SEA-NLI: Natural Language Inference as a Lens into Southeast Asian Cultural Understanding

Peerawat Chomphooyod[1,*], **Jian Gang Ngui**[2], **Yosephine Susanto**[2], **Attapol T. Rutherford**[1], **Alham Fikri Aji**[3], **Sarana Nutanong**[4], **Can Udomcharoenchaikit**[4,†], **Peerat Limkonchotiwat**[1,2,†]
[1]Chulalongkorn University [2]AI Singapore [3]MBZUAI [4]VISTEC,
canu_pro@vistec.ac.th, peerat@aisingapore.org
aisingapore/SEA-NLI

Abstract

Frontier LLMs perform well in Western contexts, but remain poorly tested on underrepresented cultures such as those in Southeast Asia (SEA). Existing NLI benchmarks are largely Western-centric, translation-derived, or monolingual, limiting their ability to measure culturally grounded reasoning. We introduce **SEA-NLI**, a native, culturally grounded NLI benchmark covering eight SEA countries in English and native regional languages, verified by native speakers. Across 17 encoder and decoder models, we observe a low performance from all models, especially for knowledge-intensive categories such as Languages and Science and Technology. Our analysis shows that failure cases mainly stem from missing SEA cultural knowledge: SEA-adapted models and culture-aware prompting improve performance, while CoT prompting offers limited gains.

1 Introduction

Frontier large language models (LLMs) excel in Western settings but often underperform in underrepresented, data-scarce cultures, largely due to training data skewed toward dominant Western linguistic and cultural paradigms. Culturally grounded evaluations—including cross-cultural reasoning (Romero et al., 2024; Satar et al., 2025; Lin et al., 2026; Kabir et al., 2026) and local knowledge retrieval (Bogdanova et al., 2026; Li et al., 2026)—are essential for surfacing these biases and improving cultural equity. Prior work also uses natural language inference (NLI) to probe cultural knowledge (Mahendra et al., 2021; Huang and Yang, 2023). In cultural NLI, the correct label can hinge on culturally situated facts; for instance, resolving the following NLI query requires knowing that *Golden Pillow* is a well-known Thai durian cultivar rather than an actual pillow:

> **Premise:** Dan bought a *Golden Pillow* from a Thai fruit stall.
> **Hypothesis:** Dan bought a bedding item.
> **Gold label:** CONTRADICTION.

As summarized in Table 1, existing NLI datasets are predominantly Western-centric, translation-derived, and monolingual, with none combining native multilingual SEA coverage and cultural knowledge. Multilingual datasets are often built by translating English texts (Conneau et al., 2018; Ham et al., 2020; Heredia et al., 2024; Htet and Dras, 2025), which can introduce errors (Agrawal et al., 2024) and obscure local meanings, idioms, sociocultural norms, and value systems (Singh et al., 2025; Susanto et al., 2025). As a result, they often retain Western-centric source perspectives, leaving a gap in resources that faithfully represent Southeast Asian (SEA) cultures. This gap is critical because SEA accounts for 10% of the world's population, with over 700 million people,[1] yet remains underrepresented in cultural understanding research.

Therefore, this paper addresses three main research questions on the SEA NLI dataset construction and evaluation:

- **RQ1:** How to build a robust SEA NLI dataset that accurately reflects SEA cultures?
- **RQ2:** How does LLM performance change when evaluated on SEA cultures rather than Western ones?
- **RQ3:** What gaps remain in current models, and how can they be mitigated?

To answer **RQ1**, we propose **SEA-NLI**, a culturally-grounded NLI dataset for Southeast Asia. SEA-NLI covers cultural topics from eight SEA countries and languages

[*]Work conducted during Research Internship at AI Singapore
[†]Corresponding authors
[1]https://population.un.org/wpp/

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2038–2062
November 6-10, 2026 ©2026 Association for Computational Linguistics

Dataset	Test Size	Lang	Culture	Data Source	Key Contribution
English Benchmarks					
SNLI (Bowman et al., 2015)	10k	EN	Western	Image captions (Flickr30k)	Pioneering large-scale NLI benchmark
MNLI (Williams et al., 2018)	20k	EN	Western	10 Genres (Letters, Fiction, etc.)	Multi-genre cross-domain evaluation
Multilingual & SEA Benchmarks					
XNLI (Conneau et al., 2018)	75k	15	Western	Translation (MNLI)	Cross-lingual consistency benchmark
IndoNLI (Mahendra et al., 2021)	5.2k	IDN	Indo.	Wikipedia, News, and Web	Challenging native-sourced IDN NLI
ViNLI (Huynh et al., 2022)	3k	VNM	Viet.	Online news articles	High-quality VNM NLI
Myanmar XNLI (Htet and Dras, 2025)	5k	MYA	Western	Translation (XNLI)	Burmese low-resource evaluation
Cultural Benchmarks					
CALI (Huang and Yang, 2023)	2.7k	EN	US/India	NormBank	Cultural label disagreement study
SEA-NLI (ours)	**2.1k**	**SEA**	**SEA**	**SEA Wikipedia**	**SEA cultural knowledge evaluation**

Table 1: NLI dataset comparison. SEA-NLI uniquely combines SEA cultural grounding with multilingual coverage.

(`COUNTRY_NAME:LANGUAGE`): Cambodia:KHM, Myanmar:MYA, Malaysia:ZSM, Thailand:THA, Singapore:TAM, Philippines:FIL, Indonesia:IND, and Vietnam:VIE, with texts in both English and SEA languages. By doing so, SEA-NLI also serves as a diagnostic tool: correct English but incorrect SEA suggests a language-understanding gap, failure in both suggests missing cultural knowledge, and correct SEA but incorrect English suggests cross-lingual misalignment.

