The Korean Society for Language and Information
https://doi.org/10.29403/LI.29.3.9

# Do Korean-Adapted LLMs Think in Korean? Analyzing Latent Language and the Preservation of Korean-Specific Knowledge

**Sangah Lee**[†]
**Seoul National University**

**Sangah Lee. 2025. Do Korean-Adapted LLMs Think in Korean? Analyzing Latent Language and the Preservation of Korean-Specific Knowledge.** *Language and Information* 29.3, 229-256. Language models trained on English-dominant corpora may not process other languages directly, but rather translate them internally before performing a task. This study investigates whether Korean-adapted large language models (LLMs) genuinely "think" in Korean by analyzing their latent language, the internal linguistic medium that emerges during task-solving. We find that models exposed to more Korean data use Korean tokens more frequently in latent sequences and translate into Korean at earlier representational stages. Using the KoBALT benchmark, we link these internal traces to model performance across linguistic subfields. Models that rely on non-Korean latent languages tend to lose Korean-specific cues, particularly in morphology and phonology, which leads to lower accuracy. In contrast, models with continued pre-training and richer Korean adaptation better preserve Korean-specific knowledge in internal representations. These results suggest that linguistic adaptation shapes not only how LLMs produce Korean text, but also how they internally structure and mediate Korean linguistic input, offering empirical insight into the boundary between translation and representation in multilingual language models.

**Key words:** latent language, internal representations, large language model, Korean-specific knowledge, linguistic adaptation

[†] (08826), Assistant Professor, Department of Linguistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, Korea, E-mail: sanalee@snu.ac.kr

## 1. Introduction

In recent years, the development and application of large language models (LLMs) have accelerated rapidly, and many pre-trained models are now employed for Korean language processing (e.g., LLaMA (Touvron et al., 2023), Qwen (Bai et al., 2023), Gemma (Gemma Team, 2024)). However, because these models were primarily trained on datasets dominated by non-Korean languages such as English or Chinese, they often lack Korean-specific cultural and linguistic knowledge and therefore exhibit limited performance on Korean-focused evaluations (Kim et al., 2024; Kim and Lee, 2025; Shin et al., 2025). While such models may produce fluent Korean text, their outputs do not necessarily align with users' intended meanings or expectations.

Accordingly, there is an increasing demand for models that can better understand and process the Korean language. One approach is to pre-train models from scratch using Korean data (LG AI Research, 2024; NAVER Cloud HyperCLOVA X Team, 2025; Kanana LLM Team, 2025), but this requires massive datasets and computational resources, making it difficult to scale. Consequently, many Korean-capable models have instead been developed by adapting existing large models through post-training methods such as continued pre-training (CPT), supervised fine-tuning (SFT), or alignment tuning, using the models as backbones.[1)2)3)]

However, even among models adapted for Korean through various post-training strategies, performance varies widely across benchmarks and linguistic domains. This variation raises important questions about how these models internally process Korean prompts beyond surface-level fluency. One way to approach this is to analyze the internal language that models employ during intermediate task-solving processes. Recent studies have adopted the Logit Lens technique, which decodes intermediate layer embeddings into token distributions, thereby revealing layer-wise linguistic representations (Wendler et al., 2024; Wang et al., 2025; Zhong et al., 2025). The linguistic form that emerges within these internal computations is referred to as the latent language.

Previous studies have shown that English-centric models such as LLaMA tend to perform tasks through a latent language close to English (Wendler et al., 2024), transforming internal representations into the target (prompted) language only at the output layer. Zhong et al. (2025) further demonstrated that the latent language of a model can vary according to the linguistic composition of its pre-training corpus. Building on these findings, we examine

1) https://huggingface.co/NCSOFT/Llama-VARCO-8B-Instruct
2) https://huggingface.co/yanolja/KoSOLAR-10.7B-v0.2
3) https://huggingface.co/skt/A.X-4.0-Light

whether models adapted from non-Korean-centric architectures to Korean exhibit a corresponding shift in their latent language, for instance, from English to Korean. Even when such models appear proficient in generating Korean text after adaptation, they may still process information internally through a latent language resembling English, Chinese, or another non-Korean language, translating the results into Korean only at the surface level. This may cause information loss in tasks that require genuine Korean linguistic reasoning. For instance, when asked whether the Korean verb '뛰놀다ttwinolta' (뛰다ttwita + 놀다nolta, "to run and play") is a compound word, a model that processes the input in English may reduce the verb to "frolic", thereby losing the morphological cues necessary to identify its compound structure.

Accordingly, we formulate the following three research questions:

**RQ1. What are the latent languages of different Korean-capable models?**
We identify the languages that emerge as latent languages across multiple Korean-capable models trained on datasets with varying linguistic compositions.

**RQ2. Does the latent language change when Korean data are injected, and at which training stage does this change occur most effectively?**
We analyze whether adaptation with Korean data shifts the latent language toward Korean, comparing models that incorporate Korean data at different post-training stages.

**RQ3. How do models with different latent languages process linguistically sensitive Korean tasks?**
We compare task-solving behaviors across models with distinct latent language patterns to examine how internal processing languages affect Korean-specific performance.

To address these research questions, we design two complementary experimental procedures. For RQ1 and RQ2, we extend the dataset inroduced by Zhong et al. (2025), which consists of simple fill-in-the-blank problems, by adding Korean prompts and adjusting the observation protocol to fit our analysis of latent languages. For RQ3, we sample and utilize questions from KoBALT (Shin et al., 2025), a benchmark that evaluates Korean linguistic knowledge across diverse subfields. Collectively, these datasets enable us to examine the latent language behaviors of various LLMs and to assess how their internal representations influence performance on linguistically grounded Korean tasks.

Through these three experiments, we reaffirm that the language distribution within the training dataset significantly influences a model's latent language. We further observe that when models with different dominant latent languages (e.g., English or Chinese) are adapted

to Korean through post-training, simple supervised fine-tuning (SFT) has only a limited effect, whereas continued pre-training (CPT) followed by SFT induces more substantial shifts, depending on the model. In particular, the A.X-4.0-Light model (SKT AI Model Lab), adapted from Qwen 2.5 through CPT, SFT, and Direct Preference Optimization (DPO), exhibits a clear transition of its latent language toward Korean. This model also achieves higher performance on the KoBALT benchmark than models whose latent languages remain non-Korean. Overall, these findings suggest that sufficiently extensive Korean-based training can reshape a model's latent language and that internal processing in Korean contributes to improved performance in linguistically grounded tasks.

