

CCL-XCoT: An Efficient Cross-Lingual Knowledge Transfer Method for Mitigating Hallucination Generation

Weihua Zheng¹ Roy Ka-Wei Lee² Zhengyuan Liu¹

Kui Wu¹ AiTi Aw¹ Bowei Zou¹

¹Institute for Infocomm Research (I²R), A*STAR, Singapore

²Singapore University of Technology and Design

{zheng_weihua, liu_zhengyuan, wuk, aaiti, zou_bowei}@i2r.a-star.edu.sg
s.roylee@gmail.com

Abstract

Multilingual Large Language Models (MLLMs) demonstrate strong generalization across languages, yet they remain prone to hallucinations, especially in low-resource languages, due to training data imbalances. These hallucinations, which include inaccurate or fabricated outputs, are particularly problematic in domain-specific generation tasks (Chataigner et al., 2024). To address this challenge, we propose CCL-XCoT (Curriculum-based Contrastive Learning-based Cross-lingual Chain-of-Thought), a two-stage fine-tuning framework for mitigating hallucination in MLLMs. Our approach first enhances cross-lingual semantic alignment through curriculum-based contrastive learning combined with next-token prediction during continued pre-training. Building on this foundation, we then introduce a cross-lingual Chain-of-Thought (XCoT) prompting strategy during instruction fine-tuning, which guides the model to reason in a high-resource language before generating answers in the target low-resource language. Experimental results show that CCL-XCoT reduces hallucination rates by up to 62% and substantially improves factual knowledge transfer across language pairs, without relying on external retrieval or multi-model ensembles.

1 Introduction

Despite the impressive performance of large language models (LLMs) across a wide range of natural language processing tasks, hallucination remains a fundamental and unresolved challenge—particularly in multilingual settings. This issue is especially pronounced in low-resource languages, where limited training data increases the likelihood of inaccurate or misleading outputs. Hallucinations in LLMs typically manifest in three forms: (1) *context-related hallucinations*, which contradict general world knowledge; (2) *self-conflicting*

hallucinations, involving internal inconsistencies such as flawed reasoning; and (3) *ungrounded hallucinations*, where the generated content diverges from source facts despite maintaining surface fluency (Lei et al., 2023). This work focuses on mitigating context-related hallucinations in low-resource languages.

A major contributor to context-related hallucinations is the asymmetry in knowledge coverage between high-resource and low-resource languages. Recent studies have shown that multilingual models often fail to share factual knowledge across languages (Hu et al., 2025; Schut et al., 2025). This limitation is partially attributed to the language-agnostic next-token prediction (NTP) pretraining objective, which does not explicitly enforce cross-lingual semantic alignment. As a result, multilingual LLMs struggle to generalize factual knowledge to underrepresented languages.

Existing solutions such as Retrieval-Augmented Generation (RAG) have demonstrated effectiveness in English, but perform poorly in low-resource settings due to unreliable retrieval and limited language coverage (Niu et al., 2024). Other strategies, such as Chain-of-Thought (CoT) prompting (Ayala and Bechard, 2024; Song et al., 2024; Sun et al., 2025; Li et al., 2024), improve factual accuracy via step-by-step reasoning. However, CoT assumes the model already possesses sufficient internal knowledge in the target language—an assumption that often breaks down in low-resource conditions.

To address these gaps, we propose CCL-XCoT, a two-stage framework for improving factual generation in low-resource languages without relying on external retrieval. Our approach comprises two key components: (i) a *curriculum-based contrastive learning* strategy combined with next-token prediction during continued pretraining to align semantic spaces between high- and low-resource languages, and (ii) a cross-lingual Chain-of-Thought strategy during instruction fine-tuning, which enables mod-

els to reason in a high-resource language (e.g., English) before generating answers in the target low-resource language. This design facilitates the transfer of **non-language-specific** knowledge, such as common facts and historical information, across languages.

This paper makes the following contributions: (i) We propose CCL-XCoT, a two-stage framework that reduces hallucinations in low-resource languages by up to 62%, without requiring retrieval systems or ensemble models. (ii) We introduce curriculum-based contrastive learning to align multilingual semantic spaces, improving factual understanding and yielding up to 20% gains on cross-lingual NLU tasks. (iii) We develop XCoT strategy that transfers reasoning patterns from high-resource to low-resource languages, enhancing answer accuracy and completeness. (iv) Our layer-wise analysis reveals that mid-level transformer layers are key to effective cross-lingual knowledge transfer, guiding future efficient fine-tuning.

2 Related Work

LLMs have achieved strong performance in text generation tasks, but remain prone to hallucinations that undermine output reliability. To mitigate this, a range of strategies have been proposed.

Retrieval-Based Hallucination Mitigation. RAG integrates external knowledge to improve LLM outputs’ factual accuracy (Izcard et al., 2022; Lewis et al., 2021; Ram et al., 2023). For instance, Ayala and Bechard (2024) retrieved relevant task templates using MiniLM embeddings, while Song et al. (2024) proposed Hallucination Aware Tuning (HAT), which uses hallucination detection and GPT-4-based revision to build a preference dataset for Direct Preference Optimization(DPO). Despite promising results in English, RAG remains less effective in low-resource languages due to poor retrieval and limited training coverage (Niu et al., 2024). Moreover, RAG-based systems may generate conflicting outputs if the retrieved context is misinterpreted.

Reasoning-Based Mitigation and Self-Correction. Another line of work explores enhancing internal reasoning to suppress hallucinations without external retrieval. Lei et al. (2023) introduced CoNLI, which applies Natural Language Inference (NLI) to detect and rewrite unsupported generations. Chain-of-Verification

(Dhuliawala et al., 2023a) prompts the model to verify its own claims via question-answering before producing final outputs. Ji et al. (2023) adopted a self-reflective, multi-turn generation process for iterative refinement. While effective, these methods rely heavily on robust reasoning and world knowledge, capabilities that often degrade in low-resource languages, making such approaches less generalizable in multilingual contexts.

Ensemble and Agreement-Based Methods. Wei et al. (2024) propose FEWL, a multi-model ensemble that assigns weights to reference LLMs based on their reliability, and introduces a laziness penalty to penalize superficial responses. While FEWL shows promise, it depends on the availability of strong evaluation models, which are often lacking for non-English languages—thereby limiting its cross-lingual effectiveness.

Translation-based Methods. Huzaifah et al. (2024) found that LLMs exhibit greater potential for handling code-switching inputs. Chai et al. and Lin et al. (2024) used machine translation to align semantic spaces by generating code-switch sentences or swapping full question-answer sequences. However, these methods risk cumulative translation errors and ambiguities from fragmented replacements.

3 Methodology

We propose a framework to improve decoder-only language models for low-resource languages by integrating NTP, two-stage cross-lingual contrastive learning, and CoT reasoning. Unlike standard causal language modeling, our method explicitly aligns multilingual semantic spaces through contrastive learning and transfers knowledge from high-resource to low-resource languages via CoT prompting. An overview of the framework is illustrated in Figure 1.

3.1 Background

Next-Token Prediction in Causal Language Modeling. Causal language models (CLMs) are typically trained using the NTP objective, where each token is predicted based on its preceding context (Radford et al., 2018, 2019). However, this objective is applied primarily on monolingual corpora without enforcing cross-lingual semantic alignment. As a result, multilingual CLMs often exhibit fragmented representations and struggle to

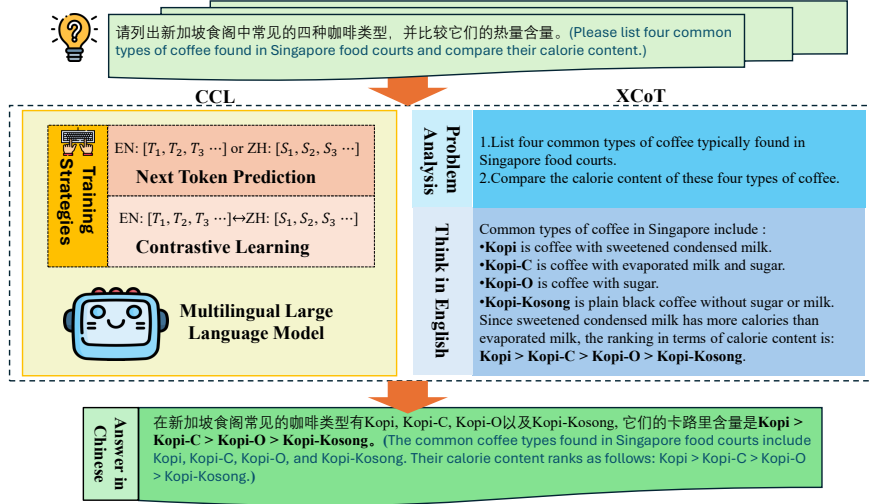


Figure 1: Overview of our proposed framework. The model is pre-trained with a combination of next-token prediction and a two-stage curriculum-based cross-lingual contrastive learning approach to align semantic spaces across languages. During the instruction fine-tuning phase, a cross-lingual CoT strategy guides the model to leverage knowledge in a high-resource language to improve responses in low-resource languages.

generalize knowledge across languages.