We construct SEA-NLI via LLM generation with human verification to ensure culturally grounded, high-quality examples. Annotators first select cultural concepts from Wikipedia and draft candidate pairs while mitigating common artifacts (e.g., negation matching, length bias, and noun-phrase variation). We then apply lexical and semantic filters to remove easy cases and regenerate samples until these artifacts are eliminated. Native annotators from all eight countries review and revise each example for cultural accuracy and relevance. Finally, we split SEA-NLI into **normal** (reduced premise–hypothesis overlap) and **hard** (regenerated pairs with no cultural topic overlap).

We further evaluate current models on SEA-NLI (**RQ2** and **RQ3**) through 5 experiments on 17 models. For **RQ2**, SEA-NLI is challenging: average performance drops by 11.60%/13.17% on the hard set in SEA/English compared to the normal set, with low scores on knowledge-intensive categories such as Languages and Science and Technology. For **RQ3**, model failure cases are driven mainly by missing SEA-specific cultural knowledge rather than generic reasoning limits: SEA-adapted models and culture-aware prompting improve performance, while CoT provides limited benefit. These results show that improving SEA-NLI performance requires SEA-specific adaptation, rather than relying only on high-quality English data or model reasoning ability.

Our contributions are as follows.

- We release SEA-NLI, a culturally-grounded NLI dataset for Southeast Asia, reviewed, corrected, and localized by native speakers across eight countries and languages.
- We establish a benchmark for evaluating cultural understanding in SEA and show that both encoder and decoder models experience substantial performance drops on culturally challenging examples, especially in our hard subset.
- By providing each example in a SEA language and English, we enable diagnostic evaluation that disentangles language-understanding gaps, missing cultural knowledge, and cross-lingual misalignment; we further identify failure modes in cultural reasoning, showing that culturally aware prompting improves LLM performance.

2 Related Works

2.1 Natural Language Inference (NLI)

NLI determines the logical relationship between a premise and a hypothesis. Given the premise as true, a model classifies the pair as entailment (the hypothesis is also true), contradiction (the hypothesis is inconsistent with the premise), or neutral (whether the hypothesis is true or not cannot be determined given the premise). Popular NLI datasets, such as SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), and SICK (Marelli et al., 2014), are largely English-centric, reflecting Anglosphere linguistic conventions and cultural perspectives. Multilingual extensions, most notably XNLI (Conneau et al., 2018), translate English texts into other languages. This translation-based paradigm remains common in later language-specific, multilingual, and domain-specific benchmarks (Wijnholds and Moortgat, 2021; Yanaka and Mineshima, 2022; Ebrahimi et al., 2022; Aggarwal et al., 2022; Heredia et al., 2024; Obadić et al., 2023; Htet and Dras, 2025; Oğul et al., 2025). While expanding linguistic coverage, these datasets

often inherit cultural assumptions from English source texts rather than representing local contexts.

To reduce reliance on translated English data, native-source NLI datasets have been developed for non-English languages (Hu et al., 2020; Mahendra et al., 2021; Huynh et al., 2022; Van Huynh et al., 2026). These resources address language-specific data scarcity and better reflect local usage, but do not systematically evaluate cultural knowledge. Beyond data sources, CALI (Huang and Yang, 2023) studies how cultural background affects NLI label disagreement across annotator groups (Plank, 2022), but is limited to English and two cultures. Thus, despite progress in multilingual and native-source NLI, no comprehensive dataset covers major Southeast Asian languages with an explicit focus on SEA cultural and regional knowledge.

2.2 NLI Synthetic Data Generation

Human annotation remains reliable for producing correct NLI examples, but scaling diverse and challenging examples is difficult (Liu et al., 2022). LLMs enable more controlled data generation, especially when combined with filtering or selection strategies to improve quality. Prior work shows that synthetic data targeting domain gaps (Hosseini et al., 2024), spurious correlations (Wu et al., 2022), and complex examples for fine-tuning or distillation (Stacey and Rei, 2024; Stacey et al., 2026) can improve robustness and generalization. Human-AI collaborative frameworks, such as WANLI (Liu et al., 2022) and CALI (Huang and Yang, 2023), further improve synthetic data by combining LLM generation with human oversight, with CALI explicitly eliciting cultural knowledge during annotation. Together, these studies suggest that strategically curated, complex, and culturally informed synthetic data can reduce dataset biases and support more reliable NLI evaluation.

3 SEA-NLI

The creation of the SEA-NLI dataset follows four key steps, as shown in Figure 1. We describe each step as follows.

3.1 Concept Gathering

Following the cultural taxonomy of prior works (Zhang et al., 2025; Alkhamissi et al., 2026), we construct SEA-NLI to evaluate *culture-as-knowledge*, i.e., whether models can apply cultural knowledge to infer the correct NLI label. As shown in Figure 1, we collect 7,904 Wikipedia topics across 10 categories: (i) landmarks, (ii) clothing, (iii) musical instruments, (iv) languages, (v) crime, (vi) science and technology, (vii) cuisine, (viii) education, (ix) politics, and (x) economy, with annotators pre-screening unrelated topics for each country. In addition, we discuss the annotator guidelines in Appendix E.2. For each topic in Wikipedia, we extract its title, category, and summary: the page title, category hierarchy, and introductory paragraph, respectively. We used Wikipedia pages and topics from 2024-2026.

