Model	Checkpoint	$\mathbf{Acc}_{\mathbf{Base}}$	$\mathbf{Acc}_{\mathbf{Semantic}}$	$\mathbf{Acc}_{\mathbf{RanAvg}}$
Q2.5-ZH	300	0.2331	0.1892	0.0247
	600	0.4425	0.3839	0.0351
	900	0.4944	0.4482	0.0415
	1200	0.5356	0.4828	0.0444
	1500	0.5485	0.4887	0.0471
	1800	0.5561	0.5079	0.0475
	2100	0.5574	0.5125	0.0486
	2400	0.5602	0.5136	0.0486
	2700	0.5598	0.5168	0.0487
	2944	0.5625	0.5192	0.0486
Q2.5-EN-ZH	300	0.1689	0.1343	0.0226
	600	0.4248	0.3446	0.0371
	900	0.5277	0.4385	0.0433
	1200	0.5843	0.4865	0.0475
	1500	0.6028	0.5243	0.0489
	1800	0.6187	0.5328	0.0512
	2100	0.6243	0.5464	0.0515
	2400	0.6264	0.5481	0.0515
	2700	0.6259	0.5523	0.0519
	3000	0.6321	0.5512	0.0523
	3300	0.6305	0.5513	0.0521
	3494	0.6315	0.5499	0.0525
Q2.5-GA	300	0.1090	0.1122	0.0641
	600	0.1740	0.1644	0.0826
	900	0.2154	0.1989	0.0909
	1200	0.2570	0.2400	0.0994
	1500	0.2925	0.2699	0.1052
	1800	0.3043	0.2835	0.1050
	2100	0.3015	0.2789	0.1036
	2400	0.3067	0.2836	0.1034
	2700	0.3084	0.2866	0.1041
	2854	0.3082	0.2863	0.1038
Q2.5-EN-GA	300	0.0943	0.0968	0.0581
	600	0.1494	0.1411	0.0732
	900	0.2024	0.1799	0.0855
	1200	0.2349	0.2193	0.0938
	1500	0.2610	0.2372	0.0960
	1800	0.2751	0.2486	0.0982
	2100	0.2853	0.2636	0.0973
	2400	0.2912	0.2658	0.0980
	2700	0.2916	0.2645	0.0967
	3000	0.2926	0.2674	0.0968
	3130	0.2930	0.2672	0.0970

Table 6: Sentence retrieval accuracy across checkpoints for CPT models. Results are shown for base, semantic, and random layer-swap configurations.

CPT Model	Checkpoint	$\mathbf{Acc_{Base}}$	$\mathbf{Acc_{Semantic}}$	$\mathbf{Acc_{RanAvg}}$
Q2.5-EU	300	0.1530	0.1568	0.0791
	600	0.1958	0.2029	0.0902
	900	0.2085	0.2158	0.0962
	1200	0.2192	0.2269	0.0967
	1500	0.2203	0.2365	0.0945
	1800	0.2247	0.2416	0.0947
	2100	0.2261	0.2424	0.0947
	2400	0.2244	0.2423	0.0946
	2424	0.2236	0.2421	0.0946
Q2.5-EN-EU	300	0.1075	0.1206	0.0755
	600	0.1609	0.1722	0.0948
	900	0.1950	0.2141	0.1042
	1200	0.2131	0.2364	0.1082
	1500	0.2369	0.2569	0.1125
	1800	0.2539	0.2753	0.1187
	2100	0.2544	0.2751	0.1151
	2400	0.2553	0.2780	0.1139
	2700	0.2550	0.2767	0.1129
	2972	0.2564	0.2787	0.1129
L2-7B-EU	300	0.0866	0.0814	0.0453
	900	0.2459	0.2673	0.0811
	1500	0.2798	0.2819	0.1086
	2100	0.2956	0.2903	0.1198
	2700	0.3358	0.3270	0.1448
	3300	0.0121	0.0123	0.0650
	3900	0.5004	0.4931	0.1437
	4500	0.4524	0.4455	0.1366
	5100	0.0143	0.0138	0.0861
	5700	0.3762	0.3699	0.1327
	6300	0.3897	0.3846	0.1337
	6900	0.3740	0.3698	0.1059
	7500	0.3638	0.3568	0.1361
	8100	0.3467	0.3454	0.1325
	8700	0.3511	0.3465	0.1313
	9300	0.3537	0.3507	0.1301
	9900	0.3497	0.3460	0.1297
	10200	0.3551	0.3492	0.1300
L2-13B-GA	100	0.0801	0.0903	0.0268
	400	0.0742	0.0720	0.1252
	700	0.4968	0.5249	0.1524
	1000	0.6056	0.6217	0.1736
	1300	0.5738	0.5973	0.1705
	1600	0.6138	0.6301	0.1761
	1900	0.0844	0.0817	0.1117
	2200	0.3343	0.3513	0.1316
	2500	0.5183	0.5413	0.1359
	2800	0.5200	0.5521	0.1395
	3100	0.6174	0.6359	0.1494
	3400	0.5778	0.6090	0.1414
	3700	0.5755	0.6072	0.1407
	4000	0.5663	0.6002	0.1401
	4280	0.5713	0.6011	0.1402