## 2. Backgrounds

### 2.1. Post-Training of LLMs

The training of large language models (LLMs) typically involves two main stages: (1) pre-training, where the model acquires general linguistic knowledge through next-word prediction, and (2) instruction tuning, a supervised fine-tuning (SFT) stage that enables it to follow human instructions (Wei et al., 2022). After pre-training, SFT functions as one of the primary post-training strategies. The data used for SFT typically consist of triplets of instructions, inputs, and desired outputs. Most publicly available models are released in two forms: a pre-trained version and an instruction-tuned version (e.g., Llama-3-8B and Llama-3-8B-Instruct, respectively). In many cases, English-centric or multilingual pre-trained models are further adapted to Korean by applying SFT with Korean instruction datasets.[4)5)6)]

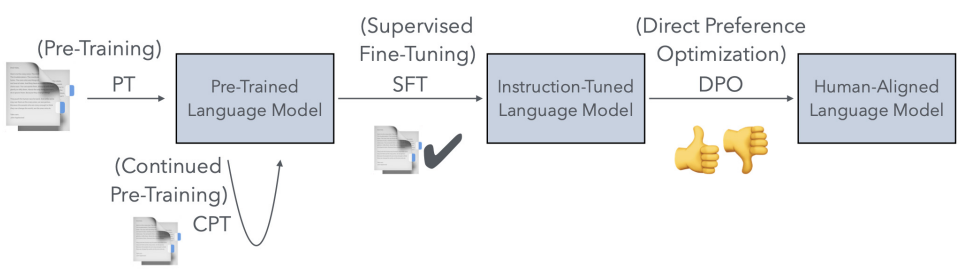
In some cases, continued pre-training (CPT) is applied by feeding additional data into a pre-trained model and continuing the next word prediction training. When Korean data are used for CPT, this process helps the model acquire richer Korean linguistic knowledge. Since the resulting model after CPT remains a pre-trained model, it is typically followed by Korean-language SFT to further adapt it for instruction following.[7)8)]

After SFT, many models undergo an additional stage known as alignment tuning, which aims to align the model's generation behavior with human preferences. One widely adopted

4) https://huggingface.co/hometax/sapie-gemma2-9B-IT
5) https://huggingface.co/AIDX-ktds/ktdsbaseLM-v0.12-based-on-openchat3.5
6) https://huggingface.co/yanolja/KoSOLAR-10.7B-v0.2
7) https://huggingface.co/Saxo/Linkbricks-Horizon-AI-Korean-Advanced-27B
8) https://huggingface.co/NCSOFT/Llama-VARCO-8B-Instruct

approach for this is Direct Preference Optimization (DPO) (Rafailov et al., 2023). DPO is a preference-based tuning method inspired by reinforcement learning, where the model is trained on pairs of responses to the same query, one labeled as 'chosen' and the other as 'rejected' by human annotators. Through this process, the model learns to produce outputs that resemble those preferred ('chosen') by humans, thereby aligning its behavior more closely with human values.

The diagram in Figure 1 illustrates the typical training sequence of an LLM: pre-training from scratch (PT), continued pre-training (CPT), supervised fine-tuning (SFT), and direct preference optimization (DPO). Following this pipeline, our study analyzes latent language patterns across three variants of the same backbone model: (1) a baseline model fine-tuned in English, (2) a model fine-tuned in Korean, and (3) a model that undergoes CPT with Korean data followed by Korean SFT. This comparison corresponds to RQ2.




**<Figure 1> Overview of the LLM training pipeline from PT to CPT, SFT, and DPO**

## 2.2. Detecting Latent Languages of LLMs

Mechanistic interpretability research offers several approaches for investigating the internal processing of language models. Some studies identify specific circuits (Wang et al., 2022; Conmy et al., 2023) or neurons (Song et al., 2024; Tang et al., 2024) within language models that govern particular tasks or languages, and then manipulate their activation to observe the resulting changes in model behavior. Other studies employ sparse autoencoders (SAEs) to decompose a model's complex internal representations into a sparse set of features, enabling the discovery of interpretable and meaningful components (Cunningham et al., 2023; Lieberum et al., 2024).

Another line of research has focused on observing a model's latent language by decoding

intermediate layer embeddings using techniques such as the Logit Lens or Tuned Lens (Nostalgebraist, 2020; Belrose et al., 2023). Wendler et al. (2024) applied this approach to the English-centric multilingual LLaMA-2 model, prompting it with simple repetition and cloze tasks to examine the latent language used across layers. Their analysis proposed that, within the model, representations progress through three stages as they move from lower to higher layers: an input space, a concept space, and an output space. While the input and output spaces are aligned with language tokens in their respective modalities, the concept space refers to a more abstract level of representation that is not bound to any specific natural language. The concepts in this space, however, were found to be biased toward English compared to other languages.

Wang et al. (2025) conducted similar experiments using the English-centric LLaMA-2 model and the multilingual BLOOM model. The English-centric model exhibited a dominance of English tokens in the middle-to-upper layers, indicating that its concept space was primarily English-centric. In contrast, the multilingual model showed a more diverse linguistic composition at the same layers, with a mixture of English, French, Spanish, and other languages, suggesting a shared concept space not aligned with any single language.

Schut et al. (2025) likewise investigated latent language representations in both English-centric models (LLaMA-3.1, Gemma) and multilingual models (Aya-23, Mixtral-8x). Using next-word prediction tasks prompted in French, German, Dutch, and Mandarin, they found that LLMs tend to first produce representations closer to English before mapping them to the target language. Furthermore, lexical (semantically loaded) items such as nouns and verbs were frequently routed through English, whereas function words belonging to other parts of speech (e.g., adpositions, determiners) were rarely mediated by English.

Zhong et al. (2025) aimed to investigate the latent languages of non-English-centric models. They trained relatively small models under two conditions: (1) continued pre-training with a non-English target language, and (2) pre-training on a balanced multilingual corpus. When these models were prompted to perform a cloze task and their latent languages were examined, the non-English-centric models were focused to dynamically select one latent language among several, depending on the similarity between the latent language and the target language.

Similarly, we observe the intermediate processing of models by unembedding their intermediate-layer embeddings using the Logit Lens. While we build upon the dataset introduced by Zhong et al. (2025), we employ Korean-translated prompts and adopt a slightly different observation method. Instead of computing the generation probability of predetermined answers for each language, we allow the model to perform free generation and then examine the language of the produced tokens. In addition, we sample questions from the KoBALT benchmark, convert them into cloze-style tasks, and apply the same method to

compare the behaviors of models that exhibit different latent languages.

Ozaki et al. (2025) examined translation and geo-cultural reasoning tasks and reported that aligning a model's latent language with the input or output language does not necessarily lead to optimal downstream performance. However, both of these tasks represent cases where internal processing in a non-target latent language does not cause the task itself to fail. In contrast, the linguistic benchmark tasks used in our study include cases where the latent language shifts and the model's "thinking" language performs internal translation, resulting in the loss or distortion of information contained in the original problem.

## 3. RQ1. Latent Languages of Korean-Capable Models

### 3.1. Experiment

To investigate the latent languages that emerge in Korean-capable models, we conduct layer-wise analyses using the Logit Lens method. We examine several LLMs that are capable of generating Korean but differ in the linguistic composition of their pre-training data. The selected models are as follows: **Llama-3.1-8B-Instruct (Llama)** is an English-centric model trained primarily on English data, with limited multilingual exposure (Llama Team, 2024); **EXAONE-3.5-7.8B-Instruct (EXAONE)** is a bilingual model pre-trained on a mixture of English and Korean data (LG AI Research, 2024); **Qwen-2.5-7B-Instruct (Qwen)** is a multilingual model pre-trained on data covering more than 29 languages, including Chinese, English, French, Spanish, and Korean, with particularly strong performance in English and Chinese (Qwen, 2024); **Aya-expanse-8B (Aya)** is a balanced multilingual model pre-trained on 23 languages, including Arabic, Chinese, English, and Korean (Dang et al., 2024).

All models are of comparable size (7B-8B parameters) and are used in their instruction-tuned versions, which are capable of understanding and following natural language instructions. Each model has undergone pre-training and subsequent supervised fine-tuning (SFT) in the respective languages of its dataset.

For the dataset, we modify the cloze test dataset introduced by Zhong et al. (2025), which consists of 166 simple fill-in-the-blank problems. The original dataset includes questions and answers in several languages such as English, Chinese, Japanese, and French. In this study, we translate the entire set into Korean for use in our experiments. The initial translation is performed using the gpt-4.1-2025-04-14 model API, followed by manual review and correction to ensure linguistic accuracy and consistency.

An example of an input prompt, including the instruction and two demonstrations, is provided below (shown in English translation as Example (1)).