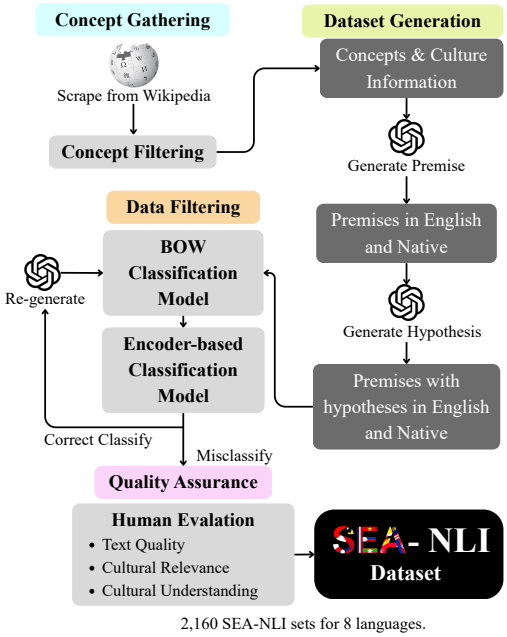


Figure 1: The complete process for creating SEA-NLI.

3.2 Dataset Generation

We use GPT-5.2 to generate SEA-NLI through three stages: (i) premise generation, (ii) hypothesis generation, and (iii) quality improvement. All prompts and examples are provided in Appendix H and C.1.

Premise Generation. Given a topic's title, category, and summary from Section 3.1, GPT-5.2 generates premises in both English and the topic's native language. According to previous works (AlKhamissi et al., 2024; Wang et al., 2024), the system prompt includes the following constraining instructions, to improve cultural accuracy and generation quality:

- **Role-prompting and persona adoption:** Assigning a professional identity and cultural background to ensure culturally relevant premises.
- **Constraint-based generation:** Enforcing strict rules regarding sentence length and the inclusion of localized keywords.

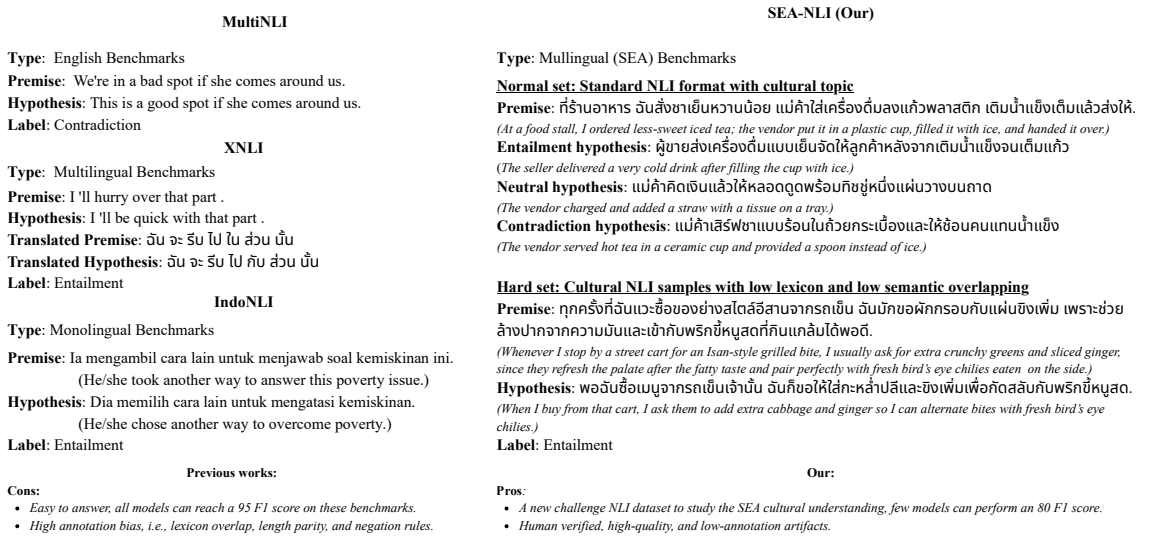


Figure 2: Comparison of SEA-NLI with the existing NLI datasets

- **Concept adaptation:** Allowing the LLM to rewrite concept titles for each country, e.g., shrimp curry to "Phuket shrimp curry", which is clearer for Thai native speakers. If the original title is already specific, it remains unchanged.
- **Concept summary:** Since Wikipedia summaries may include information irrelevant to the target culture, e.g., "Phuket shrimp curry" may include Burmese or Indonesian variants, we let the LLM select only paragraphs related to the target concept.

Hypothesis Generation. Given each generated premise, GPT-5.2 generates hypotheses for the three NLI labels: `entailment`, `neutral`, and `contradiction`. As in premise generation, we apply constraint-based prompting to avoid trivial samples solvable through superficial linguistic patterns. Furthermore, to mitigate annotation artifacts (Zhou and Bansal, 2020; McCoy et al., 2019; Belinkov et al., 2019), we enforce the following rules:

- **Length parity:** To prevent length-based bias, the word count difference between any two hypotheses for the same premise is strictly limited to a maximum of four words.
- **Lexical overlap mitigation:** The prompt explicitly forbids high word overlap between the premise and the entailment hypothesis, a common issue in standard NLI datasets.
- **Anti-negation rules:** Contradictions cannot be formed by simple negations (e.g., merely adding "not" or "no").
- **Noun phrase variation:** The prompt prevents the reliance on generic-to-specific noun phrase substitutions (e.g., replacing "animal" with "dog") to trivially generate entailments.
- **Hedging and intensity bans:** For neutral hypotheses, unverifiable hedging words (e.g., "might," "possibly") and intensity cues are strictly prohibited.
- **Avoidance of superlatives:** Absolute quantifiers (e.g., "all," "none," "always," "only") are banned in neutral and contradiction labels, as models frequently exploit these statistical cues to predict non-entailment.