To address this, we incorporate contrastive learning during continued pretraining to encourage cross-lingual semantic alignment and improve multilingual transfer. At the same time, continued training may cause catastrophic forgetting of previously acquired knowledge (McCloskey and Cohen, 1989; Ratcliff, 1990; Gururangan et al., 2020; Ke and Liu, 2022; Liu, 2019). To mitigate this, we retain the NTP objective during pretraining and jointly optimize it alongside contrastive learning.

Contrastive Learning for Cross-Lingual Alignment. Contrastive learning has shown promise for aligning semantic spaces, particularly in encoder-decoder models (Lewis et al., 2019; Raffel et al., 2023; Liu and Liu, 2021). However, its application to decoder-only models remains limited and is largely restricted to English (Jain et al., 2022; Su et al., 2022). Its potential for cross-lingual alignment in such models is still underexplored.

To address this, we compute contrastive loss at the sequence level—rather than token level—to accommodate varying tokenization granularities across languages. For two semantically equivalent sentences S_m^i and S_n^i from languages m and n , we treat them as a positive pair. Batches for each language are defined as:

$$B_m = [S_m^i, S_m^{i+1}, \dots, S_m^{i+T}], \quad (1)$$

$$B_n = [S_n^i, S_n^{i+1}, \dots, S_n^{i+T}] \quad (2)$$

where sentences with matching indices are positives, and all others are treated as negatives.

3.2 Two-Stage Curriculum-Based Cross-Lingual Contrastive Learning

To align semantic representations across languages in decoder-only CLMs, we propose a two-stage curriculum-based contrastive learning framework, which progressively enhances cross-lingual understanding by addressing the scarcity of paragraph-level bilingual data and accommodating varying tokenization granularities across languages.

Stage 1: Sentence-Level Contrastive Pre-training. The first stage leverages large-scale sentence-aligned parallel corpora. Given bilingual sentence pairs $\{(S_m^i, S_n^i)\}_{i=1}^T$ from languages m and n , we compute sentence embeddings using the final hidden state of the decoder. Each aligned pair (S_m^i, S_n^i) is treated as a positive example; all others in the batch are treated as negatives. The contrastive loss is:

$$\mathcal{L}_{\text{contrastive}} = -\frac{1}{T} \sum_{i=1}^T \left[\log \frac{\exp(\text{sim}(S_m^i, S_n^i)/\tau)}{\sum_{j=1}^T \exp(\frac{\text{sim}(S_m^i, S_n^j)}{\tau})} + \log \frac{\exp(\text{sim}(S_n^i, S_m^i)/\tau)}{\sum_{j=1}^T \exp(\frac{\text{sim}(S_n^i, S_m^j)}{\tau})} \right]. \quad (3)$$

where $\text{sim}(\cdot)$ denotes cosine similarity and τ is the temperature. This stage promotes coarse-

grained cross-lingual alignment without altering the decoder-only architecture.

Stage 2: Paragraph-Level Finetuning. The second stage uses a smaller corpus of paragraph-aligned bilingual texts, consisting of multi-sentence or document-level segments. We apply the same contrastive loss, but compute paragraph embeddings via mean pooling over token-level representations. Despite the smaller scale, this stage is crucial for: (1) extending alignment to longer discourse spans for real-world multilingual applications, and (2) preserving long-text modeling capabilities that may degrade with sentence-only training. This curriculum, moving from short to long text, supports hierarchical multilingual representation learning.

To prevent catastrophic forgetting and retain general language modeling capabilities, we jointly optimize contrastive loss with the NTP loss during both stages:

$$\mathcal{L}_{\text{PT}}(\theta) = \mathcal{L}_{\text{NTP}}(\theta) + \mathcal{L}_{\text{contrastive}}(\theta). \quad (4)$$

3.3 Cross-lingual Chain-of-Thought (XCoT)

To further enhance knowledge transfer from high-resource to low-resource languages, we introduce a Cross-lingual Chain-of-Thought (XCoT) prompting strategy during instruction fine-tuning. XCoT guides the model to decompose and solve queries in low-resource languages by leveraging the reasoning capabilities and factual knowledge it has primarily learned in English.

During instruction fine-tuning and inference, XCoT operates in three steps:

1. **Reasoning in English.** Given a question in a low-resource language, the model is first prompted to outline the reasoning steps in English, drawing on clearer semantic structures and well-formed logical patterns.
2. **Answering in English.** It then generates a concise English answer based on the reasoning trace, capturing the core factual or inferential content.
3. **Respond in target language.** Finally, the model returns the answer in the original low-resource language, preserving both accuracy and linguistic appropriateness.

This cross-lingual strategy complements contrastive pretraining by explicitly bridging linguistic

gaps during reasoning and generation. Illustrative examples of XCoT are provided in Appendix A, Table 2.

4 Experiments

4.1 Experimental Settings

We focus on Singaporean cultural knowledge as our target domain, using English as the high-resource language due to its broad data availability and strong representation in existing LLMs. For multilingual evaluation, we adopt Gemma-7B and LLaMA-3.1-8B, two decoder-only models with multilingual capabilities.

To simulate a realistic scenario where the high-resource language contains domain-specific knowledge, we inject cultural content exclusively into English during instruction fine-tuning. The English instruction fine-tuning dataset is constructed by combining 10,000 samples from CRAFT, a QA dataset focused on Singaporean cultural knowledge (Wang et al., 2024), with the Alpaca English dataset (Taori et al., 2023). From CRAFT, we manually select 500 fact-based QA pairs as the test set and translate them into Chinese and Malay (mid-resource languages) and Tamil (low-resource language), categorized according to their relative representation in the pretraining corpora.

All models are fine-tuned using Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient training. During continued pretraining, we update layers 9–19 for Gemma-7B and layers 11–22 for LLaMA-3.1-8B. In contrast, instruction fine-tuning is applied to all layers in both models. We evaluate hallucination mitigation performance on both models and provide further analysis on the impact of layer-specific tuning. For ablation and exploratory experiments, we use Gemma-7B as the default model. Full dataset construction details are provided in Section A.4.

4.2 Cross-Lingual Hallucination Analysis

We evaluate hallucination rates in non-English QA tasks and compare our proposed method against several baselines. Responses are manually assessed by linguistic experts using authoritative sources. A response is considered **hallucination-free** if it satisfies: (1) It accurately and fully answers the question. (2) It contains no factual errors, including in any elaborations.

Representative outputs are shown in Appendix A, Table 2, covering hallucination-free

cases and three common error types: (1) incomplete but seemingly sufficient answers, (2) factually incorrect but topically relevant responses, and (3) irrelevant answers.

Table 1 reports hallucination-free rates for Chinese, Malay, and Tamil. We observe that standard instruction fine-tuning using English domain-specific data fails to generalize effectively to non-English outputs. This underscores a key limitation of current multilingual LLMs: knowledge acquired in high-resource languages does not naturally transfer to low-resource ones. The problem is especially severe for Tamil, where hallucination-free rates are just 1% (Gemma) and 2% (LLaMA-3.1). Malay, sharing more lexical overlap with English, achieves the highest baseline performance, 16% and 18%, for Gemma and LLaMA-3.1, respectively.

Applying our XCoT strategy yields substantial improvements. Hallucination-free responses increase by 38% (Gemma) and 36.6% (LLaMA-3.1) on average across all three languages. The effectiveness of this approach is primarily driven by XCoT’s ability to scaffold the reasoning process using high-resource language knowledge and well-learned reasoning patterns, reducing both factual inaccuracies and incomplete answers.