Table 7: Sentence retrieval accuracy across checkpoints for CPT models. Results are shown for base, semantic, and random layer-swap configurations.

CPT Model	Checkpoint	$\mathbf{Acc}_{\mathbf{Base}}$	$\mathbf{Acc}_{\mathbf{Semantic}}$	$\mathbf{Acc}_{\mathbf{RanAvg}}$
Q2.5-ZH	300	0.1765	0.1863	0.1137
	600	0.2451	0.1814	0.1393
	900	0.3529	0.2108	0.1775
	1200	0.3529	0.2255	0.1775
	1500	0.3627	0.2206	0.1824
	1800	0.3480	0.2451	0.1941
	2100	0.3676	0.2500	0.2059
	2400	0.3578	0.2353	0.2079
	2700	0.3480	0.2353	0.2059
	2944	0.3578	0.2402	0.2059
Q2.5-EN-ZH	300	0.1961	0.1716	0.0941
	600	0.2549	0.2010	0.1206
	900	0.2990	0.1961	0.1500
	1200	0.3186	0.2255	0.1608
	1500	0.3284	0.2108	0.1696
	1800	0.3284	0.2206	0.1785
	2100	0.2843	0.2157	0.1863
	2400	0.3088	0.2206	0.1951
	2700	0.3235	0.2059	0.1951
	3000	0.3382	0.2010	0.1932
	3300	0.3431	0.2010	0.1922
	3494	0.3382	0.2010	0.1922
Q2.5-GA	300	0.1373	0.1373	0.1137
	600	0.1618	0.1471	0.1275
	900	0.1422	0.1373	0.1226
	1200	0.1667	0.1520	0.1353
	1500	0.1618	0.1471	0.1383
	1800	0.2255	0.1569	0.1392
	2100	0.2010	0.1618	0.1392
	2400	0.1863	0.1569	0.1373
	2700	0.1863	0.1471	0.1383
	2854	0.1912	0.1471	0.1402
Q2.5-EN-GA	300	0.1176	0.1176	0.1009
	600	0.1422	0.1373	0.1166
	900	0.1618	0.1471	0.1176
	1200	0.1716	0.1422	0.1166
	1500	0.1912	0.1520	0.1461
	1800	0.1765	0.1422	0.1314
	2100	0.2059	0.1667	0.1343
	2400	0.1961	0.1618	0.1314
	2700	0.2010	0.1765	0.1363
	3000	0.2010	0.1618	0.1363
	3130	0.2059	0.1618	0.1363

Table 8: SIB200 dataset: Model accuracy across checkpoints. Results are shown for base, semantic, and random layer-swap configurations.

CPT Model	Checkpoint	$\mathbf{Acc_{Base}}$	$\mathbf{Acc_{Semantic}}$	$\mathbf{Acc_{RanAvg}}$
Q2.5-EU	300	0.1667	0.1765	0.1638
	600	0.1667	0.2108	0.1588
	900	0.2157	0.2206	0.1490
	1200	0.1569	0.1863	0.1363
	1500	0.1716	0.1912	0.1569
	1800	0.1814	0.1863	0.1481
	2100	0.1863	0.1814	0.1510
	2424	0.1814	0.1814	0.1500
Q2.5-EN-EU	300	0.1471	0.1422	0.1245
	600	0.1765	0.1618	0.1422
	900	0.1716	0.1618	0.1344
	1200	0.1912	0.1667	0.1491
	1500	0.2010	0.2010	0.1559
	1800	0.2010	0.1863	0.1510
	2100	0.1912	0.1814	0.1353
	2400	0.1961	0.2059	0.1422
	2700	0.1912	0.1961	0.1422
	2972	0.1961	0.2010	0.1412
L2-7B-EU	300	0.7304	0.7353	0.6284
	2100	0.7598	0.7059	0.6304
	3900	0.7402	0.7451	0.5284
	5700	0.7696	0.7108	0.5686
	7500	0.7843	0.6667	0.5578
	9300	0.7843	0.6618	0.5559
	10200	0.7696	0.6667	0.5608
L2-13B-GA	100	0.4020	0.4510	0.1961
	1000	0.7059	0.6667	0.1873
	1900	0.7010	0.6520	0.1745
	2800	0.7108	0.6716	0.2108
	3700	0.6961	0.6716	0.2079
	4280	0.7108	0.6716	0.2098