(1) Please fill in the blank "__".
"__" is when two people officially become a married couple. Answer: "marriage"
"__" is a device used for communication. Answer: "telephone"
"__" is a diagram that shows a location. Answer:

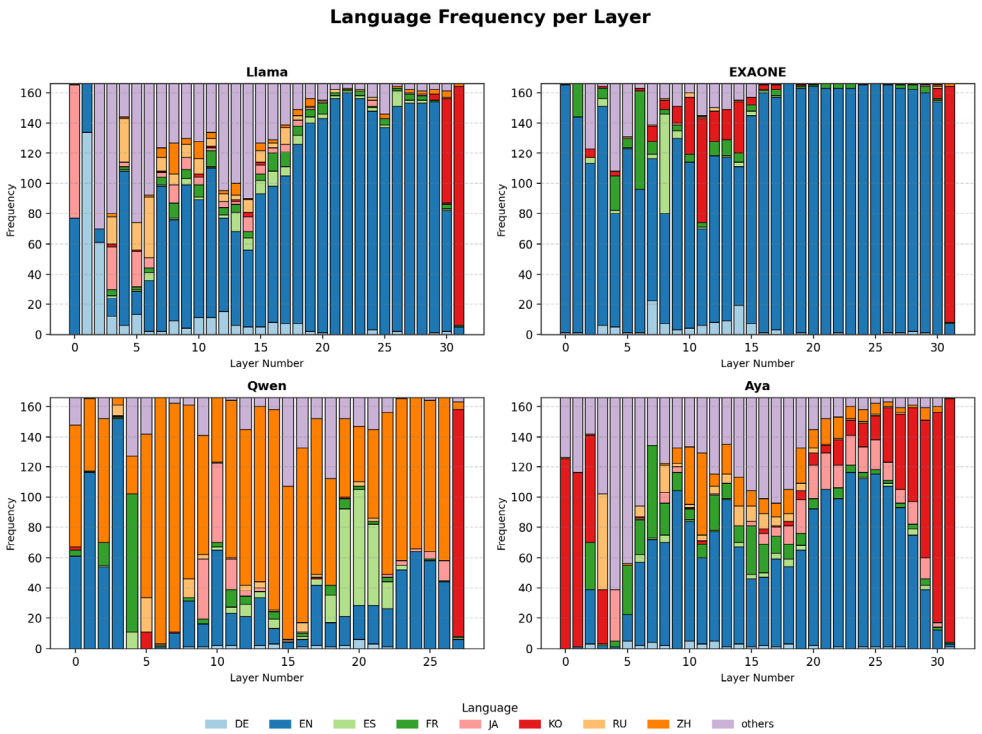
Following previous studies, we adopt a 2-shot setting, providing two examples per prompt to prevent cases where incorrect answers would make observation impossible. Model generation is performed using beam search with one beam and a generation length of five tokens. Although the correct answers are not always exactly five tokens long, we specify free generation up to five tokens to accommodate answers of varying lengths across different questions.

While Zhong et al. (2025) predefined translated answers in each candidate latent language and computed their layer-wise generation probabilities, this approach does not account for the possibility that models may produce surface forms with similar meanings but different lexical realizations. Therefore, we instead adopt a free-generation setting and detect which language the generated output most closely resembles.

We decode the representations at each layer using the Logit Lens method, then perform language identification by classifying the decoded output into ISO-639-1 language codes using the gpt-4.1-2025-04-14 model API. This allows to observe the dominant latent language in which each model "thinks" at each layer. Although the GPT model may produce classification errors and tends to force ambiguous early-layer outputs into specific language categories, we select it as it offers the most robust language identification performance among currently available methods.

### 3.2. Results

Figure 2 presents the distribution of latent languages across layers for each model on the 166 cloze test problems described above. In each subplot, the x-axis represents the model layers, and the y-axis represents the proportion of each latent language observed across the 166 problems. The language that appear with high frequency in the four models, including German (DE), English (EN), Spanish (ES), French (FR), Japanese (JA), Korean (KO), Russian (RU), and Chinese (ZH), are defined as the candidate latent languages, while all other detected languages are grouped under "others."




**<Figure 2> Layer-wise distributions of latent languages across four Korean-capable models on 166 cloze task problems**

We first note that in all models, up to approximately layer 20, that is, in the lower to middle layers, the generated outputs often do not correspond to tokens in any identifiable language and are generally not semantically meaningful as strings. For example, outputs such as "ㅎ Verm gr fr fr" or "ICT통계isticsMV라" are frequently observed. In some cases, code snippets, function names, or repetitions of prompts and in-context examples are also generated. After these layers, the middle to upper layers begin to produce semantically meaningful tokens that often resemble or match the correct answers.

In all four models, the final layer produces answers in the target language, Korean (KO). However, the dominant latent language in the preceding layers differs across models. In both Llama and EXAONE, English (EN) is highly dominant in the middle to upper layers. This is expected for Llama, which is an English-centric model. Suprisingly, in the case of EXAONE, whose training data consist of roughly equal proportions of English and Korean, English still appears overwhelmingly dominant as the latent language, surpassing both Korean and other

languages.

Although Qwen is described as supporting multilingual capabilities, its technical report appears to emphasize performance primarily in Chinese and English. Similarly, in our analyis, Chinese (ZH) emerges as the dominant latent language, with English (EN) also appearing to a lesser extent. In contrast, Aya exhibits a more diverse distribution of latent languages, including English, Korean, Japanese, Chinese, and others, with Korean appearing in relatively earlier layers as well. (Note that the Korean tokens in layers 0 to 4 are likely misclassifications of GPT.) This pattern in Aya aligns with findings reported by Wang et al. (2025), in which multilingual models demonstrate more diverse latent language behaviors.

Taken together, the results indicate that the linguistic composition of the training data, especially the language that dominates it, may largely determine which latent language emerges during internal processing. Building on this observation, the next section (RQ2) investigates whether such latent language tendencies can be shifted through post-training adaptation using Korean data, and how the timing of this adaptation affects the degree of change.

## 4. RQ2. Effects of Korean Data Injection and Training Stage

### 4.1. Experiment

To examine whether a model's latent language can shfit through adaptation with Korean data and to determine at which stage this shift occurs most effectively, we compare models trained with different adaptation strategies. Following the pre-training → CPT → SFT → DPO pipeline, we focus on three configurations of the same backbone model: **Original-SFT** is a model fine-tuned only in its primary language (English or Chinese), without the addition of Korean data; **KO-SFT** is a model fine-tuned in Korean without additional pre-training; **KO-CPT-SFT** is a model that underwent continued pre-training (CPT) with Korean data prior to Korean SFT.

This experimental design enables us to disentangle the effect of the training stage at which Korean data are incorporated on the model's latent language dynamics. As in RQ1, we select models with comparable sizes for which all three configurations are available based on the same foundation model. Using the same cloze task dataset and Logit Lens-based analysis method described in Section 3, we observe the latent language distributions of these models. The selected models are presented in Table 1.

**<Table 1> Selected models for RQ2 experiments and their training configurations**

| backbone model | Original-SFT | KO-SFT | KO-CPT-SFT |
|---|---|---|---|
| Llama-3.1-8B | Llama-3.1-8B-Instruct | llama3.1_korean_v1.1_sft_by_aidx | Llama-VARCO-8B-Instruct |
| Qwen2.5-7B | Qwen2.5-7B-Instruct | Qwen2.5-7B-Instruct-kowiki-qa | A.X-4.0-Light |
| SOLAR-10.7B-v1.0 | SOLAR-10.7B-Instruct-v1.0 | KoSOLAR-10.7B-v0.2 | OPEN-SOLAR-KO-10.7B-S-Core |

Using Llama-3.1-8B as a backbone model, the Original-SFT model is fine-tuned primarily on English data with a small amount of multilingual data. The KO-SFT model is fine-tuned on Korean data starting from the foundation model.[9] The KO-CPT-SFT model first undergoes continued pre-trained on a mixed Korean-English corpus, followed by Korean fine-tuning and subsequent DPO using Korean data.