Quality Improvement. In a pilot study, we found that frequent words were spuriously associated with specific NLI labels, enabling shortcut-based prediction. This problem also occurred in previous NLI datasets (Bowman et al., 2015; Mahendra et al., 2021). To mitigate this problem, we banned these terms in the final generation prompt and used GPT-5.2 to rewrite premises that explicitly explained cultural concepts. These steps reduce lexical and contextual shortcuts, requiring models to rely more on intrinsic cultural knowledge for NLI classification. In addition, we reuse the adapted concept and summary from the premise generation step to improve cultural grounding. Without rewriting, output quality was substantially less reliable, as shown in Appendix C.

3.3 Data Filtering

Despite our best efforts to remove annotation artifacts, there will still be some left in the dataset, including word matching, sentence length, negation, etc. These artifacts may allow for models to try to find shortcuts for prediction (Zhou and Bansal, 2020; McCoy et al., 2019; Belinkov et al., 2019). To remove these problems, we propose a data filtering system using lexical and semantic filters to identify *easy examples* for all premise-hypothesis pairs.

Lexicon Filtering. We extract the lexical features using a Bag-of-Words (BoW) from the English examples of our test set, and train a logistic regression (LR) using BoW as the feature to predict the class of premise-hypothesis pairs[2]. We remove samples that the LR model can correctly predict since these examples are trivially solvable with lexical matching without cultural knowledge requirements.

Semantic Filtering. We further filter easy samples using DeBERTa-v3 (He et al., 2023), a strong NLI encoder model provided by Laurer et al. (2024). Because DeBERTa-v3 lacks SEA cultural knowledge, correctly predicted samples are treated as shortcut-solvable and regenerated until both LR and DeBERTa-v3 misclassify them.

Regenerate Samples. We regenerate samples that fail either lexical or semantic filtering. The regeneration prompt uses stricter premise-hypothesis constraints to reduce annotation bias (Appendix C). Regenerated samples are then filtered again with the same systems. We repeat this for four rounds: only 2.07% (492/23,712) pass in the first round, with later rounds adding 1.23%, 1.28%, and 0.82%; we stop after four rounds due to diminishing returns.

Sample Splitting. As shown in Figure 2, we split SEA-NLI into *normal* and *hard* sets. The *normal* set contains samples that pass the first filtering round without regeneration, with each premise paired with three hypotheses to follow the standard NLI format. The *hard* set contains regenerated premise-hypothesis pairs, making it imbalanced but more culturally demanding. Together, the *normal* and *hard* sets test whether models can move beyond lexical or semantic shortcuts as crutches and rely on deeper SEA cultural understanding.

3.4 Quality Assurance

To ensure all samples met our quality threshold, we recruited 27 annotators from 8 countries to validate the dataset. We describe the annotator metrics and process below. Full annotator details and guidelines are provided in Appendix E.

Task Metrics. Annotators checked four aspects: (i) label correctness, (ii) cultural relevance of each topic (Likert scale 1-5), (iii) cultural understanding required for each review topic (Likert scale 1-5), and (iv) issue flags (see Appendix E.3). We use a majority vote for the final label and average scores within each annotator group.

[2]We cannot apply this to SEA languages because no robust word tokenizer covers all of them.

Task Guidelines. Annotators reviewed SEA-NLI in three steps. First, they evaluated GPT-5.2 output generated using pilot prompt techniques from Section 3.2 (120 samples per country) to validate baseline translation, labels, and cultural relevance (see Appendix B). Second, we verified labels and LLM outputs by reviewing examples generated without filtering or refinement. Third, they reviewed the final SEA-NLI to ensure that regenerated samples are high-quality and require stronger cultural knowledge. We expected that agreement and cultural-relevance scores of the *second* and *third* tasks from annotators would be better than those of the first task.

Results from Annotators. Annotator agreement is consistently high across the three steps (0.97, 0.99, and 0.96), while cultural relevance increases from 4.15 to 4.36 and 4.47. This highlights that our datasets are more culturally relevant than vanilla GPT-5.2-generated texts. Also, the regenerated samples did not damage any semantic or quality of data, as we got higher quality scores from annotators compared to samples before the regenerated process (the third task). We also find that 8.89% of labels were changed and 1.48% of samples were flagged for quality issues. These results show that GPT-5.2 can generate reliable SEA-NLI candidates when properly prompted, but human validation remains necessary to ensure label correctness, quality, and cultural relevance. All the scores and how we calculate them are discussed in Appendix E.4.

3.5 SEA-NLI Summary

We summarize the dataset statistics and distributions in Figure 4. SEA-NLI contains 1,443 *normal* and 717 *hard* samples across 10 cultural categories, 8 countries, and 8 SEA languages. SEA-NLI is the first NLI dataset focused on Southeast Asian cultural knowledge. We ensure dataset quality through full annotator review of labels, cultural relevance, and translations, with annotator results confirming that the samples are culturally relevant and of high quality. Each sample includes both the target culture's native language and an annotator-verified English translation from GPT-5.2, enabling analysis of SEA-English inconsistency. Full statistics are provided in Appendix 7. In addition, we conduct a bias study in SEA-NLI in Appendix K.