Table 9: SIB200 dataset: Model accuracy across checkpoints. Results are shown for base, semantic, and random layer-swap configurations.

Model	Checkpoint	$\mathbf{Acc_{Base}}$	$\mathbf{Acc_{Semantic}}$	$\mathbf{Acc_{RanAvg}}$
L2-13B-GA	100	0.3196	0.3073	0.2550
	400	0.3721	0.3631	0.2476
	700	0.4514	0.4458	0.2527
	1000	0.4860	0.4804	0.2606
	1300	0.4804	0.4838	0.2530
	1600	0.4894	0.4916	0.2603
	1900	0.4905	0.4860	0.2644
	2200	0.4659	0.4804	0.2619
	2500	0.4860	0.4760	0.2623
	2800	0.4883	0.4816	0.2677
	3100	0.4749	0.4682	0.2552
	3400	0.4749	0.4793	0.2556
	3700	0.4883	0.4849	0.2606
	4000	0.4860	0.4883	0.2583
	4280	0.4816	0.4838	0.2570
L2-7B-EU	300	0.4324	0.4279	0.3459
	900	0.4514	0.4425	0.3437
	1500	0.4078	0.4156	0.3234
	2100	0.4447	0.4603	0.3511
	2700	0.4156	0.4313	0.3466
	3300	0.2670	0.2626	0.2742
	3900	0.4436	0.4313	0.2856
	4500	0.4112	0.4123	0.2932
	5100	0.3821	0.3866	0.2932
	5700	0.4346	0.4101	0.3012
	6300	0.4246	0.4078	0.3111
	6900	0.4000	0.4067	0.2961
	7500	0.4000	0.4101	0.3028
	8100	0.3944	0.4112	0.2959
	8700	0.3978	0.4022	0.2939
	9300	0.3955	0.3989	0.2930
	9900	0.3866	0.3899	0.2947
	10200	0.3855	0.3922	0.2954

Table 10: Belebele dataset: Model accuracy across checkpoints for L2-13B-GA and L2-7B-EU. Results are shown for base, semantic, and random layer-swap configurations.

Model	Checkpoint	$\mathbf{Acc_{Base}}$	$\mathbf{Acc_{Semantic}}$	$\mathbf{Acc_{RanAvg}}$
L2-13B-GA	100	0.7873	0.7975	0.2744
	400	0.7495	0.8253	0.2410
	700	0.9034	0.9215	0.2476
	1000	0.9165	0.9367	0.2532
	1300	0.9089	0.9215	0.2471
	1600	0.9266	0.9342	0.2370
	1900	0.9114	0.9215	0.2380
	2200	0.8886	0.9114	0.2380
	2500	0.9089	0.9190	0.2466
	2800	0.9266	0.9114	0.2531
	3100	0.9342	0.9089	0.2522
	3400	0.9342	0.9215	0.2501
	3700	0.9316	0.9165	0.2496
	4000	0.9291	0.9063	0.2516
	4280	0.9215	0.9089	0.2527

Table 11: IrishQA dataset: Model accuracy across checkpoints for L2-13B-GA. Results are shown for base, semantic, and random layer-swap configurations.