Integrating XCoT with our cross-lingual contrastive learning objective (CL-XCoT) further amplifies performance gains. The combined framework aligns multilingual representations while enhancing factual transfer from English to other languages. We achieve average increase of hallucination-free rates across both models, reaching 59% (Chinese), 46% (Malay), 54% (Tamil) for Gemma and 65% (Chinese), 41% (Malay), 49% (Tamil) for LLaMA-3.1, significantly outperforming all baselines. Notably, these improvements are consistent across typologically diverse languages, demonstrating the generality of our approach. Paragraph-level contrastive learning (CCL-XCoT) yields additional benefits, with an average improvement of 4% for Gemma and 4.33% for LLaMA-3.1 across the three languages. These improvements are primarily attributed to enhanced capabilities in handling long texts and a deeper semantic understanding enabled by longer sequences. These results highlight that with proper alignment and structured reasoning, multilingual LLMs can bridge the semantic and factual gaps between vastly different languages, without needing task-specific data in each target language.

We also benchmark against two competitive baselines:

- **CrossIn** (Lin et al., 2024): a translation-based fine-tuning approach. Our method surpasses it by over 32% (Gemma) and 29% (LLaMA-3.1) in Tamil, indicating that contrastive semantic alignment paired with reasoning transfer is more effective than direct translation in low-resource scenarios.
- **CoV** (Dhuliawala et al., 2023b): a verification-based generation method. Our model outperforms CoV by an average of 53.8% across all three languages. This suggests that grounding generation in high-resource reasoning may be more effective than post-hoc verification mechanisms in multilingual contexts.

To estimate the upper bound of hallucination mitigation, we implement an RAG baseline (Gemma-7B+SFT+RAG), using Sentence-BERT (Reimers and Gurevych, 2020) to retrieve top-2 relevant paragraphs as in-context prompts. While RAG offers modest improvements in Chinese and Malay, it performs poorly in Tamil. This highlights a critical limitation: retrieval-based methods depend not only on evidence quality, but also on the model’s ability to comprehend and integrate it, which remains fragile in low-resource languages. Our method, in contrast, requires no retrieval infrastructure and delivers stronger, more consistent results.

We further show that larger model scale and more training data fail to reduce hallucinations in low-resource languages, where hallucination-free rates remain low (24%) despite high rates in English (85%). Our 8B model outperforms the 70B model in domain-specific QA. See Section A.3 for details.

4.3 Impact of Contrastive Learning on Semantic Alignment

Semantic space alignment is critical for cross-lingual generalization in multilingual LLMs, particularly in zero-shot scenarios (Devlin et al., 2019; Conneau et al., 2019). To assess this, we adopt the metric from Li et al. (2025), which measures cosine similarity between average sentence embeddings of semantically aligned sentence pairs across model layers. We use 5,000 such pairs to compute both similarity scores and variance, with lower variance indicating more consistent alignment.

Method	Chinese	Malay	Tamil
Gemma-7B+SFT	9%	16%	1%
CrossIn(Lin et al., 2024)	50%	51%	28%
CoV(Dhuliawala et al., 2023b)	12%	19%	3%
Gemma-7B+SFT+RAG	53%	50%	30%
Gemma-7B+XCōT	48%	55%	38%
Gemma-7B+CL-XCōT	68%	62%	55%
Gemma-7B+CCL-XCōT	71%	66%	60%

Table 1: Hallucination-free rates (%) in Chinese, Malay, and Tamil under different training strategies. +SFT: direct instruction fine-tuning; +XCōT: apply XCōT during fine-tuning; +CL: integrate sentence-level contrastive learning in pretraining; +CCL: integrate curriculum-based contrastive learning in pretraining. Best results are in **bold**.

Method	Chinese	Malay	Tamil
LLaMA-3.1-8B+SFT	4%	18%	2%
CrossIn(Lin et al., 2024)	49%	53%	26%
CoV(Dhuliawala et al., 2023b)	10%	22%	4%
LLaMA-3.1-8B+SFT+RAG	58%	52%	31%
LLaMA-3.1-8B+XCōT	50%	41%	43%
LLaMA-3.1-8B+CL-XCōT	69%	59%	51%
LLaMA-3.1-8B+CCL-XCōT	74%	63%	55%

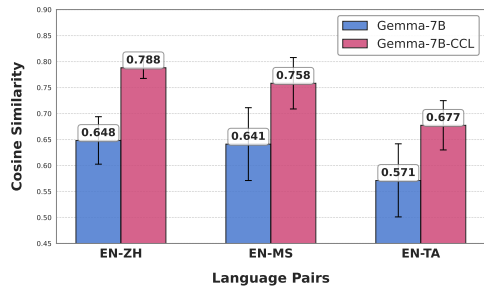


Figure 2: Semantic space alignment across pre-trained models. Bars indicate average cosine similarity between English and target languages; error bars represent sample variance.

As shown in Figure 2, incorporating contrastive learning during continued pretraining substantially improves semantic alignment. Compared to the baseline Gemma-7B model, our contrastive variant yields similarity gains of 14.0% (English–Chinese), 11.7% (English–Malay), and 10.6% (English–Tamil), along with marked reductions in variance.

These results highlight the role of contrastive learning in improving multilingual representations, especially in underrepresented language pairs. By encouraging the model to anchor semantically equivalent sentences in a shared latent space, our approach enables more reliable transfer of factual and contextual knowledge across languages. This alignment effect underpins downstream improvements in generation quality and hallucination mitigation observed in earlier experiments.

4.4 Evaluating Cross-Lingual Understanding and Consistency

To understand which capabilities are most enhanced by cross-lingual semantic alignment, we categorize NLP tasks into three stages: (1) input understanding, (2) reasoning and inference, and (3)

language-specific output. We hypothesize that improved alignment primarily benefits the first two stages, especially for languages with limited pretraining exposure.

We test this by comparing the baseline Gemma-7B model against Gemma-7B+CCL, which incorporates our curriculum-based contrastive learning during continued pretraining. Both models are instruction-tuned on the same dataset. Evaluation is conducted on two multilingual benchmarks: Cross-MMLU (knowledge understanding) and Cross-LogiQA (logical reasoning), drawn from SEA-Eval (Wang et al., 2023), with test sets in English, Chinese, and Malay. We also manually translated and validated Tamil versions of both datasets, resulting in a four-language evaluation.

As shown in Figure 3, Gemma-7B+CCL outperforms the baseline across nearly all settings. The largest gains are in Cross-MMLU: Chinese (+20%) and Malay (+10%). Cross-LogiQA shows smaller, though generally positive, improvements. These results confirm that semantic alignment significantly boosts multilingual comprehension, while its effect on reasoning is more moderate. Interestingly, English performance also improves (+12.66% on MMLU), suggesting that multilingual alignment benefits cross-lingual transfer and strengthens English representations, possibly by refining structural understanding and reducing representational redundancy. Tamil, as a low-resource language, shows smaller gains. While MMLU accuracy remains low, Cross-LogiQA improves by 3.98%. This underscores that alignment helps even when monolingual training is minimal, though limitations in base representation quality cap the benefits.

We also assess cross-lingual consistency, whether the same question yields semantically equivalent answers across languages. As shown

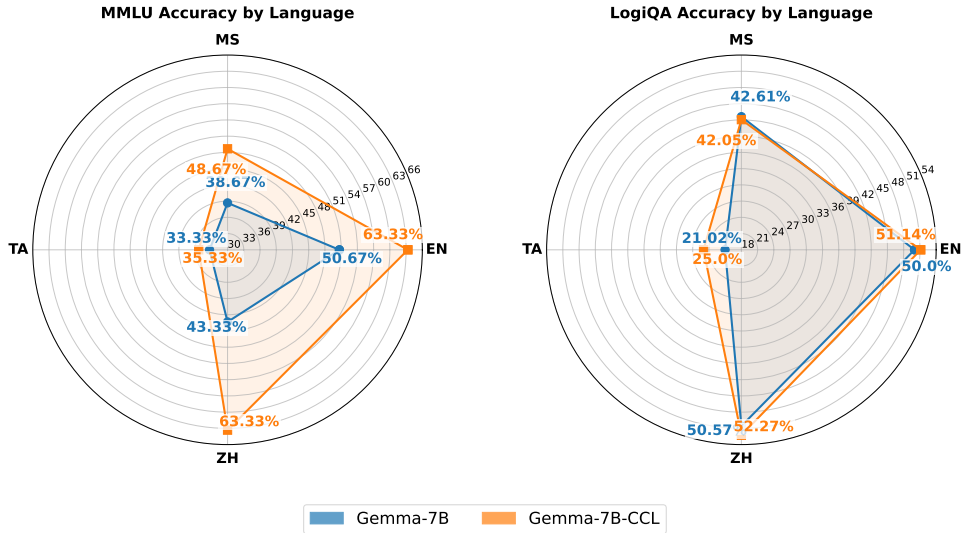


Figure 3: Accuracy comparison on Cross-MMLU and Cross-LogiQA across four languages using Gemma-7B and Gemma-7B+CCL.