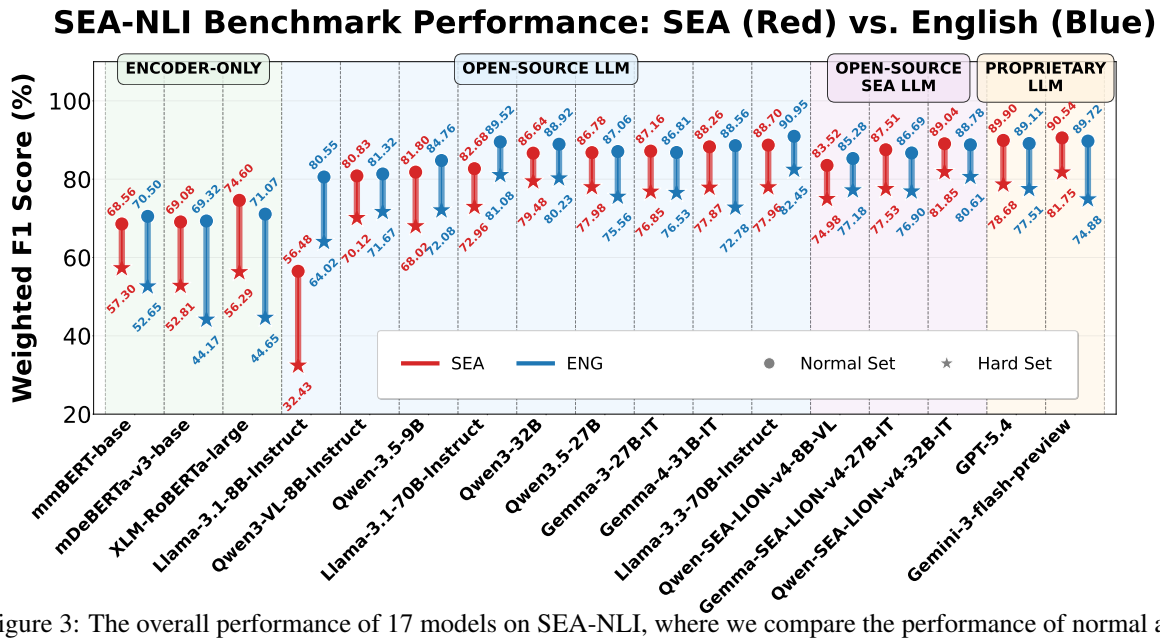


Figure 3: The overall performance of 17 models on SEA-NLI, where we compare the performance of normal and hard sets on SEA and English languages.

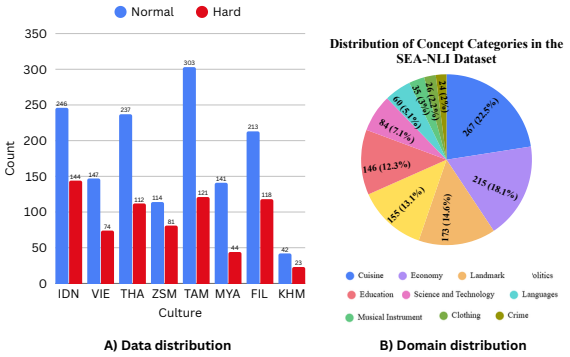


Figure 4: Data statistics of SEA-NLI.

4 Experimental Setup

Models. To answer **RQ2** and **RQ3**, we employ 17 models across encoder-based and decoder-based models to be evaluated on the SEA-NLI benchmark. For the encoder-based evaluation, we fine-tune the pre-trained base models: XLM-R (Conneau et al., 2020), mmBERT (Marone et al., 2025), and mDeBERTa (He et al., 2023), using SNLI (Bowman et al., 2015) and XNLI (Conneau et al., 2018) datasets. For the decoder-based architectures, we evaluate the models in a zero-shot setting utilizing Qwen-3.5 (Yang et al., 2025), Llama-3.1/3.3 (Grattafiori et al., 2024), and the SEA regionally specialized LLMs: SEA-LION-Qwen-v4 and Gemma-SEA-LION-v4 (Ng et al., 2025), where the prompt is detailed in Figure 20a. Additionally, we evaluate leading proprietary models via API, namely GPT-5.4 and Gemini-3-Flash.

Metrics. We evaluate the NLI classification performance using the weighted-averaged F1-score, computed on both SEA languages and English-translated versions of the dataset separately.

5 Experimental Results

5.1 Main Results

To answer **RQ2**, we summarize results from 17 models on SEA-NLI normal and hard sets in Figure 3. Full results are reported in Table 15.

Normal vs. Hard Sets. The hard set is substantially more difficult than the normal set for all model families. Averaged across all models, performance drops from 81.89/84.05 on the normal set to 70.29/70.88 (11.60%/13.17% drop) on the hard set in SEA/English. This degradation is especially pronounced for encoder-based models, indicating that the hard set removes many shortcut-solvable examples and requires stronger reasoning beyond superficial lexicon or generic semantic cues.

English vs. SEA Languages. Cross-lingual robustness remains inconsistent. On the normal set, many models perform slightly better in English, but this pattern is less stable on the hard set, where the SEA–English gap widens. For example, `Llama-3.1-8B-Instruct` and other Llama models show much larger SEA–English gaps than other models. This suggests that strong English NLI performance does not necessarily transfer to SEA languages in culturally challenging settings.

Effect of SEA Adaptation. As shown in Figure 3, we also observe that SEA-adapted models generally improve over their corresponding base models, especially on SEA-language evaluation. For example, `Qwen-SEA-LION-v4-8B-VL` outperforms `Qwen3-VL-8B-Instruct` by 2.69/3.96 points on the normal set and 4.86/5.51 points on the hard set. These results emphasize that both SEA-specific

adaptation and model scaling are important for robust culturally grounded NLI.

Summary. The hard set introduces a new NLI challenge that is less solvable by lexical overlap, sentence patterns, or generic NLI reasoning alone. Because these samples survive both the lexical and semantic filters, models must rely more heavily on culturally-grounded knowledge to infer the presence of an entailment relation correctly. The broad performance drop across all models, including frontier models, highlights cultural understanding as a key remaining challenge in SEA-NLI.