Using Qwen-2.5-7B as the backbone, the Original-SFT model is a multilingual fine-tuned model whose training data are assumed to be dominated by English and Chinese. The KO-SFT model is fine-tuned on Korean data starting from the foundation model.[10] The KO-CPT-SFT model first undergoes continued pre-training primarily on Korean and English data, with approximately 7% of the corpus consisting of other languages and code, before subsequent fine-tuning.

Using SOLAR-10.7B as the backbone, which is based on and scaled up from Mistral-7B (Jiang et al., 2023) and represents an English-centric architecture (Kim et al., 2024), the Original-SFT model is fine-tuned primarily on English data. The KO-SFT model is fine-tuned on a corpus consisting of 83.46% Korean data, 10.69% multilingual data (primarily English), and 5.86% English-Korean paragraph pairs. The KO-CPT-SFT model first undergoes continued pre-training on Korean data, followed by Korean SFT.[11]

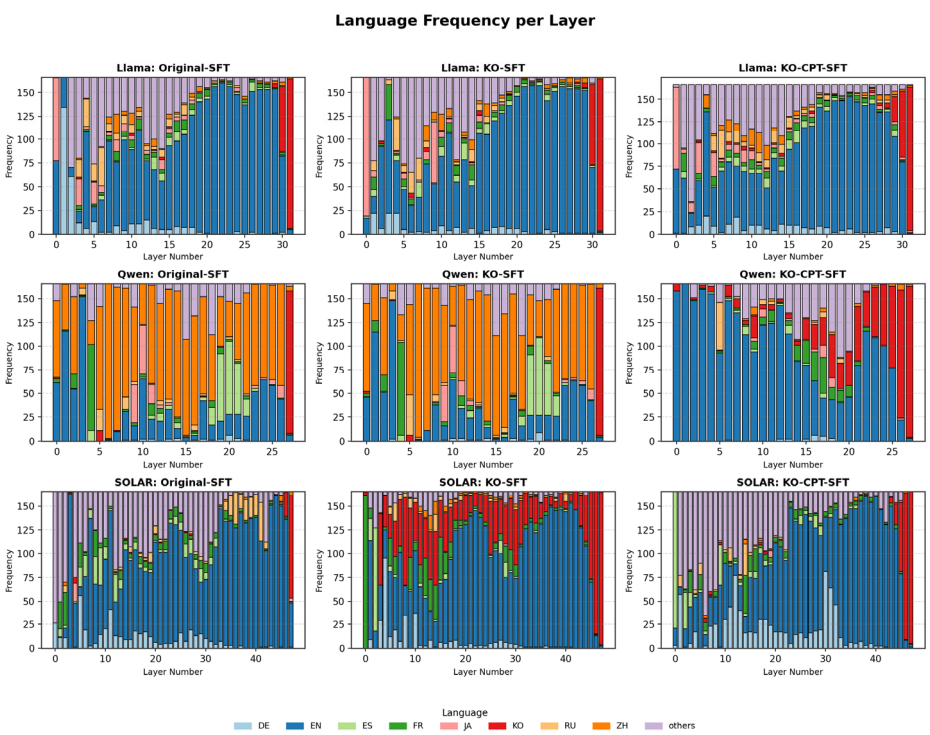
Although the Original-SFT, KO-SFT, and KO-CPT-SFT versions of each model are based on the same foundation architecture, they were independently trained by different institutions using separate datasets. As such, they are not part of a continuous or unified training pipeline.

We now analyze whether the injection of Korean data at different stages of the training pipeline leads to observable shifts in the models' latent languages.

9) https://huggingface.co/SEOKDONG/llama3.1_korean_v1.1_sft_by_aidx
10) https://huggingface.co/beomi/Qwen2.5-7B-Instruct-kowiki-qa
11) https://huggingface.co/refarde/OPEN-SOLAR-KO-10.7B-S-Core

## 4.2. Results

Figure 3 presents the distribution of latent languages across layers for the three versions of each model described in Table 1. As in RQ1, the x-axis represents the model layers, and the y-axis indicates the proportion of each detected latent language across the 166 cloze test problems. The same set of candidate latent languages, German (DE), English (EN), Spanish (ES), French (FR), Japanese (JA), Korean (KO), Russian (RU), and Chinese (ZH), is used for consistency, with all other detected languages grouped as "others." As in the previous analysis, only the middle to upper layers produce tokens that appear semantically meaningful to human inspection.




**<Figure 3> Layer-wise latent language distributions for three adaptation configurations of each model**

According to the observations in Figure 3, the Llama Original-SFT model consistently shows English (EN) as the dominant latent language across all cases. The KO-SFT model, despite

being fine-tuned on Korean data, also exhibits English dominance, showing a pattern very similar to that of the Original-SFT model. In the case of KO-CPT-SFT, since the training data consist of a mixture of English and Korean, it is difficult to isolate the effect of Korean post-training alone. However, the overall dominance of English slightly decreases, and Korean (KO) begins to appear from somewhat earlier layers.

For Qwen, the Original-SFT model shows Chinese (ZH) as the dominant latent language, with English (EN) also appearing to a lesser extent. The KO-SFT model, although fine-tuned on Korean data, displays a latent language pattern very similar to that of the Original-SFT model. In contrast, the KO-CPT-SFT model, whose training data combine English and Korean, exhibits a substantial reduction in the proportion of Chinese as the latent language, while English and Korean appear in roughly similar proportions. Korean tokens also begin to emerge from much earlier layers, and their proportion continues to increase toward the output layer.

For SOLAR, the Original-SFT model, being English-centric, shows a strong dominance of English (EN) as the latent language. In the KO-SFT model, Korean (KO) appears slightly more frequently, although the Korean tokens in the earlier layers are largely nonsensical strings that only happen to be written in Hangul. The overall trend remains similar to the English-dominant pattern, with the model processing primarily in English before shifting to Korean near the output layers. In the KO-CPT-SFT model, this shift from English to Korean persists, but Korean tokens begin to appear in somewhat earlier layers compared to the English-fine-tuned model.

Taken together, the results from Llama and Qwen suggest that KO-SFT alone is not sufficient to substantially alter the latent language patterns established during pre-training. This may be due to the fact that pre-training involves significantly larger volumes of data and longer training durations, which shape the model's foundational linguistic representations. However, as observed in SOLAR, even in layers that produce semantically noisy outputs, Korean tokens begin to appear, and the transition to Korean latent language occurs slightly earlier.

When models undergo CPT followed by SFT with Korean data, subtle yet consistent shifts in latent language patterns emerge. In both Llama and SOLAR, Korean tokens begin to appear slightly earlier across layers. Notably, for Qwen, which was originally centered on Chinese, applying CPT → SFT → DPO with a mixture of Korean and English data almost eliminates Chinese dominance from the latent language distribution, while English and Korean become the two prevailing languages. These findings suggest that, although applying only SFT has limited influence, a more extensive training process that includes both CPT and SFT can contribute to meaningful changes in a model's latent language patterns. A shift toward Korean as the latent language provides further evidence that the model has internally adapted to process Korean more natively.