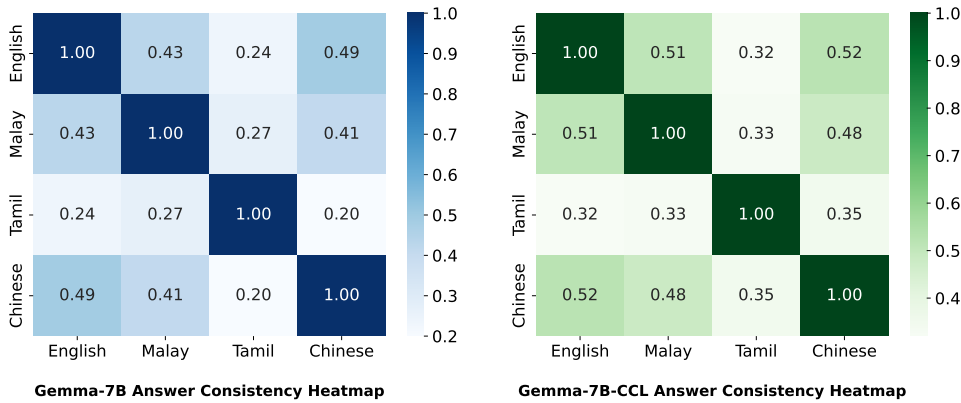


Figure 4: Answer consistency heatmap across language pairs for Gemma-7B and Gemma-7B-CCL.

in Figure 4, consistency with English increases by 8% for both Tamil and Malay, and 3% for Chinese. More notably, alignment improves even among non-English pairs (e.g., ZH-TA: +15%, MS-TA: +6%, ZH-MS: +7%) despite no direct training signal between them. These findings reveal that English can act as a pivot language, indirectly aligning low-resource language pairs through shared semantic structure. This opens up promising directions for multilingual model training without relying on costly parallel corpora.

4.5 Layer-Wise Analysis of Cross-Lingual Knowledge Transfer

Recent studies suggest that large language models (LLMs) exhibit a tripartite functional architecture, with distinct computational roles across different layer hierarchies. The lower layers, forming the input space, primarily handle raw token pro-

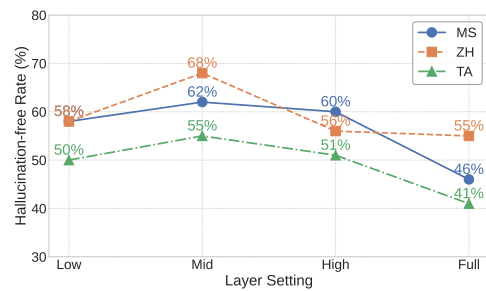


Figure 5: Hallucination-free rates by language across different layer-wise fine-tuning settings.

cessing and initial contextualization, as indicated by their high prediction entropy (Wendler et al., 2024). The intermediate layers, or concept space, operate as a language-agnostic region where factual and relational knowledge is retrieved. This characterization is supported by observed cross-lingual latent similarity peaks and the encoding

of abstract relational structures within these layers (Wang et al., 2025; Fierro et al., 2025). Moreover, these layers are hypothesized to be the main locus of cross-lingual transfer, mediated through an English-centric “latent language” representation (Wendler et al., 2024). Finally, the upper layers, constituting the language transition component, map abstract knowledge representations onto language-specific token distributions.

To investigate the role of each layer group in knowledge transfer, we partition the Gemma-7B architecture into three segments: lower layers (0–8), intermediate layers (9–19), and upper layers (20–27). We then fine-tune each segment independently and evaluate its effect on hallucination-free generation across Chinese, Malay, and Tamil.

As shown in Figure 5, fine-tuning only the **mid-level layers** achieves the best performance across all languages. High-level tuning ranks second, while full-model fine-tuning surprisingly underperforms both. These findings suggest that mid-level layers play a pivotal role in aligning semantic representations across languages, likely functioning as the core bridge between surface-level linguistic features and higher-order reasoning.

In contrast, fine-tuning low-level layers leads to a noticeable performance drop. We hypothesize that these layers encode language-specific syntactic and morphological patterns; altering them may distort the model’s capacity to process diverse linguistic inputs. Worse, updates in these early layers can propagate undesirable shifts to downstream layers, disrupting the semantic abstraction necessary for effective knowledge transfer. These results highlight the importance of targeted mid-layer adaptation in multilingual transfer settings—offering a more efficient and stable alternative to full-model tuning.

5 Logit Lens Analysis of CCL

To further investigate the impact of incorporating contrastive learning during the pre-training stage on the base model, we employ Logit Lens (nostalgebraist, 2020), a tool that projects hidden states at each layer onto the vocabulary space, allowing us to analyze the evolution of intermediate states during inference. As discussed earlier, cross-lingual knowledge transfer and reasoning typically occur in the middle-to-upper layers of the model. Therefore, we focus on layers 18–26 (out of 28 layers in Gemma-7B, indexed from 0 to 27). In the fig-

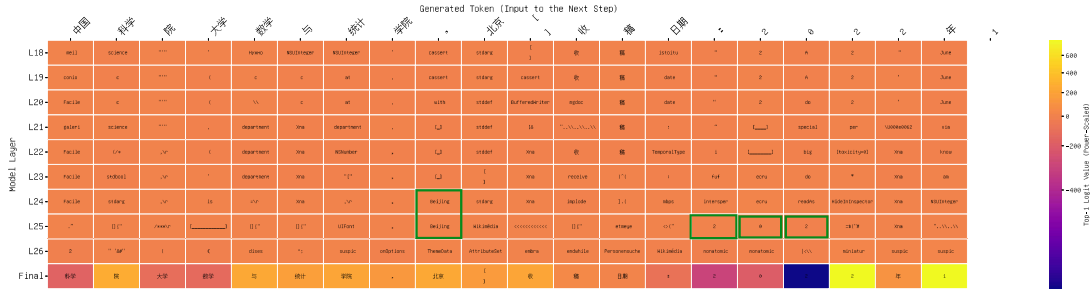
ure, each column represents a time step: except for the initial step, the input at each subsequent step is derived from the output of the previous step, with the token at the top of each column denoting the current input.

As shown in Figure 6a, when the initial token “中国” is given and the model is prompted to continue generation, the intermediate layers of the baseline Gemma-7B, when projected through Logit Lens, often correspond to semantically irrelevant or meaningless tokens. Its reasoning process appears to occur in a language-agnostic space. In contrast, after applying CCL, as illustrated by the light green highlight in Figure 6b, even when the prompt is in Chinese, the model’s intermediate reasoning process tends to proceed in English. Before generating the next Chinese token, it frequently associates with corresponding or related English tokens. This demonstrates that CCL significantly enhances cross-lingual alignment, enabling the model to leverage high-resource languages during inference for low-resource language prompts, thereby facilitating more effective cross-lingual knowledge transfer.