5.2 Category Results

We answer **RQ2** by evaluating the performance of models using the category results. As shown in Figure 5, model performance varies substantially across cultural concept categories. Overall, models achieve the strongest results on more visually or commonly represented cultural topics, such as Musical Instrument, Cuisine, and Landmark, with average F1 scores around 72%-78% across the dataset. In contrast, categories that require more fine-grained cultural or contextual knowledge, especially Languages (e.g., SEA languages and dialects such as Malayo-Polynesian, Wa, and Cua) and Science and Technology (i.e., Singtel, AXN, and Kompas), are consistently more challenging, with average F1 scores dropping to 58%-62%. This suggests that SEA-NLI not only tests surface-level cultural recognition, but crucially also exposes weaknesses in models' ability to reason over culturally grounded and knowledge-intensive concepts. Note that we demonstrate the category results using English texts in Appendix G, where we also observe the same trend as SEA languages.

6 Insights for Future Model Development

To answer **RQ3**, we conduct three studies to identify challenges and gaps, and to improve the model's ability to understand more SEA cultures.

6.1 Understand the Challenge of Hard Set

To explain the performance drop on the *hard* set, we test whether success depends on explicit cultural keyword overlap. Many *hard* examples omit the target concept, so correct predictions require cultural understanding rather than surface lexical cues. For instance, given a premise containing *Hokkien mee* and a hypothesis mentioning "noodles", a model can predict entailment only if it knows that "mee" refers to "noodles" in this cultural context. We therefore compare F1 between samples with and without concept occurrence in the premise or hypothesis.

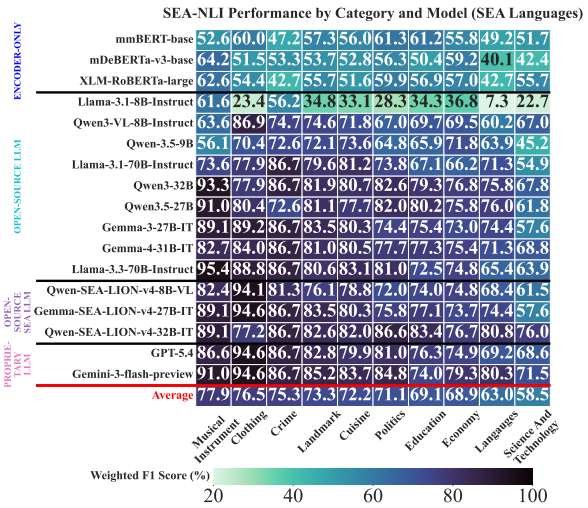


Figure 5: Weighted F1 performance across cultural concept categories on SEA languages (Hard set). We mixed all languages within the same category.

As shown in Table 2, models perform worse when the cultural concept appears in neither the premise nor the hypothesis. When the concept is absent, models need to understand the concept behind those topics without relying on superficial cues. In addition, we notice that the SEA-adapted model, Gemma-SEALION-27B, always performs better than the original model in all cases. These results suggest that SEA adaptation injects useful regional knowledge into the model, making it more resilient to the lack of explicit cultural keywords. Thus, the hard set is challenging because models must infer cultural background beyond surface-level lexical matching, but SEA-specialized models are better equipped to handle this problem.

Model	Concept Occurrence	Count	SEA	ENG
Gemma-3-27B-it	Both + Once	256	77.41	77.66
	Neither	461	76.51	75.90
Gemma-SEA-LION-v4-27B	Both + Once	256	78.39	78.46
	Neither	461	77.00	76.02

Table 2: The analysis of hard set on cultural keyword overlapping using weighted F1 score.

6.2 Culture-aware Prompting

LLMs are trained on large-scale corpora, including Wikipedia (Grattafiori et al., 2024; Team et al., 2025; Ng et al., 2025), which may contain substantial knowledge about SEA cultures. However, such knowledge may remain latent and require targeted prompting to be reliably elicited. We therefore study culture-aware prompting strategies for

extracting SEA-specific knowledge from LLMs, including cultural persona (`Cult.`), topic summaries (`Sum.`), target topics (`Topic`), in-context examples (`ICL`), and reasoning tokens (`CoT`); full prompts are provided in Appendix 20.

In Table 3, `Cult.+Sum.` yields the largest gains, and other culture-enriched variants also improve SEA adaptation, suggesting that explicit SEA context helps models retrieve relevant knowledge. In contrast, `CoT` does not improve performance in most cases. Appendix J shows that reasoning prompts often shift models toward lexical overlap, leading to over-prediction of `neutral`. For example, in Figure 21, Gemma-SEA-LION-v4-27B understands the Thai uniform context but tries to map the premises keywords to the hypothesis. Without a keyword match, it defaults to `neutral` instead of inferring the cultural concept. We observe a similar pattern with `ICL`, where demonstrations encourage lexical matching over cultural understanding. Overall, these results suggest that SEA-NLI calls for cultural-knowledge adaptation more than generic reasoning elicitation.