CPT Model	Checkpoint	Attention	FFN	ΔAttentionFFN
Q2.5-ZH	300	0.001277	0.001135	0.000141
	600	0.002162	0.001930	0.000232
	900	0.002716	0.002421	0.000295
	1200	0.003087	0.002743	0.000344
	1500	0.003342	0.002970	0.000372
	1800	0.003514	0.003117	0.000397
	2100	0.003615	0.003202	0.000413
	2400	0.003659	0.003243	0.000416
	2700	0.003676	0.003256	0.000420
	2944	0.003676	0.003257	0.000419
Q2.5-EN-ZH	300	0.001166	0.001036	0.000130
	600	0.002012	0.001793	0.000219
	900	0.002547	0.002265	0.000282
	1200	0.002924	0.002596	0.000328
	1500	0.003204	0.002839	0.000365
	1800	0.003407	0.003018	0.000389
	2100	0.003559	0.003145	0.000414
	2400	0.003656	0.003225	0.000431
	2700	0.003710	0.003268	0.000442
	3000	0.003734	0.003291	0.000443
	3300	0.003740	0.003295	0.000445
	3494	0.003740	0.003295	0.000445
Q2.5-GA	300	0.001169	0.000916	0.000253
	600	0.001814	0.001360	0.000454
	900	0.002253	0.001680	0.000573
	1200	0.002565	0.001912	0.000653
	1500	0.002779	0.002075	0.000704
	1800	0.002915	0.002179	0.000736
	2100	0.002989	0.002236	0.000753
	2400	0.003021	0.002261	0.000760
	2700	0.003031	0.002266	0.000765
	2854	0.003031	0.002266	0.000765
Q2.5-EN-GA	300	0.001109	0.000886	0.000223
	600	0.001733	0.001323	0.000410
	900	0.002164	0.001640	0.000524
	1200	0.002480	0.001880	0.000600
	1500	0.002711	0.002055	0.000656
	1800	0.002865	0.002174	0.000691
	2100	0.002961	0.002248	0.000713
	2400	0.003012	0.002288	0.000724
	2700	0.003036	0.002305	0.000731
	3000	0.003042	0.002309	0.000733
	3130	0.003042	0.002309	0.000733

Table 12: Attention weight change comparison with FFN weight change.

CPT Model	Checkpoint	Attention	FFN	ΔAttentionFFN
Q2.5-EU	300	0.001290	0.001048	0.000242
	600	0.002017	0.001552	0.000465
	900	0.002480	0.001885	0.000595
	1200	0.002784	0.002106	0.000678
	1500	0.002964	0.002240	0.000724
	1800	0.003058	0.002311	0.000747
	2100	0.003094	0.002334	0.000760
	2400	0.003098	0.002340	0.000758
	2424	0.003098	0.002340	0.000758
Q2.5-EN-EU	300	0.001129	0.000971	0.000158
	600	0.001799	0.001462	0.000337
	900	0.002253	0.001799	0.000454
	1200	0.002574	0.002044	0.000530
	1500	0.002798	0.002217	0.000581
	1800	0.002944	0.002329	0.000615
	2100	0.003028	0.002394	0.000634
	2400	0.003071	0.002425	0.000647
	2700	0.003085	0.002435	0.000650
	2972	0.003085	0.002435	0.000650
L2-7B-EU	300	0.000286	0.000277	0.000008
	900	0.000976	0.000895	0.000081
	1500	0.001820	0.001607	0.000213
	2100	0.002346	0.002126	0.000220
	2700	0.002806	0.002582	0.000224
	3300	0.003022	0.002822	0.000200
	3900	0.003461	0.003256	0.000205
	4500	0.003812	0.003607	0.000205
	5100	0.004057	0.003857	0.000200
	5700	0.004242	0.004042	0.000200
	6300	0.004365	0.004176	0.000189
	6900	0.004459	0.004266	0.000194
	7500	0.004531	0.004334	0.000197
	8100	0.004599	0.004393	0.000207
	8700	0.004636	0.004421	0.000215
	9300	0.004656	0.004439	0.000217
	9900	0.004663	0.004445	0.000219
	10200	0.004664	0.004445	0.000219
L2-13B-GA	100	0.000301	0.000255	0.000046
	400	0.002323	0.002024	0.000299
	700	0.002822	0.002381	0.000441
	1000	0.003212	0.002658	0.000554
	1300	0.003629	0.002998	0.000631
	1600	0.003983	0.003305	0.000678
	1900	0.004309	0.003645	0.000664
	2200	0.004485	0.003813	0.000672
	2500	0.004649	0.004010	0.000639
	2800	0.004752	0.004124	0.000628
	3100	0.004804	0.004184	0.000620
	3400	0.004834	0.004225	0.000609
	3700	0.004848	0.004244	0.000604
	4000	0.004850	0.004249	0.000601
	4280	0.004850	0.004251	0.000599

Table 13: Attention weight change comparison with FFN weight change.