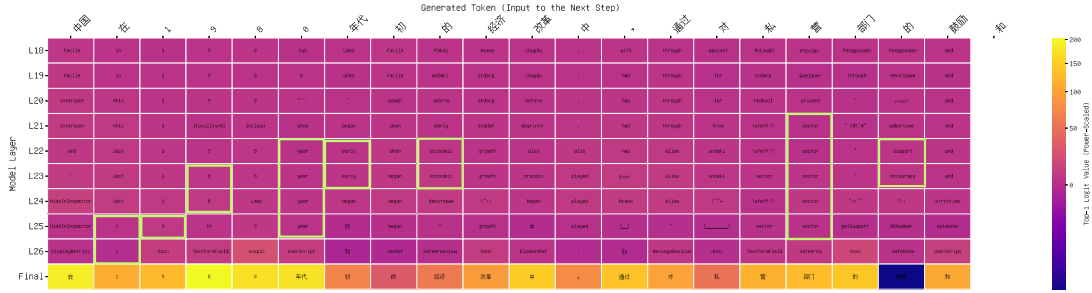
6 Case Study

To better understand hallucination behavior across languages, we conduct a targeted case study comparing model outputs in medium-resource languages (Chinese and Malay) and a low-resource language (Tamil). After direct instruction fine-tuning, we observe distinct error profiles. In Chinese and Malay, hallucinations typically fall into two categories: unnecessary elaborations beyond the scope of the prompt, and misinterpretations of culturally specific or domain-relevant terms, such as *Cendol* or the *Home Ownership Scheme*. In contrast, hallucinations in Tamil are predominantly severe, with over 90% of cases producing responses that are entirely unrelated to the input question, highlighting the fragility of the model’s representations in truly low-resource settings.

Introducing the XCoT strategy leads to substantial improvements across all languages. For Chinese and Malay, XCoT reduces over-generation errors by encouraging more focused and structured reasoning. However, challenges with domain-specific terminology persist, as these terms are often underrepresented in the training data. In Tamil, XCoT improves response relevance by providing a clearer reasoning scaffold, but some hallucinations



(a) Latent space for Gemma-7B



(b) Latent space for Gemma-7B-CCL

Figure 6: Logit lens maps of Gemma-7B model (a) and Gemma-7B-CCL model (b) highlight key features in the model’s intermediate layers with green boxes.

remain, often manifesting as topic drift or vague associations. When XCoT is combined with cross-lingual contrastive learning, we observe further improvements in all three languages, particularly in handling terminology. In medium-resource settings, hallucinations are reduced to minor factual inaccuracies, such as incorrect years or road names, indicating a transition from gross misunderstanding to finer-grained errors. In Tamil, hallucinations become more diverse but notably less frequent, suggesting better control over output semantics despite limited language-specific training signals.

These findings highlight the importance of tailoring multilingual LLM interventions based on language resource availability. For medium-resource languages, enhancing factual precision and grounding in culturally specific terminology should be prioritized. For low-resource languages like Tamil, future improvements will likely require more robust semantic alignment mechanisms and adaptive reasoning strategies to address a broader range of hallucination types and maintain answer relevance. Additional qualitative examples are included in Section A.5 to further illustrate these patterns.

7 Conclusion

This paper present CCL-XCoT, an efficient two-stage framework for cross-lingual knowledge transfer in multilingual LLMs. Our approach combines curriculum-based contrastive learning during continued pretraining with XCoT strategy during instruction fine-tuning. Together, these components align semantic representations across languages and scaffold reasoning through high-resource language traces, substantially reducing hallucinations in low-resource question answering tasks.

Extensive experiments show that contrastive learning improves semantic alignment and consistency, while XCoT enables more accurate and faithful generation across diverse language pairs. Our layer-wise analysis reveals that cross-lingual transfer is most effective when concentrated in mid-level layers, offering a practical path toward efficient multilingual adaptation. Case studies further demonstrate how hallucination patterns vary by resource level, highlighting the need for differentiated strategies.

Looking ahead, future work may explore integrating multilingual retrieval augmentation, expanding XCoT-style reasoning to additional modalities, and developing lightweight adapters to support broader cross-lingual generalization, especially in zero-shot or resource-sparse scenarios.

Limitations

Although our proposed training strategy demonstrates high effectiveness in promoting cross-lingual sharing of factual knowledge, its impact on enhancing capabilities for complex logical reasoning tasks is limited. Furthermore, while the strategy primarily focuses on reducing the overall hallucination rate, it lacks effective suppression mechanisms for more fine-grained hallucination types, such as those involving years, designer names, or road numbers, which remain prevalent even in high-resource languages.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

This research project is also supported by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (Award No: T2EP20222-0036).

References

- Orlando Ayala and Patrice Bechar. 2024. [Reducing hallucination in structured outputs via retrieval-augmented generation](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, page 228–238. Association for Computational Linguistics.
- Linzheng Chai, Jian Yang, Tao Sun, Hongcheng Guo, Jiaheng Liu, Bing Wang, Xiannian Liang, Jiaqi Bai, Tongliang Li, Qiyao Peng, and 1 others. `xcot`: Cross-lingual instruction tuning for cross-lingual chain-of-thought reasoning, 2024. URL <https://arxiv.org/abs/2401.7037>.
- Cléa Chataigner, Afaf Taïk, and Golnoosh Farnadi. 2024. [Multilingual hallucination gaps in large language models](#). *Preprint*, arXiv:2410.18270.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023a. [Chain-of-verification reduces hallucination in large language models](#). *Preprint*, arXiv:2309.11495.
- Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023b. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.
- Constanza Fierro, Negar Foroutan, Desmond Elliott, and Anders Søgaard. 2025. [How do multilingual language models remember facts?](#) *Preprint*, arXiv:2410.14387.
- 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. *arXiv preprint arXiv:2004.10964*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.
- Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. [Large language models are cross-lingual knowledge-free reasoners](#). *Preprint*, arXiv:2406.16655.
- Muhammad Huzaifah, Weihua Zheng, Nattapol Chanpaisit, and Kui Wu. 2024. [Evaluating code-switching translation with large language models](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6381–6394, Torino, Italia. ELRA and ICCL.
- Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. [Atlas: Few-shot learning with retrieval augmented language models](#). *Preprint*, arXiv:2208.03299.
- Nihal Jain, Dejiao Zhang, Wasi Uddin Ahmad, Zijian Wang, Feng Nan, Xiaopeng Li, Ming Tan, Ramesh Nallapati, Baishakhi Ray, Parminder Bhatta, and 1 others. 2022. Contraclm: Contrastive learning for causal language model. *arXiv preprint arXiv:2210.01185*.

- Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. [Towards mitigating LLM hallucination via self reflection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.
- Zixuan Ke and Bing Liu. 2022. Continual learning of natural language processing tasks: A survey. *arXiv preprint arXiv:2211.12701*.
- Deren Lei, Yaxi Li, Mengya Hu, Mingyu Wang, Vincent Yun, Emily Ching, and Eslam Kamal. 2023. [Chain of natural language inference for reducing large language model ungrounded hallucinations](#). *Preprint*, arXiv:2310.03951.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Preprint*, arXiv:2005.11401.
- Johnny Li, Saksham Consul, Eda Zhou, James Wong, Naila Farooqui, Yuxin Ye, Nithyashree Manohar, Zhuxiaona Wei, Tian Wu, Ben Echols, Sharon Zhou, and Gregory Diamos. 2024. [Banishing llm hallucinations requires rethinking generalization](#). *Preprint*, arXiv:2406.17642.
- Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2025. Language ranker: A metric for quantifying llm performance across high and low-resource languages. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 28186–28194.
- Geyu Lin, Bin Wang, Zhengyuan Liu, and Nancy F Chen. 2024. Crossin: An efficient instruction tuning approach for cross-lingual knowledge alignment. *arXiv preprint arXiv:2404.11932*.
- Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Yixin Liu and Pengfei Liu. 2021. Simcls: A simple framework for contrastive learning of abstractive summarization. *arXiv preprint arXiv:2106.01890*.
- Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. [Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models](#). *Preprint*, arXiv:2401.00396.
- 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 John Hoffman Semarley Jarrett Kaushik Ram Sadagopan Dirk Rowe Shannon Spruit Chau Tran Pierre Andrews Necip Fazil Ayan Shruti Bhosale Sergey Edunov Angela Fan Cynthia Gao Vedanuj Goswami Francisco Guzmán Philipp Koehn Alexandre Mourachko Christophe Ropers Safiyyah Saleem Holger Schwenk Jeff Wang NLLB Team, Marta R. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation.
- nostalgebraist. 2020. Interpreting gpt: The logit lens. <https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens>. Accessed: 2025-07-19.
- Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer](#). *Preprint*, arXiv:1910.10683.
- Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlga, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. [In-context retrieval-augmented language models](#). *Transactions of the Association for Computational Linguistics*, 11:1316–1331.
- Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. 2025. [Do multilingual llms think in english?](#) *Preprint*, arXiv:2502.15603.
- Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. [RAG-HAT: A hallucination-aware tuning](#)

pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. *Advances in Neural Information Processing Systems*, 35:21548–21561.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2025. *Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability*. *Preprint*, arXiv:2410.11414.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

Bin Wang, Geyu Lin, Zhengyuan Liu, Chengwei Wei, and Nancy F. Chen. 2024. *Craft: Extracting and tuning cultural instructions from the wild*. *Preprint*, arXiv:2405.03138.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, AiTi Aw, and Nancy F Chen. 2023. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *arXiv preprint arXiv:2309.04766*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. *Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models*. *Preprint*, arXiv:2504.04264.