		Gemma-3 27B-IT		Gemma-SEA LION-v4-27B	
Set	**Method**	**SEA**	**ENG**	**SEA**	**ENG**
Normal Set	Base	87.16	87.06	87.51	86.69
	Base+Cult.	87.12	87.92	87.19	88.06
	Base+Cult.+Topic	87.76	88.18	88.18	88.68
	Base+Cult.+Sum.	**89.99**	**91.46**	**90.48**	**91.46**
	CoT+Cult.	84.59	86.05	86.48	86.81
	CoT+Cult.+Sum.	86.67	88.17	87.30	89.28
	CoT+ICL+Cult.	85.31	85.96	85.94	86.70
Hard Set	Base	76.85	76.53	77.53	76.90
	Base+Cult.	76.42	76.04	77.27	76.34
	Base+Cult.+Topic	75.92	77.59	76.58	77.64
	Base+Cult.+Sum.	**77.59**	78.15	**77.85**	**78.44**
	CoT+Cult.	75.73	76.17	76.06	75.61
	CoT+Cult.+Sum.	75.95	**78.17**	77.68	78.28
	CoT+ICL+Cult.	76.28	75.56	76.57	75.48

Table 3: Weighted F1-score performance (%) for each prompt variant.

6.3 Error Analysis

To analyze model failure cases, we compare predictions on parallel English and SEA language NLI samples using 17 models, as shown in Figure 6. We group errors into three categories: *language misunderstanding* (ENG ✓/SEA ×), *cultural knowledge deficits* (ENG ×/SEA ×), and *cross-lingual misalignment or translation loss* (ENG ×/SEA ✓). The full results are shown in Table 8. We summarize the results as follows.

Cultural Knowledge Deficits Dominate Model Errors. Most errors are cultural knowledge deficits, accounting for 8.43%-25.56% of cases across models. Even top models show non-trivial rates (∼9–11%), suggesting that many failures stem from gaps in SEA-specific semantic and cultural knowledge rather than translation artifacts. For example, on "Laksa," all models fail to distinguish sour and spicy laksa, which differ in taste and cooking method, and predict `neutral`. Models that fail on English samples also typically fail on SEA-language samples, indicating shared deficiencies rather than language-specific noise.

SEA Language Understanding Remains a Major Bottleneck. Language misunderstanding is the second-largest error category, accounting for 3.15%-30.56% of errors. This is especially common in smaller models such as LLaMA-3.1-8B (30.56%), suggesting reliance on English-centric representations. By contrast, cross-lingual misalignment remains low (3%-8%), indicating that models rarely succeed in SEA languages when failing in English. Thus, model knowledge appears more accessible in English, limiting real-world SEA applications where users need reliable answers in their native languages.

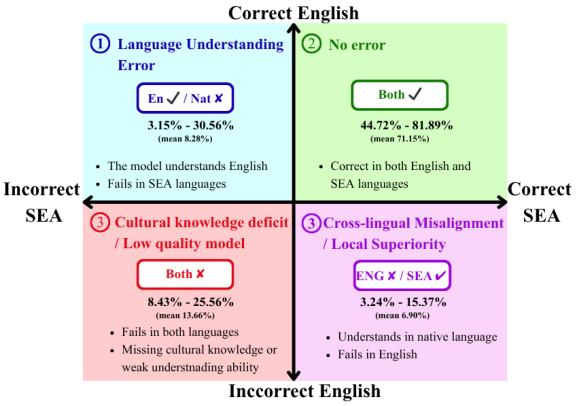


Figure 6: Error taxonomy mapping English and SEA prediction outcomes to distinct model deficiencies.

7 Conclusion

We introduce SEA-NLI, a culturally grounded NLI benchmark for evaluating encoder- and decoder-based models in Southeast Asian contexts. Our evaluation of 17 models shows that SEA-NLI remains challenging even for strong frontier and SEA-adapted models. Our analysis suggests that this degradation is driven less by generic reasoning failures than by missing SEA cultural knowledge: models struggle when cultural concepts are implicit, while culture-aware prompting improves performance more reliably than chain-of-

thought. Cross-lingual error analysis further indicates that failures are dominated by cultural knowledge deficits and SEA language misunderstanding, rather than translation misalignment. Together, these findings highlight the need for models adapted not only to Southeast Asian languages but also to Southeast Asian cultural knowledge.

Limitation

This study is subject to limitations related to the full range of linguistic and cultural diversity across all Southeast Asia countries. The evaluation is restricted to a selected subset of languages and does not account for all dialects, registers, or code-switching practices. These constraints limit the generalization of the findings and suggest that further work is encouraged to develop more comprehensive and locally grounded NLI datasets.

Similar to other benchmark works, we did not present a new model that mitigates the SEA safety problem. However, we dedicate the whole Section 6 to how to achieve a high score on our benchmark. We present both errors and cultural knowledge studies for future work that are interesting to work on the SEA NLI problems. This is an insight for future works to improve the SEA cultural understanding.

Due to constrained annotator availability within the project timeframe, the dataset is limited to 2,160 samples and framed as a targeted diagnostic benchmark rather than a broad-coverage corpus. To ensure reliability, analysis is restricted to major aggregate categories or cultural levels independently, rather than their isolated, micro-level intersections. Furthermore, we conduct Cross-model variance analysis confirms these smaller cultural subsets exhibit lower variance than larger subsets, proving a stable and statistically reliable signal despite their compact size in Appendix N.

Moreover, since we use generative AI for data creation, there might be another bias in the dataset. However, we did check with humans and previous works (Gururangan et al., 2018; Zhou and Bansal, 2020; McCoy et al., 2019; Belinkov et al., 2019) and not found any patterns for models to find a shortcut in our dataset, i.e., negation rule, length prediction, lexicon matching, semantic matching, etc., as shown in Appendix K.

Ethical Statement

All annotators were native speakers of SEA languages and were of legal working age (above 18 years old). The majority of annotators were students enrolled at local universities and participated on a voluntary, compensated basis. They were recruited through a rigorous screening and evaluation process. Prior to their participation, they received comprehensive training to ensure informed, accurate, and responsible contributions. Annotators were compensated at a rate of $17 USD per hour, which exceeds the typical market rate for comparable annotation work, reflecting a commitment to fair and ethical labor practices. In addition, we also have an IRB for this project, allowing us to leverage the annotator to help us check and revise the cultural datasets for SEA languages.