However, as noted above, although these models share the same backbone architecture, they are distinct versions trained independently by different institutions, without continuity across training stages. This effect is particularly evident in comparisons among SOLAR-based models. Specifically, the SOLAR KO-SFT model reports additional vocabulary extension during fine-tuning based on various Korean web-crawled datasets. In contrast, the KO-CPT-SFT model underwent sufficient continued pre-training on diverse Korean texts, but is fine-tuning relied solely on a Korean-translated version of the relatively small Stanford Alpaca dataset (Taori et al., 2023). Contrary to our initial hypothesis, the higher proportion of Korean in the latent language observed in the KO-SFT model compared to the KO-CPT-SFT model may therefore be attributable to these differences in training data, as well as to additional methodological choices applied in the KO-SFT model.

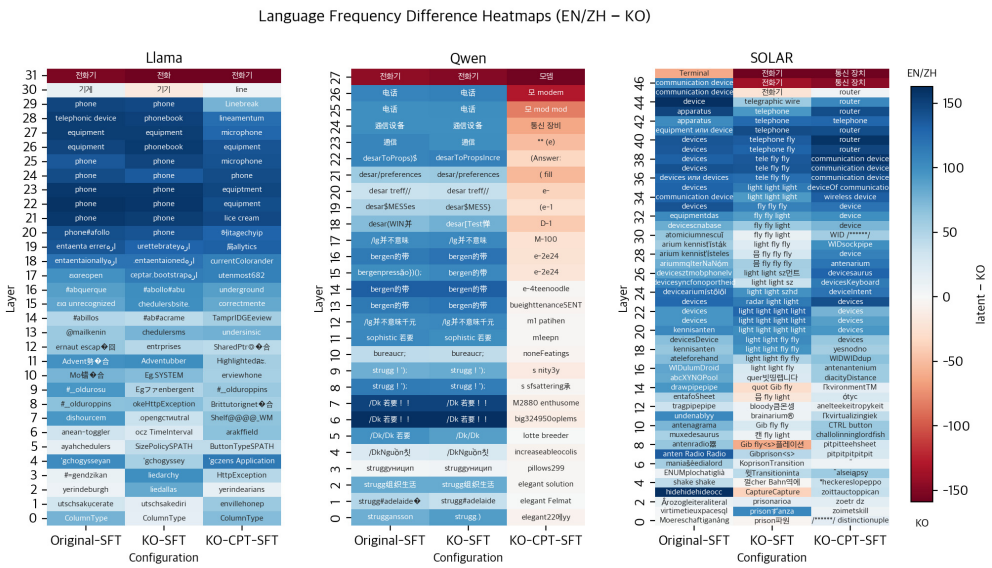
In the next analysis, we visualize the relative frequency of Korean responses compared to the top latent languages (English and Chinese) across layers when the same Korean prompts are given to the Original-SFT, KO-SFT, and KO-CPT-SFT models built on the same foundation architecture.

Figure 4 presents the results for each foundation model (Llama, Qwen, and SOLAR) and their corresponding model variants. The text within each cell illustrates an example of layer-wise unembedded responses for a single sample prompt. While the overall trends are consistent with those reported above in this section, this visualization allows for a clearer observation of cases where Korean tokens are used as the latent language compared to the major latent languages.

For Llama, whose latent language shift pattern follows EN → KO, the KO-SFT model shows little noticeable change, but in the KO-CPT-SFT model the proportion of English (EN) decreases slightly, indicating a weakened English dominance. For Qwen, whose latent language shift pattern mainly follows ZH → KO, the KO-SFT model again shows minimal difference from the original, whereas the KO-CPT-SFT model exhibits a clear shift, with Chinese (ZH) decreasing sharply and Korean (KO) becoming more prominent. For SOLAR, whose latent language shift pattern follows EN → KO, the KO-SFT model displays an increase in Korean tokens even in lower to middle layers, where outputs are generally semantically unclear. In the KO-CPT-SFT model, compared with the Original-SFT version, the red cells near the output layers deepen in color and an increased number of light-blue cells together reflect a stronger representation of Korean in the latent language.

The analyses in RQ2 demonstrate that sufficient adaptation with Korean data, particularly through the combination of continued pre-training (CPT) and supervised fine-tuning (SFT), and optionally alignment tuning (DPO), can induce a noticeable shift in a model's latent language toward Korean. This raises a further question: when such internal linguistic representations differ, do they also affect how the model processes and interprets Korean-specific linguistic

information? To explore this, RQ3 examines how models with distinct latent language profiles behave in tasks that require understanding of Korean linguistic features.




**<Figure 4> Layer-wise visualization of latent language shifts across model variants**

## 5. RQ3. Effects of Latent Language on Korean Linguistic Processing

### 5.1. Differences in field-wise scores on the KoBALT benchmark

As observed in RQ1 and RQ2, many models process inputs using tokens from non-target languages before "translating" their internal representations into Korean near the output layer. The tasks used in RQ1 and RQ2 were designed such that this translation-based processing did not hinder task performance, regardless of the internal language used. However, it remains unclear how these models behave when translation during processing leads to a loss of information that is crucial for solving the problem. Can they still perform correctly in such cases?

For instance, consider the Korean verb '뛰놀다ttwinolta' ("to run and play") again, which is a compound verb formed from '뛰다ttwida' ("to run") and '놀다nolta' ("to play"). If this word were internally translated into the English word 'frolic' or the Chinese word '游玩', the compound structure would be lost, making it impossible for the model to correctly determine

whether the original expression is a compound word.

To systematically probe such Korean-specific linguistic knowledge in LLMs, we adopt the KoBALT benchmark (Shin et al., 2025). Although several publicly available Korean exam-style question sets were also considered, KoBALT is explicitly structured by linguistic subfields, which better suits our research goals. This design allows us to compare model behavior across tasks that can largely be solved via translation (e.g., lexical and semantic problems, as used in most prior works) and those that cannot, such as tasks involving syntax and phonology.

The KoBALT benchmark provides a collection of tasks spanning five domains of linguistics: Semantics, Syntax, Phonetics/Phonology, Pragmatics, and Morphology. The benchmark also reports performance scores for several LLMs, including GPT, Claude, Qwen 2.5, Gemma, and Llama. The results show a consistent trend across models, with the highest scores in Semantics and Pragmatics, followed by Syntax, and considerably lower performances in Morphology and Phonetics/Phonology.

We hypothesize that the differences in performance across linguistic domains may stem from the loss of language-specific information when models process inputs through translated, non-Korean latent languages. For example, tasks in the Semantics domain, such as choosing the appropriate conjunction between sentences or selecting the correct classifier based on quantity, can likely be solved even if the model internally translates the prompt into another language. A model could, for instance, generate the English token 'but' or the Chinese token '但是' as its latent representation and then translate it into the Korean token '하지만' at the output layer, successfully matching the expected answer.

In contrast, tasks in domains such as Phonetics/Phonology and Morphology, which require reasoning over the internal structure or phonological constraints of Korean words, are far more vulnerable to information loss during translation. For instance, to identify a word that undergoes liquid assimilation in Korean phonology, the correct answer '달님talnim[달림tallim]' ("moon") would lose the relevant phonological information if it were translated into the English 'moon' or the Japanese 'お月様'. As a result, the model would no longer be able to determine whether assimilation occurs.

Since most of the models evaluated in the KoBALT paper are English-centric or multilingual, but not models in which Korean constitutes a major portion of the training data, the observed differences in performance across linguistic domains may partly reflect this imbalance in language representation. Therefore, to examine how differences in latent language patterns relate to model performance on the KoBALT dataset, we selected several models of comparable scale that exhibit distinct latent language characteristics.