CPT Model	Checkpoint	$\mathbf{Acc_{Base}}$	$\mathbf{Acc_{Attention}}$	$\mathbf{Acc_{FFN}}$
Q2.5-ZH	300	0.2331	0.1388	0.2437
	600	0.4425	0.2311	0.4655
	900	0.4944	0.2620	0.5219
	1200	0.5356	0.2796	0.5647
	1500	0.5485	0.2578	0.5752
	1800	0.5561	0.2580	0.5847
	2100	0.5574	0.2536	0.5869
	2400	0.5602	0.2497	0.5911
	2700	0.5598	0.2499	0.5894
	2944	0.5625	0.2512	0.5919
Q2.5-EN-ZH	300	0.1689	0.0985	0.1803
	600	0.4248	0.2313	0.4517
	900	0.5277	0.2719	0.5630
	1200	0.5843	0.3047	0.6185
	1500	0.6028	0.2999	0.6390
	1800	0.6187	0.3001	0.6573
	2100	0.6243	0.2950	0.6630
	2400	0.6264	0.2806	0.6625
	2700	0.6259	0.2765	0.6661
	3000	0.6321	0.2826	0.6701
	3300	0.6305	0.2796	0.6681
	3494	0.6315	0.2802	0.6678
Q2.5-GA	300	0.1090	0.0696	0.1121
	600	0.1740	0.0966	0.1756
	900	0.2154	0.1120	0.2170
	1200	0.2570	0.1217	0.2617
	1500	0.2925	0.1262	0.2961
	1800	0.3043	0.1278	0.3061
	2100	0.3015	0.1247	0.3040
	2400	0.3067	0.1239	0.3074
	2700	0.3084	0.1237	0.3115
	2854	0.3082	0.1240	0.3109
Q2.5-EN-GA	300	0.0943	0.0646	0.0979
	600	0.1494	0.0875	0.1512
	900	0.2024	0.1080	0.2028
	1200	0.2349	0.1207	0.2380
	1500	0.2610	0.1184	0.2619
	1800	0.2751	0.1210	0.2731
	2100	0.2853	0.1223	0.2878
	2400	0.2912	0.1199	0.2935
	2700	0.2916	0.1179	0.2918
	3000	0.2926	0.1184	0.2947
	3130	0.2930	0.1192	0.2939

Table 14: Sentence retrieval accuracy across checkpoints for CPT models. Results are shown for base, attention, and FFN component-swap configurations.

CPT Model	Checkpoint	$\mathbf{Acc_{Base}}$	$\mathbf{Acc_{Attention}}$	$\mathbf{Acc_{FFN}}$
Q2.5-EU	300	0.1530	0.1090	0.1534
	600	0.1958	0.1268	0.1973
	900	0.2085	0.1378	0.2088
	1200	0.2192	0.1357	0.2216
	1500	0.2203	0.1336	0.2242
	1800	0.2247	0.1320	0.2286
	2100	0.2261	0.1344	0.2304
	2400	0.2244	0.1323	0.2292
	2424	0.2236	0.1318	0.2288
Q2.5-EN-EU	300	0.1075	0.0826	0.1126
	600	0.1609	0.1017	0.1656
	900	0.1950	0.1110	0.2018
	1200	0.2131	0.1179	0.2231
	1500	0.2369	0.1232	0.2442
	1800	0.2539	0.1317	0.2633
	2100	0.2544	0.1266	0.2625
	2400	0.2553	0.1232	0.2648
	2700	0.2550	0.1224	0.2643
	2972	0.2564	0.1211	0.2665
L2-7B-EU	300	0.0866	0.0809	0.0869
	900	0.2459	0.2305	0.2676
	1500	0.2798	0.3254	0.2959
	2100	0.2956	0.3454	0.3155
	2700	0.3358	0.3560	0.3528
	3300	0.0121	0.0818	0.0122
	3900	0.5004	0.4190	0.5030
	4500	0.4524	0.4127	0.4571
	5100	0.0143	0.3799	0.0138
	5700	0.3762	0.3500	0.3855
	6300	0.3897	0.3454	0.3969
	6900	0.3740	0.3319	0.3840
	7500	0.3638	0.3397	0.3763
	8100	0.3467	0.3311	0.3611
	8700	0.3511	0.3276	0.3646
	9300	0.3537	0.3264	0.3679
	9900	0.3497	0.3263	0.3637
	10200	0.3551	0.3253	0.3665
L2-13B-GA	100	0.0801	0.1366	0.1296
	400	0.0742	0.1794	0.0212
	700	0.4967	0.1936	0.1238
	1000	0.6056	0.1874	0.1854
	1300	0.5738	0.1833	0.1557
	1600	0.6138	0.1715	0.1885
	1900	0.0844	0.1257	0.0303
	2200	0.3343	0.1065	0.0868
	2500	0.5183	0.1079	0.1221
	2800	0.5200	0.1038	0.3060
	3100	0.6174	0.1000	0.2512
	3400	0.5778	0.0908	0.2649
	3700	0.5755	0.0907	0.2462
	4000	0.5663	0.0926	0.2528
	4280	0.5713	0.0922	0.2545

Table 15: Sentence retrieval accuracy across checkpoints for CPT models. Results are shown for base, attention, and FFN component-swap configurations.