Jiaheng Wei, Yuanshun Yao, Jean-Francois Ton, Hongyi Guo, Andrew Estornell, and Yang Liu. 2024. *Measuring and reducing llm hallucination without gold-standard answers*. *Preprint*, arXiv:2402.10412.

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.

A Appendix

A.1 Gold Answer and Hallucination Examples

The hallucination-free cases and three common error types are shown in Table 2.

A.2 Sample Prompt Template for XCoT

Table 3 shows an example of Instruction and Output for XCoT strategy. XCoT strategy consists of three steps:

1. Perform decomposition and analysis in English for the non-English question.
2. Generate the corresponding English content based on the analysis from the first step.
3. Produce a non-English answer by leveraging English knowledge and reasoning capabilities.

A.3 Comparison of Hallucination-Free Rate with Large-Scale LLMs

We demonstrate that scaling model size and expanding training corpora do not effectively reduce hallucinations in low-resource languages. While English achieves a high hallucination-free rate of 85%, low-resource languages persist at a low 24%. As presented in Table 4, our approach, applied to LLaMA-3.1-8B, consistently surpasses LLaMA-3.1-70B-Instruct in domain-specific question-answering tasks related to Singaporean culture, yielding improvements of 32%, 12%, and 31% in Chinese, Malay, and Tamil, respectively. These findings indicate that merely increasing model size and training data is insufficient for enhancing cross-lingual knowledge generalization, particularly for low-resource languages. In the future, we will apply our method to larger-scale LLMs once sufficient computational resources are available.

A.4 Data Details

In the contrastive learning experiments, we selected a subset of sentence pairs from the WMT-2024 dataset as training data for the EN-ZH language pair. To increase data diversity and ensure sufficient semantic distinction among negative samples within each batch, we extracted 100K sentence pairs with varying lengths and degrees of semantic similarity. For the EN-MS and EN-TA pairs, due to the absence of high-quality open-source datasets, we relied on data collected internally by our team. The data selection criteria and overall volume were kept consistent with those used for EN-ZH.

For the Chain-of-Thought (CoT) data used during the instruction fine-tuning phase, we extracted 3,000 QA pairs from the CRAFT dataset that are non-overlapping with the test set, both in terms of surface question formulations and the underlying knowledge content they assess. CoT annotations were generated using ChatGPT-4o based on these

Golden Answer	Incomplete Response	Factual Error	Irrelevant response
Kampong Spirit refers to the strong sense of community, mutual support, and togetherness found in traditional village life. It originated in early Singapore, where residents of close-knit kampongs relied on one another for daily needs and survival.	Kampong Spirit refers to the strong sense of community, mutual support, and togetherness found in traditional village life.	Kampong Spirit refers to the strong drive to compete and persevere, rooted in the experiences of early Chinese immigrants to Singapore who strived to improve their lives through hard work and mutual competition.	The Kiasu spirit in Singapore reflects a fear of missing out and a strong desire to stay ahead of others. It drives people to be highly competitive, hard-working, and resourceful in all aspects of life.
✓	✗	✗	✗

Table 2: Example responses to question *What is Singapore’s Kampong Spirit? And what are its origins?* for hallucination evaluation.

QA pairs through a 3-shot in-context learning approach. The generated data was manually verified by linguistic experts to ensure quality.

The semantic space alignment test data consists of two parts: the complete FLORES-200 dataset (NLLB Team, 2022) and a subset sampled from our self-collected data. Each test set for a language pair contains 5,000 sentence pairs, with no overlap between the test and training data.

A.5 Case Study

Tables 6, 7, and 8 present the model outputs for the question *What is the origin of bak kut teh, a popular Singaporean dish, and which cultural groups are credited with its creation?* obtained from Gemma-7B after (1) direct fine-tuning, (2) application of the XCoT strategy, and (3) application of the CCL-XCoT strategy, respectively.

Table 6 presents the results of directly fine-tuning large language models. In the Chinese response, the additional explanation of Bak Kut Teh –“它的特色是用红糖、黑糖和白糖等糖混合炖煮而成的浓郁香甜的汤汁 (It features a rich and sweet broth made by stewing with brown sugar, black sugar, and white sugar)”– contains a clear hallucination. Similarly, in the Malay output –“Bak Kut Teh adalah hidangan yang sangat popular di Singapura, dengan akar-akar sejarahnya terletak dalam masakan Cina (Bak Kut Teh is a very popular dish in Singapore, with its historical roots in Chinese cuisine)”– the origin of Bak Kut Teh is not accurately addressed. Furthermore, the Tamil output –“பாக் குட் டே என்பது தாயிலாந்தில் இருந்து வந்த ஒரு தேயிலை வகையாகும். இது பொதுவாக எலுமிச்சை இலைகளும் அரோமாட்டிகான மூலிகைகளும் கலந்து தயாரிக்கப்படுகிறது, பிறகு

ஐஸ் சேர்த்து பரிமாறப்படுகிறது. உள்ளூர் மக்கள் இதைப் பெரும்பாலும் வெப்பத்தை குறைக்கும் ஒரு முறையாகப் பயன்படுத்துகிறார்கள். (Bak Kut Teh is a type of tea that originated from Thailand. It is usually prepared by mixing lemon leaves and aromatic herbs, then served with ice. Locals often use it as a method to reduce heat.)” –completely fails to answer our question, as it incorrectly identifies Bak Kut Teh as a type of Thai tea.

After applying the XCoT strategy, the Chinese response showed some improvement – the obvious hallucination from Table 6 disappeared. However, a new hallucination about the dish’s origin was introduced: “这种菜肴的起源可以追溯到中国，特别是海南人社区。(The origin of this dish can be traced back to China, particularly the Hainanese community.)” In fact, Bak Kut Teh was originally developed by Chinese immigrants from Fujian and Teochew regions in Malaysia and Singapore. We also observed improvements in the Malay response, which correctly included the origin of Bak Kut Teh. However, there is a minor hallucination regarding its ingredients: “Hidangan ini dimasak menggunakan tulang rusuk babi, herba perubatan, rempah ratus, dan sup hitam yang pekat serta dipercayai mempunyai khasiat untuk kesihatan. (This dish is cooked using pork ribs, medicinal herbs, spices, and thick dark soup, and is believed to have health benefits.)” In Tamil, with the application of XCoT, the output started generating more relevant responses. However, it still failed to address the question about which ethnic group is associated with Bak Kut Teh, and included hallucinations in its ingredient descriptions: “பாக் குட் டே என்பது சிங்கப்பூர், மலேசியா போன்ற இடங்களில்

Instruction	请列出新加坡食阁中常见的四种咖啡类型，并比较它们的热量含量。
Output	<p>Problem Analysis:</p> <p>List four common types of coffee typically found in Singapore food courts.</p> <p>Compare the calorie content of these four types of coffee.</p> <p>Think in English:</p> <p>Common types of coffee in Singapore include:</p> <p>Kopi is coffee with sweetened condensed milk.</p> <p>Kopi-C is coffee with evaporated milk and sugar.</p> <p>Kopi-O is coffee with sugar.</p> <p>Kopi-Kosong is plain black coffee without sugar or milk.</p> <p>Since sweetened condensed milk has more calories than evaporated milk, the ranking in terms of calorie content is:</p> <p>Kopi > Kopi-C > Kopi-O > Kopi-Kosong.</p> <p>Answer in Chinese:</p> <p>在新加坡食阁常见的咖啡类型有 Kopi, Kopi-C, Kopi-O 以及 Kopi-Kosong, 它们的卡路里含量是 Kopi > Kopi-C > Kopi-O > Kopi-Kosong。</p>

Table 3: Example of Instruction and Output for XCoT.

Model	English	Chinese	Malay	Tamil
Llama-3.1-70B-Instruct	85%	42%	50%	24%
LLaMA-3.1-8B+CCL-XCoT	94%	74%(+32%)	63%(+13%)	55%(+31%)

Table 4: Comparison of hallucination-free rate between our method and large-scale LLMs. Best results are in **bold**.