The selected models exhibit distinct latent language shift patterns. **Aya-expanse-8B (Aya)** initially processes representations across multiple languages (e.g., EN, ZH, JA, and KO) before

converging on Korean (KO). **A.X-Light-4.0 (A.X)** displays mixed latent representations of English and Korean in earlier layers and subsequently shifts to KO; in some cases, meaningful outputs are expressed exclusively in Korean. Among all models examined in RQ2, A.X shows the strongest shift toward a Korean-dominant latent language. **Qwen-2.5-7B-Instruct (Qwen)** primarily processes content in Chinese (ZH) or English (EN) before mapping it to Korean. Finally, **Llama-3.1-8B-Instruct (Llama)** follows a predominantly English-centric pattern, shifting from EN to KO.

Using the publicly available KoBALT benchmark, we examine the field-wise scores (accuracy) of each model (Figure 5).[12] The KoBALT benchmark consists of multiple-choice questions with ten options, designed to test linguistic knowledge across five major fields. Each field contains a variety of subfields, for example, Syntax covers Embedded Clauses, Ellipsis, Agreement, Scrambling, and Argument Structure. In Figure 5, for Llama, Qwen, and Aya, the results are taken from those reported in the KoBALT paper, while for A.X, the results are obtained through inference using the same configuration.

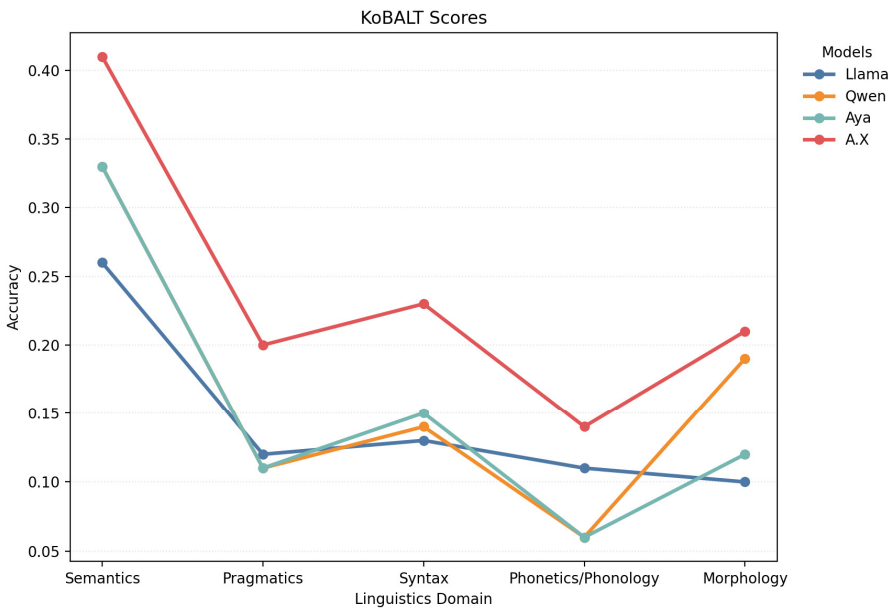
Similar to the overall trends reported in the KoBALT paper, all four models show notably higher performance in the Semantics field, followed by Syntax, and then Pragmatics and Morphology. Only Llama exhibits higher scores in Phonetics/Phonology than in Morphology, while Qwen shows relatively strong performance in Morphology, ranking second after Semantics. Overall, however, performance in the Phonetics/Phonology field remains distinctly low across all models.

The differences observed across these fields, particularly the clear gap between Semantics and Phonetics/Phonology, are likely closely related to whether the information required to solve each problem is preserved during processing. This relationship is explored further in Section 5.2.

## 5.2. Exploring latent languages across linguistic fields in the KoBALT benchmark

In this subsection, we examine the latent languages and corresponding information processing patterns of the four models while they solve tasks from different field of the KoBALT benchmark. In particular, we compare how models generate intermediate tokens, specifically which languages they use, and how these intermediate representations lead to the final answers in the Semantics and Phonetics/Phonology domains.

12) https://huggingface.co/datasets/snunlp/KoBALT-700




**<Figure 5> Accuracy scores of each model across five linguistic fields in the KoBALT benchmark**

The original KoBALT benchmark is designed as a multiple-choice question format with English alphabet options, which can blur the distinction of the target language and make the latent language more difficult to observe clearly. To address this, we sample questions from each linguistic field and convert them into a free-generation format. Four questions are selected from each subfield within the five major fields, resulting in a total of 92 questions: Semantics (28), Pragmatics (20), Syntax (16), Phonetics/Phonology (16), and Morphology (12). The Scrambling subfield of Syntax field is excluded because it contains only three questions. The tasks include various response formats, such as fill-in-the-blank, yes/no questions, and extractive or open-ended textual responses. Table 2 presents the English-translated examples of each question type.

**<Table 2> The English-translated examples of each question type**

| Question Type | Question | Answer |
|---|---|---|
| fill-in-the-blank | Passage: One [ ] of cigarettes should be enough.<br>Question: Fill in the blank [ ] in the passage with the appropriate word.<br>Answer: | 보루<br>(“carton”) |

| Question Type | Question | Answer |
|---|---|---|
| yes/no question | Passage: He knows the title of a unique song.<br>Question: Determine whether the passage can be interpreted ambiguously and answer yes or no.<br>Answer: | 예 (“yes”) |
| extractive QA | Passage: Because of the war, the people were driven into a deadly situation.<br>Question: Identify and write the passive expression in the passage.<br>Answer: | 내몰렸다 (“driven”) |
| open-ended QA | Passage: It became difficult to apply for the subsidy overnight.<br>Question: Describe the syntactic role of the embedded clause in the passage.<br>Answer: | 주어 (“subject”) |

We use the same instructions provided in the original KoBALT benchmark. An example input including the instruction is as follows (shown in English translation as Example (2)).

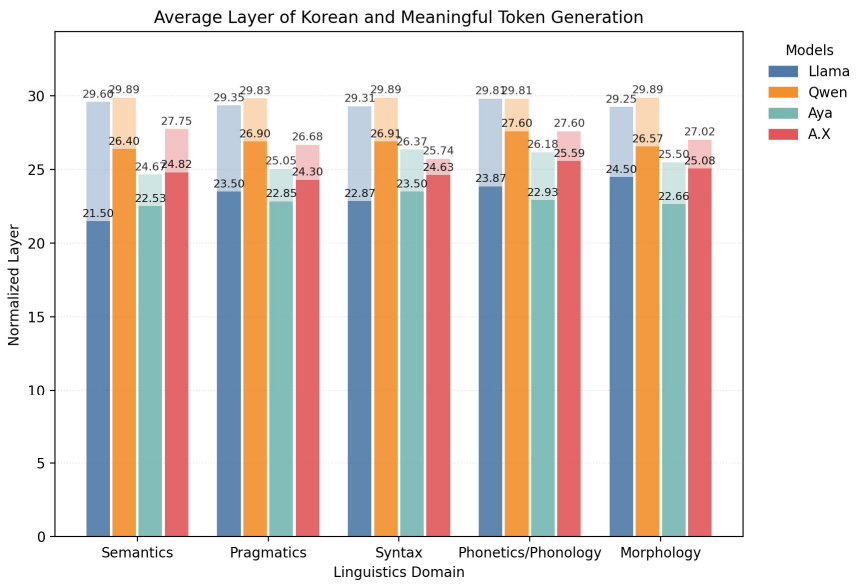
(2) You are an expert problem solver.
Think carefully and reason through the following problem before answering.
To solve the problem, let us think step by step.
Passage: One [ ] of cigarettes should be enough.
Question: Fill in the blank [ ] in the passage with the appropriate word.
Answer:

The experiment is conducted in a zero-shot setting, since the dataset is relatively small and contains diverse question types. Model generation follows the same configuration as in the previous experiments, using beam search with one beam and a generation length of five tokens. The outputs are decoded from each model layer based on the Logit Lens method, and the decoded results are classified into ISO-639-1 language codes using the gpt-4.1-2025-04-14 model API for language identification.