Table 16: Confusion matrix for L2-13B-GA at checkpoint 4280 on SIB200 (204 samples).

Base

	a	b	c	d	e	f	g	Total
a	5	0	1	2	4	1	6	19
b	0	6	1	1	7	0	2	17
c	0	0	15	0	6	0	1	22
d	0	0	0	26	3	0	1	30
e	1	0	7	0	40	0	3	51
f	2	1	2	0	0	18	2	25
g	0	0	0	1	3	1	35	40
Total	8	7	26	30	63	20	50	204

Semantic Hub Swap

	a	b	c	d	e	f	g	Total
a	6	0	0	0	5	2	6	19
b	0	1	0	0	10	0	6	17
c	2	0	2	0	13	2	3	22
d	1	0	0	27	1	0	1	30
e	6	0	0	0	43	0	2	51
f	1	1	0	0	0	20	3	25
g	0	0	0	0	1	1	38	40
Total	16	2	2	27	73	25	59	204

Random Swap

	a	b	c	d	e	f	g	Total
a	0	0	0	0	15	1	3	19
b	0	0	0	0	9	1	7	17
c	0	0	0	0	14	1	7	22
d	0	0	0	0	17	5	8	30
e	0	0	0	1	33	2	15	51
f	0	0	0	0	17	0	8	25
g	0	0	0	0	21	7	12	40
Total	0	0	0	1	126	17	60	204

	Sample 1	Sample 2
Question	Cad í an phríomhchreideamh in Éirinn? *(What is the predominant religion in Ireland?)*	Cad é córas grádaithe scrúduithe na hArdteistiméireachta? *(What is the Leaving Certificate examination grading system?)*
Context	Is í an Chríostaíocht an príomhreiligiún in Éirinn. *(Christianity is the predominant religion in Ireland.)*	Tá 8 ngrád ar scála scrúdaithe na hArdteistiméireachta. Is é Grád 1 an grád is airde, agus is é Grád 8 an grád is ísle. Tháinig an scála grádaithe 8 bpointe in ionad an scála 14 bpointe ag an dá leibhéal Ardteistiméireachta agus Gnáthleibhéil in 2017. Roinntear marcanna idir 100% agus 30% i seacht mbanna grád (1-7). Tá gach banna 10% ar leithead. Soláthraíonn an córas grádaithe seo bealach soiléir agus struchtúrtha chun feidhmíocht na ndaltaí a mheas, rud a chabhraíonn leo tuiscinta fháil ar a gcuid buanna acadúla agus réimsí le haghaidh feabhsúcháin. *(The Leaving Certificate examination scale has 8 grades, the highest grade is Grade 1 and the lowest grade is Grade 8. The 8-point grading scale replaced the 14-point scale at both Higher and Ordinary levels in 2017. Marks between 100% and 30% are divided into seven grade bands (1–7). Each band is 10% wide.)*
Choices	(A) Ioslam *(Islam)* (B) Búdachas *(Buddhism)* (C) Críostaíocht *(Christianity)* (D) Hiondúchas *(Hinduism)*	(A) 10 ngrád ó A go J *(10 grades A–J)* (B) 12 ghrád ó 1 go 12 *(12 grades 1–12)* (C) 14 ghrád ó A1 go H8 *(14 grades A1–H8)* (D) 8 ngrád ó 1 go 8 *(8 grades 1–8)*
Answer	**C**	**D**

Table 17: IrishQA samples used for qualitative analysis.

Setting	Sample 1				Sample 2			
	Pred.	**LL**	**LL (2nd)**	**Gap**	**Pred.**	**LL**	**LL (2nd)**	**Gap**
Base	C	−0.010	−5.810	5.800	D	−0.095	−2.780	2.685
Semantic Swap	C	−0.014	−5.814	5.800	D	−0.710	−1.310	0.600
Random Swap	A	−1.306	−1.499	0.193	B	−1.220	−1.370	0.150

Table 18: Log-likelihood of predicted answer (LL), log-likelihood of second-highest scoring option (LL (2nd)), and their gap (Gap) for both IrishQA samples across the three settings.