பிரபலமான பாரம்பரிய உணவாகும். இது பொதுவாக சீனாவின் புஜியான் பகுதியில் தோன்றியதாகக் கருதப்படுகிறது. இதன் தயாரிப்பு முறை: சூடுகாட்டில் இருந்து பெறப்படும் மூலிகைகள் மற்றும் மணமூட்டும் வாசனைகள் கொண்ட மூலிகைகள் —தாங்குய், கோக்ஸ், டாங்சேன் போன்றவை சேர்த்து காய்ந்த சாறு தயாரிக்கப்படுகிறது, அதில் பன்றி இறைச்சி சேர்த்து பரிமாறப்படுகிறது. (Bak Kut Teh is a traditional dish popular in places like Singapore and Malaysia. It is generally believed to have originated from China’s Fujian region. The preparation involves making a decoction from herbs collected from forests and aromatic herbs like tangkui, goji, and dangshen, then adding pork to serve.)”

After fully applying our method, we observe

that both the Chinese and Malay responses are completely free of hallucinations. The Tamil response, while still somewhat vague regarding the origin —“இதன் தோற்றம் சீன உணவின் வேறுகைகளைக் கொண்டுள்ளது. (Its origin is influenced by variations of Chinese cuisine.)”— still manages to provide a largely complete and correct answer to our question. Compared to the responses generated by directly fine-tuning the large language model, the hallucination rate is significantly reduced.

A.6 Training Details

Our training was conducted on 8 NVIDIA H100 (80GB) GPUs. For continued pre-training, the training time for each model was approximately 3.5 to 4.5 hours. For instruction fine-tuning, the training time for each model was about 30 to 40 minutes. All training was performed using bf16 precision with Low-Rank Adaptation (LoRA)(Hu

Task Name	Language / Language Pairs	Data Volume	Data Description	Data Source	Human-verified
Sentence-level contrastive learning	EN-ZH	100K	Sentence Pair	WMT2024	No
	EN-MS	100K	Sentence Pair	self-collected data	No
	EN-TA	100K	Sentence Pair	self-collected data	No
Paragraph-level contrastive learning	EN-ZH	10K	Paragraph Pair	WMT2024	No
	EN-MS	10K	Paragraph Pair	self-collected data	No
	EN-TA	10K	Paragraph Pair	self-collected data	No
CoT Instruction Finetune	EN-ZH	3K	Question-Answering	GPT-generated(3-shots)	Yes
	EN-MS	3K	Question-Answering	GPT-generated(3-shots)	Yes
	EN-TA	3K	Question-Answering	GPT-generated(3-shots)	Yes
Semantic space alignment test data	EN-ZH	5K	Sentence Pair	FLORES-200+self-collected	No
	EN-MS	5K	Sentence Pair	FLORES-200+self-collected	No
	EN-TA	5K	Sentence Pair	FLORES-200+self-collected	No

Table 5: Task Details and Data Statistics

et al., 2022). The specific training parameters are provided in Table 9 and Table 10.

A.7 Baseline Reproducibility Details

The specific parameters and implementation steps for the experiments in Table 1 are as follows. Gemma-7B+SFT+RAG: we use all-mpnet-base-v2 as our encoding model to calculate the sentence similarity. CrossIn(Lin et al., 2024): The dataset in the original paper is mainly divided into two parts: the Complex Task Dataset and the Linguistic Uniformity Dataset. The Linguistic Uniformity Dataset is closely related to cross-lingual consistency and cross-lingual knowledge transfer. Following the procedures outlined in the paper, we used our original data for CoT instruction tuning data generation and generated CrossX2en and Crossen2X datasets using ChatGPT-4o. These were then combined with Alpaca English for instruction fine-tuning of Gemma-7B and LLaMA-3.1-8B. The instruction fine-tuning parameters were kept consistent with those used in our CoT tasks, as shown in Table 10. CoV(Dhuliawala et al., 2023a): We adopt the best-performing method from the original paper, Factor+Revise, as our validation approach. For each low-resource language, we select five examples from the CRAFT dataset as the basis for 5-shot learning. The Chinese, Malay, and Tamil versions of these five exam-

ples are then input into the base models (Gemma-7B+SFT and Llama-3.1-8B+SFT) to generate initial outputs. We subsequently feed both the questions and their corresponding model outputs into GPT-4o, using the following prompt: "Based on the question and the model output, please generate sub-questions that could be used to verify all factual claims in the output." Each GPT-4o response is then manually verified and revised. For each language, we use the five corresponding examples to guide Gemma-7B+SFT and Llama-3.1-8B+SFT in generating verification sub-questions for each input-output pair. The ground truth answer for each question serves as the final revised output in the Revise step.

A.8 Potential Risks of Our Work

The models and datasets used in our research are sourced entirely from publicly available channels. Our self-collected dataset was obtained through the purchase of data usage rights from publicly available sources. Potential risks are primarily inherent to these existing models and datasets themselves, rather than the methods we have proposed. In the process of constructing our dataset, we carefully selected and manually reviewed the data to ensure it contains no potential risks. Another unavoidable potential risk lies in the generation of model hallucinations. Although we proposed a

method to reduce hallucinations in LLMs, they cannot be completely eliminated. We hope that future researchers who use our method will pay attention to the risks mentioned above.

A.9 Data Sensitivity and Privacy Discussion

The WMT-2024, FLORES-200(NLLB Team, 2022), CRAFT(Wang et al., 2024), SeaEval(Wang et al., 2023) datasets and our self-collected data may contain sensitive data such as personal information or hate speech. At the same time, this problem may also exist in the original training data of Llama-3.1-8B and Gemma-7B. However, our main purpose is to verify the effectiveness of the method and we do not report any experimental results that contain sensitive information or privacy.

A.10 Data Review and Participants

In our work, the data verification process consists of three parts:

1. Verifying whether the model outputs contain hallucinations.
2. Identifying and correcting any errors in the XCoT data generated by ChatGPT.
3. Identifying and correcting any errors in translations of MMLU and LogiQA produced by ChatGPT.

We engaged three linguists proficient in Chinese and English, Malay and English, and Tamil and English, respectively, to assist with the data verification and correction. The requirements for each task are outlined in Table 11. The three annotators are a Chinese national, a Singaporean of Indian descent, and a Singaporean of Malay descent.

A.11 License of Datasets and Models

The licenses of datasets and models used in this paper are provided in Table 12. All of these datasets and tools are public and available for research purposes.

Method	Chinese	Malay	Tamil
Direct Supervised Fine-tuning	<p>肉骨茶是新加坡的一道传统美食，起源于中国广东潮州地区。它是由潮州人在新加坡创造的。肉骨茶的主要原料包括猪肋排骨、豆腐干、芽菜、豆腐皮和辣椒等。它的特色是用红糖、黑糖和白糖等糖混合炖煮而成的浓郁香甜的汤汁。肉骨茶的口味浓郁、香甜，且有浓郁的辣椒味道，非常受新加坡人喜爱。肉骨茶最初是在新加坡的茶楼、餐厅中出现的。潮州人在新加坡创造了肉骨茶，是为了适应当地潮汕移民的口味，融合了潮州菜和当地饮食文化的特点。随着肉骨茶的流行，它也成为新加坡的一道特色美食，吸引了众多食客。</p>	<p>Bak Kut Teh adalah hidangan yang sangat popular di Singapura, dengan akar-akar sejarahnya terletak dalam masakan Cina. Nama "Bak Kut Teh" berasal dari dialek Hokkien, yang bermaksud "teh tulang daging". Hidangan ini terdiri daripada tulang rusuk babi yang direbus dalam kuah herba dan rempah, termasuk bawang putih, bunga lawang, kayu manis, cengkih, dang gui, dan lada hitam. Asal usul Bak Kut Teh agak kontroversi, terdapat dua kisah utama: 1. Pengaruh Hokkien: Sesetengah sumber mengatakan bahawa hidangan ini berasal dari komuniti Hokkien, terutamanya di wilayah Fujian, China. Dikatakan bahawa pendatang Cina dari Fujian membawa hidangan ini ke Malaysia dan Singapura, dan mereka menyesuaikan hidangan ini mengikut rasa dan bahan tempatan. 2. Pengaruh Kantonis: Sumber lain mengatakan bahawa hidangan ini berasal dari komuniti Kantonis, dan versi Fujian adalah variasi. Dikatakan bahawa pendatang Cina dari Guangdong membawa versi Kantonis ke Malaysia dan Singapura, yang mempunyai rasa sedikit berbeza.</p>	<p>பாக் குட் டே" என்பது தாயிலாந்தில் இருந்து வந்த ஒரு தேயிலை வகையாகும். இது பொதுவாக எலுமிச்சை இலைகளும் அரோமாடிகான மூலிகைகளும் கலந்து தயாரிக்கப்படுகிறது, பிறகு ஐஸ் சேர்த்து பரிமாறப்படுகிறது. உள்ளூர் மக்கள் இதைப் பெரும்பாலும் வெப்பத்தை குறைக்கும் ஒரு முறையாகப் பயன்படுத்துகிறார்கள்.</p>

Table 6: The output of directly fine-tuned Gemma-7B.