The results show that each model exhibits latent language distributions very similar to those observed in the dataset analyzed in RQ1 and RQ2. Overall, there were no substantial differences across models in the latent language distributions or its layer-wise shift patterns across the linguistic fields.

For Llama, the latent language predominantly shifts from English (EN) to Korean (KO) in the upper layers. Qwen shows two tendencies, either shifting from English and Chinese (EN+ZH) to Korean (KO) or directly from Chinese (ZH) to Korean (KO). Aya demonstrates a transition from multiple languages to Korean, with Korean tokens appearing relatively earlier than in other models. Finally, A.X shows a shift from English and Korean (EN+KO) to Korean,

and in some cases, Korean appears as the sole latent language from much earlier layers.




**<Figure 6> Average layer at which models begin generating meaningful tokens and Korean tokens across linguistic fields**

In this analysis, we visualize and report the average layer at which each model begins to generate semantically interpretable ("meaningful") tokens and the average layer at which Korean tokens first appear across different linguistic fields (Figure 6). Comparing these two values allows us to identify whether the model undergoes an intermediate processing phase in a non-target latent language and to estimate the length of that phase. Because the four models differ in the total number of layers (Llama and Aya: 32 layers, Qwen and A.X; 28 layers), for comparison, we normalize all models so that layer 0 corresponds to the input layer and layer 31 to the output layer. In the figure, the darker bars indicate the average layer where meaningful tokens first emerge, and the lighter bars indicate where Korean tokens begin to appear. For example, the leftmost blue bars show that the Llama model, when solving Semantics problems, starts producing meaningful tokens around layer 21.50 and begins switching to Korean tokens at approximately layer 29.60.

According to the figure, Korean tokens begin to appear in earlier layers in the order A.X > Aya $\geq$ Qwen > Llama. The same ordering is observed in the size of the gap between the layer where meaningful tokens first emerge and the layer where Korean tokens appear, with

A.X exhibiting the smallest gap and Llama the largest.

The expected order of models in terms of exposure to Korean data during training and post-training is A.X > Aya > Qwen > Llama, where Qwen is primarily centered on English and Chinese but described as multilingual. This ordering closely corresponds to the pattern observed in the onset of Korean token generation and the magnitude of the gap between the layers generating meaningful tokens and those generating Korean tokens. The overall KoBALT benchmark scores also follow a similar trend, with A.X > Aya = Qwen > Llama. These results suggest that the relative dominance of Korean data in a model's training corpus, that is, language distribution, likely influences both its latent language behavior and its performance on Korean linguistic knowledge benchmarks.

However, the gap size and the layers where meaningful and Korean tokens first appear do not show any clear tendencies across linguistic fields, although the Llama model exhibits a relatively large gap in the Semantics field. This indicates that when solving problems from different fields, the models do not display distinct differences in their internal processing or latent language patterns. The type of question also appears to have little effect, as the models generally process inputs using their main latent language (for example, EN for Llama) and then translate the output into Korean tokens in a similar manner.

Nevertheless, the degree to which this language-shifting process influences problem-solving performance appears to vary across linguistic fields. For example, in the Semantics field, the models follow intermediate reasoning processes similar to those shown in Example (3) (translated into English) and eventually arrive at their respective answers. Although not all responses are perfectly correct, they are generally semantically appropriate. In such cases, even when the model "thinks" in a non-Korean latent language and later converts its output into Korean, the meaning is preserved, and performance is not adversely affected.

(3) Passage: Jeongmin has just started working life and still lacks experience, often struggling when a senior colleague does not provide detailed guidance.

Question: Write an appropriate expression or metaphor that describes Jeongmin in this situation.

Answer:

**Gold label**: 햇병아리 ("rookie")

**Llama**: learning (EN) → fresher (EN) → novice (EN) → 초보자 (KO)

**Qwen**: 经验不足 (ZH) → 迷路的 (ZH) → 막내 (KO)

**Aya**: beginner (EN) → 缺乏经验者 (ZH) → 初初たる者 (JA) → 학습成果 부족 (mixed) → 초初 stages의 (mixed) → 초보자 (KO)

**A.X**: learning (EN) → fresh (EN) → 초 beginner (mixed) → 신입 (KO)

Meanwhile, Example (4) below illustrates a problem from the Phonetics/Phonology field, along with the model's intermediate processing and final answer.

(4) Question: In the word '생산량' ("production"), which type of Korean assimilation phenomenon occurs?

Answer:

**Gold label**: 비음화 ("nasalization")

**Llama**: Korean language's assimilation (EN) → Korean language has a phenomenon (EN) → 한국어의 동화 (KO)

**Qwen**: 生产量 (ZH) → 生産量 (JA) → 단어 '생산 (KO)

**Aya**: word 'productivity' (EN) → word '생산量' (mixed) → 단어 '생산량' (KO)

**A.X**: 생산량 (KO)

In this case, all models produced incorrect answers, likely due to the lack of knowledge about Korean assimilation phenomena. However, their latent language patterns during the reasoning process show clear differences. Notably, in the case of A.X, while many examples still exhibit mixed latent languages such as English (EN) or EN+KO, certain cases reveal the model initiating meaningful token generation directly in Korean, bypassing any intermediary language. The other three models, by contrast, exhibit intermediate stages in English (EN), Chinese (ZH), or Japanese (JA). In such cases, the models would be unable to recognize the nasalization phenomenon that occurs in the Korean word '생산량sayngsanlyang [생산냥 sayngsannyang].' In another problem requiring the model to recognize that the word '밤pam' ("chestnut") is pronounced with a long vowel, Aya generates the Turkish intermediate token 'gece' ("night"), whose Korean surface form also corresponds to '밤pam'. As a result, these models fail to produce correct answers for such problems.

Another interesting observation is that, in several cases, models generate correct or nearly correct answers in their middle-to-upper layers (approximately the fifth layer from the top) when reasoning in their latent language, but errors occur during the subsequent conversion into Korean tokens. As a result, the final Korean outputs are sometimes incorrect or contain irrelevant responses, such as prompt regeneration. This phenomenon appears across all four models (13 cases in Llama, 5 cases in Qwen, 2 cases in Aya, and 5 cases in A.X).

This can be explained in connection with the translation barrier hypothesis formalized by Bafna et al. (2025), which proposes that models process tasks in a target-language-agnostic manner and then subsequently translate the answer concepts into the intended target language. The observed errors occur precisely at this translation stage, even when task-solving itself has succeeded. Although the authors did not identify the exact cause of this

phenomenon, they reported that translation barriers account for the majority of failures in multilingual LLMs. It is therefore plausible that when the latent language is non-Korean, the existence of this translation step can negatively affect the model's performance.

Based on these observations, we can conclude that the main language of a model's training data, and consequently the latent language shaped by it, contributes to differences in performance across linguistic domains. The A.X model, whose latent language is closest to Korean, achieves the highest KoBALT benchmark score among all models experimented in this work and outperforms others even in more challenging domains such as Phonetics/Phonology, where most models tend to struggle. This suggests that a model's exposure to Korean as a dominant language during training has a significant influence on both its internal processing and its linguistic knowledge performance.

However, as shown in Example (3) above, the A.X model still frequently processes through English (EN) or mixed EN+KO latent languages, and the presence of a translation barrier continues to result in many incorrect answers. This may be due to the model's post-training on a mixture of English and Korean data. Future research should further investigate how to effectively induce larger shifts in latent language for models trained on different linguistic bases, as well as the validity and impact of such interventions.

## 6. Conclusion

In this study, we examined the latent language, or the "thinking language," that emerges in the internal token representations of LLMs as they process prompts and solve tasks. By analyzing models trained under different training data configurations, we compared their latent language patterns and investigated how the introduction of Korean data during adaptation and post-training stages affects these patterns. The results show that models likely exposed to larger amounts of Korean data exhibit a higher proportion of Korean tokens as their latent language. Moreover, even when non-Korean latent languages are employed, these models tend to perform translation into the target language, Korean, at earlier representational layers. Similarly, in post-training, models that underwent continued pre-training (CPT) followed by supervised fine-tuning (SFT), rather than SFT alone, demonstrated more frequent involvement of Korean in their latent languages, suggesting deeper internal adaptation to Korean linguistic representations.

We also quantitatively and qualitatively examined how variations in latent language distribution and the transition patterns into the target language influence performance on KoBALT, a benchmark designed to test Korean-specific linguistic knowledge across fields. The

results show that when the model relies on a non-Korean latent language and performs a translation into Korean, Korean-specific information embedded in the input can be lost, particularly in areas such as Morphology and Phonetics/Phonology, leading to a lower accuracy in these fields. In contrast, models that underwent sufficient post-training with a mixture of Korean and English data, and whose latent language involved more Korean tokens, achieved higher performance in these linguistically sensitive fields. These findings suggest that if a model is thoroughly adpated to Korean at the latent language level through sufficient Korean-language exposure, it is more likely to preserve and effectively process Korean-specific linguistic information.

However, since our study did not train the SFT and CPT → SFT models under fully controlled conditions using identical datasets and hyperparameters, but instead relied on publicly available models released by different sources, the results may not be entirely consistent or conclusive. Differences in model performance across fields in the KoBALT benchmark were also likely influenced by variations in the linguistic knowledge each model possesses.

Moreover, some of these models were trained on mixed-language data (primarily Korean but also including English and other languages), which makes it difficult to clearly disentangle the degree of adaptation attributable specifically to Korean training. In future work, we plan to conduct controlled experiments using smaller-scale models to systematically compare the effects of different training stages and degrees of adaptation. Factors such as the number of CPT and SFT epochs and the data budget allocated at each stage can also be explored to better understand their impact on latent language shifts and linguistic adaptation.

In RQ3, the low benchmark scores and frequent generation of incorrect answers may also be attirbuted to the fact that, regardless of the proportion of Korean in the latent language, the LLMs themselves do not yet possess sufficient Korean-specific lingusitic knowledge to reliably solve the tasks. In addition, unlike our experiments that reused the dataset from Zhong et al. (2025), which were conducted in a 2-shot setting, the KoBALT-based tasks in RQ3 were evaluated in a zero-shot setting and were also intrinsically more challenging. To mitigate this gap, future work will increase the number of sampled items per linguistics subfield (or simplify the tasks) and adopt settings with at least two shots, in order to better disentangle the effects of insufficient model knowledge and general problem-solving ability from the independent influence of latent language.

In addition to linguistic knowledge benchmarks such as KoBALT, future work could extend to cultural benchmarks such as CLIcK (Kim et al., 2024) and Nunchi-Bench (Kim and Lee, 2025) to examine whether a model's latent language influences its ability to produce culturally specific reasons. We are also interested in exploring how latent language operates in more

complex reasoning tasks and tool-calling processes, which have recently received increasing research attention. Future investigations will aim to determine whether the emergence of non-Korean latent languages and the resulting translation barriers affect model performance in these higher-level reasoning and interaction settings.

**<References>**

Bafna, Niyati, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. 2025. The Translation Barrier Hypothesis: Multilingual Generation with Large Language Models Suffers from Implicit Translation Failure. arXiv:2506.22724.

Bai, Jinze, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. arXiv:2309.16609.

Belrose, Nora, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting Latent Predictions from Transformers with the Tuned Lens. arXiv:2303.08112.

Conmy, Arthur, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track.

Cunningham, Hoagy, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse Autoencoders Find Highly Interpretable Features in Language Models. arXiv:2309.08600.

Dang, John, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas,

Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. arXiv:2412.04261.

Gemma Team. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295.

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

Kanana LLM Team. 2025. Kanana: Compute-efficient Bilingual Language Models. arXiv:2502.18934.

Kim, Eunsu, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp.3335-3346.

Kim, Kyuhee and Sangah Lee. 2025. Nunchi-Bench: Benchmarking Language Models on Cultural Reasoning with a Focus on Korean Superstition. Findings of the Association for Computational Linguistics: ACL 2025. pp.15328-15342.

Kim, Sanghoon, Dahyun Kim, Chanjun Park, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track). pp.23-35.

LG AI Research. 2024. EXAONE 3.0 7.8B Instruction Tuned Language Model. arXiv:2408.03541.

LG AI Research. 2024. EXAONE 3.5: Series of Large Language Models for Real-world Use Cases. arXiv:2412.04862.

Llama Team, 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Lieberum, Tom, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for

NLP. pp.278-300.

NAVER Cloud HyperCLOVA X Team. 2025. HyperCLOVA X THINK Technical Report. arXiv:2506.22403.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. LessWrong.

Ozaki, Shintaro, Tatsuya Hiraoka, Hiroto Otake, Hiroki Ouchi, Masaru Isonuma, Benjamin Heinzerling, Kentaro Inui, Taro Watanabe, Yusuke Miyao, Yohei Oseki, and Yu Takagi. 2025. Do LLMs Need to Think in One Language? Correlation between Latent Language and Task Performance. arXiv:2505.21458.

Qwen. 2024. Qwen 2.5 Technical Report. arXiv:2412.15115.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems. pp.53728-53741.

Schut, Lisa, Yarin Gal, and Sebastian Farquhar. 2025. Do Multilingual LLMs Think In English?. arXiv:2502.15603.

Shin, Hyopil, Sangah Lee, Dongjun Jang, Wooseok Song, Jaeyoon Kim, Chaeyoung Oh, Hyemi Jo, Youngchae Ahn, Sihyun Oh, Hyohyeong Chang, Sunkyoung Kim, and Jinsik Lee. 2025. KoBALT: Korean Benchmark For Advanced Linguistic Tasks. arXiv:2505.16125.

SKT AI Model Lab. 2025. A.X 4.0 Light. https://huggingface.co/skt/A.X-4.0-Light.

Song, Ran, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2024. Does Large Language Model Contain Task-Specific Neurons?. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp.7101-7113.

Tang, Tianyi, 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. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp.5701-5715.

Taori, Rohan, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechan Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An Instruction-following Llama model. Github Repository. https://github.com/tatsu-lab/stanford-alpaca.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, Kevin, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the Wild: a Circuit for Indirect Object Identification in GPT-2

small. arXiv:2211.00593.

Wang, Mingyang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schuetze. 2025. Lost in Multilinguality: Dissecting Cross-lingual Factual Inconsistency in Transformer Language Models. Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp.5075-5094.

Wei, Jason, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. International Conference on Learning Representations.

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp.15366-15394.

Zhong, Chengzhi, Qianying Liu, Fei Cheng, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2025. What Language Do Non-English-Centric Large Language Models Think in?. Findings of the Association for Computational Linguistics: ACL 2025. pp.26333-26346.

Received on: November 05, 2025
Revised on: December 14, 2025
Accepted on: December 18, 2025