Method	Chinese	Malay	Tamil
XCoT	<p>肉骨茶是一道受欢迎的新加坡土司菜，影响力不仅仅局限于新加坡，甚至在东南亚其他地区也很受欢迎。这种菜肴的起源可以追溯到中国，特别是海南人社区，他们被认为在 20 世纪初将这种菜肴带到了新加坡。肉骨茶由排骨、药材、香料和一种被认为具有健康益处的浓郁的黑色汤底制成。虽然海南人通常被认为是肉骨茶的创造者，但需要注意的是，这道菜肴随着时间的推移而不断演变和改良，各种文化群体都在其中融合了本地风味和技术。在新加坡，肉骨茶受到不同族群的喜爱，有不同区域的版本，包括潮州、广东和福建等。肉骨茶如今被认为是新加坡的标志性菜肴，既受本地人欢迎，也受访客喜爱。</p>	<p>Bak kut teh ialah hidangan tradisional Singapura yang sangat digemari dan turut popular di seluruh Asia Tenggara. Ia dicipta pada awal abad ke-20 oleh para pendatang awal dari wilayah Chaozhou atau Fujian di China. Hidangan ini dimasak menggunakan tulang rusuk babi, herba perubatan, rempah ratus, dan sup hitam yang pekat serta dipercayai mempunyai khasiat untuk kesihatan. Dengan peredaran masa, pelbagai kumpulan etnik telah membangunkan versi mereka sendiri yang unik, menjadikan bak kut teh sebagai salah satu makanan tempatan ikonik Singapura.</p>	<p>பாக் குட் டே என்பது சிங்கப்பூர், மலேசியா போன்ற இடங்களில் பிரபலமான பாரம்பரிய உணவாகும். இது பொதுவாக சீனாவின் புஜியான் பகுதியில் தோன்றியதாகக் கருதப்படுகிறது. இதன் தயாரிப்பு முறை: சுடுகாட்டில் இருந்து பெறப்படும் மூலிகைகள் மற்றும் மணமூட்டும் வாசனைகள் கொண்ட மூலிகைகள் — தாங்குயி, கோக்சி, டாங் சேன் போன்றவை சேர்த்துக் காய்ந்த சாறு தயாரிக்கப்படுகிறது, அதில் பன்றி இறைச்சி சேர்த்துப் பரிமாறப்படுகின்றது.</p>

Table 7: Model output from Gemma-7B after applying the XCoT method.

Method	Chinese	Malay	Tamil
CCL+XCoT	肉骨茶是一道广受欢迎的新加坡传统菜肴，也风靡于东南亚其他地区。它由早期来自中国潮州或福建地区的移民在 20 世纪初创制，由排骨、药材、香料和浓郁的黑色汤底煮成，具有养生功效。随着时间演变，不同族群发展出各自风味的版本，使肉骨茶成为新加坡标志性的本地美食。	Bak kut teh ialah hidangan tradisional Singapura yang sangat popular dan turut digemari di rantau Asia Tenggara yang lain. Ia dipercayai dicipta pada awal abad ke-20 oleh pendatang awal dari wilayah Chaozhou atau Fujian di China. Hidangan ini dimasak menggunakan tulang rusuk babi, herba, rempah ratus dan sup hitam pekat yang dikatakan mempunyai khasiat untuk kesihatan. Seiring dengan peredaran masa, pelbagai komuniti etnik telah membangunkan versi mereka sendiri yang unik, menjadikan bak kut teh salah satu makanan tempatan ikonik di Singapura. Hidangan ini mudah didapati di medan selera atau pusat penjaja di seluruh Singapura.	பாக் குட் டே, இதன் பொருள் "இறைச்சி எலும்பு தேநீர்" என்பதாகும், சிங்கப்பூரின் பிரபலமான உணவுகளில் ஒன்றாகும். இதன் தோற்றம் சீன உணவின் வேர்களைக் கொண்டுள்ளது. இதன் தோற்றம் சிங்கப்பூரில் உள்ள சீன சமூகத்தில், குறிப்பாக டியோச்சுவே மற்றும் ஹொக்கியன் குழுக்களிடையே பாரம்பரியமாக செய்யப்படும் பாரம்பரிய சீன மூலிகை சூப்ப்களிலிருந்து வந்தது. இந்த சூப்ப்களில் பொதுவாக பன்றி விலா எலும்புகள் அல்லது பிற இறைச்சி வகைகள், மேலும் பல்வேறு மூலிகைகள் மற்றும் மசாலாக்கள் இருந்தன.

Table 8: Model output from Gemma-7B after applying the CCL-XCoT method.

Parameter	Value
Training Batch Size	256
Max Training Epochs	4
Warm Up Steps	500
Lora Rank	16
Lora Alpha	32
Maximum sequence length	4000
Temperature for Contrastive Learning	0.05

Table 9: Training hyperparameters for pretraining stage

Parameter	Value
Training Batch Size	192
Max Training Epochs	3
Warm Up Steps	500
Lora Rank	32
Lora Alpha	64
Maximum sequence length	4000

Table 10: Training hyperparameters for instruction fine-tuning stage

Task	Instruction for Linguists
Hallucination detection	<p>Please help verify whether the model’s output contains hallucinations. We consider the output free of hallucinations if it meets the following two criteria:</p> <ol style="list-style-type: none"> 1. The model’s response accurately and completely addresses all questions in the prompt. 2. The response, including any elaboration beyond the core answer, contains no factual inaccuracies. <p>If it is unclear whether the model’s output contains hallucinations, please refer to the materials we provided or use web search to obtain relevant factual evidence.</p> <p>If the answer contains hallucinations, put ”Y”, otherwise put ”N”.</p>
XCoT data verification and correction	<p>Please help check whether the model’s cross-lingual CoT output is correct. The following aspects need to be reviewed:</p> <ol style="list-style-type: none"> 1. Whether the model’s analysis of the question is accurate. 2. Whether the output in the target language is consistent with the English content. <p>If any issues are found in the model’s output, please help revise them.</p>
Translation verification and correction	<p>Please help verify whether the translations of MMLU, LogiQA and CRAFT are accurate.</p> <ol style="list-style-type: none"> 1. The translations must be free of grammatical errors. 2. The translations should fully and correctly convey the original meaning of the English version. <p>If you find any issues in the translation, please assist by performing post-editing or providing a manual translation.</p>

Table 11: Instructions for Linguists on Evaluation Tasks

Table 12: Model and Dataset Licenses

Model or Dataset	License
meta-llama/Llama-3.1-8B	https://huggingface.co/meta-llama/Llama-3.1-8B/blob/main/LICENSE
meta-llama/Llama-3.1-70B-Instruct	https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct/blob/main/LICENSE
google/gemma-7b	https://ai.google.dev/gemma/terms
CRAFT	https://github.com/SeaEval/CRAFT?tab=Apache-2.0-1-ov-file
WMT2024	<p>Licensing of Data</p> <p>The data released for the General MT task can be freely used for research purposes, we ask that you cite the WMT24 shared task overview findings paper, and respect any additional citation requirements on the individual data sets. For other uses of the data, you should consult with original owners of the data sets.</p>
FLORES-200	https://creativecommons.org/licenses/by-sa/4.0/
Sentence Transformers	https://github.com/UKPLab/sentence-transformers?tab=Apache-2.0-1-ov-file
SeaEval	https://github.com/SeaEval/SeaEval?tab=License-1-ov-file