Acknowledgments

This project is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. We also acknowledge the support of the Faculty of Arts, Chulalongkorn University in the form of an international travel grant, under the Overseas Conference Presentation Grant (Fiscal Year 2026).

References

Divyanshu Aggarwal, Vivek Gupta, and Anoop Kunchukuttan. 2022. IndicXNLI: Evaluating multilingual inference for Indian languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10994–11006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Agrawal, Barah Fazili, and Preethi Jyothi. 2024. Translation errors significantly impact low-resource languages in cross-lingual learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 319–329, St. Julian's, Malta. Association for Computational Linguistics.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Mai Alkhamissi, Yunze Xiao, Badr AlKhamissi, and Mona T. Diab. 2026. Hire your anthropologist! rethinking culture benchmarks through an anthropological lens. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1218–1235, Rabat, Morocco. Association for Computational Linguistics.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. On adversarial removal of hypothesis-only bias in natural language inference. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 256–262, Minneapolis, Minnesota. Association for Computational Linguistics.

Liliia Bogdanova, Shiran Sun, Lifeng Han, Natalia Amat Lefort, and Flor Miriam Plaza del Arco. 2026. Flans at semeval-2026 task 7: Rag with open-sourced smaller llms for everyday knowledge across diverse languages and cultures. *Preprint*, arXiv:2603.01910.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, Vicky Feliren, Bahrul Ilmi Nasution, Manuel Antonio Rufino, Genta Indra Winata, Rian Adam Rajagede, Carlos Rafael Catalan, and 73 others. 2025. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Ronald A Fisher. 1921. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Jiyeon Ham, Yo Joong Choe, Kyubyong Park, Ilji Choi, and Hyungjoon Soh. 2020. KorNLI and KorSTS: New benchmark datasets for Korean natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 422–430, Online. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Maite Heredia, Julen Etxaniz, Muitze Zulaika, Xabier Saralegi, Jeremy Barnes, and Aitor Soroa. 2024. XNLIeu: a dataset for cross-lingual NLI in Basque. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4177–4188, Mexico City, Mexico. Association for Computational Linguistics.

Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. A synthetic data approach for domain generalization of NLI models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 2212–2226, Bangkok, Thailand. Association for Computational Linguistics.

Aung Kyaw Htet and Mark Dras. 2025. Myanmar xnli: building a dataset and exploring low-resource approaches to natural language inference with myanmar. *Language Resources and Evaluation*, 59(3):3267–3310.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence Moss. 2020. OCNLI: Original Chinese Natural Language Inference. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3512–3526, Online. Association for Computational Linguistics.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2022. ViNLI: A Vietnamese corpus for studies on open-domain natural language inference. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3858–3872, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Mohsinul Kabir, Tasnim Ahmed, Md Mezbaur Rahman, Shaoxiong Ji, Hassan Alhuzali, and Sophia Ananiadou. 2026. Xcr-bench: A multi-task benchmark for evaluating cultural reasoning in llms. *Preprint*, arXiv:2601.14063.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and bert-nli. *Political Analysis*, 32(1):84–100.

Jiaang Li, Yifei Yuan, Wenyan Li, Mohammad Aliannejadi, Daniel Hershcovich, Anders Søgaard, Ivan Vulić, Wenxuan Zhang, Paul Pu Liang, Yang Deng, and Serge Belongie. 2026. Ravenea: A benchmark for multimodal retrieval-augmented visual culture understanding. *Preprint*, arXiv:2505.14462.

Peiqin Lin, Chenyang Lyu, Wenjiang Luo, Haotian Ye, Md Mehrab Hossain, Chunlan Ma, Shaoxiong Ji, Younes Samih, Bo Zeng, Fan Jiang, Yuanbin Cao, Dilda Duisenbek, Adrian Neo Sau Xun, Daria Pozdniakova, Liubou Misevich, Nevena Marinković, Ngoc Gia Linh Nguyen, Thi Khanh Linh Do, Sarakmatak Sophy, and 6 others. 2026. Culturall: Benchmarking multilingual and multicultural competence of llms on grounded tasks. *Preprint*, arXiv:2604.19262.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rahmad Mahendra, Alham Fikri Aji, Samuel Louvan, Fahrurrozi Rahman, and Clara Vania. 2021. IndoNLI: A natural language inference dataset for Indonesian. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10511–10527, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Marc Marone, Orion Weller, William Fleshman, Eugene Yang, Dawn Lawrie, and Benjamin Van Durme. 2025. mmbert: A modern multilingual encoder with annealed language learning. *Preprint*, arXiv:2509.06888.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Raymond Ng, Thanh Ngan Nguyen, Huang Yuli, Tai Ngee Chia, Leong Wai Yi, Wei Qi Leong, Xianbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Adithya Venkatadri Hulagadri, Kok Wai Teng, Yeo Yeow Tong, Bryan Siow, Wei Yi Teo, Tan Choon Meng, Brandon Ong, and 11 others. 2025. SEA-LION: Southeast Asian languages in one network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 512–526, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Leo Obadić, Andrej Jertec, Marko Rajnović, and Branimir Dropuljić. 2023. C-XNLI: Croatian extension of XNLI dataset. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2258–2267, Toronto, Canada. Association for Computational Linguistics.

İskender Ülgen Oğul, Fatih Soygazi, and Belgin Ergenç Bostanoğlu. 2025. TurkMedNLI: a turkish medical natural language inference dataset through large language model based translation. *PeerJ Comput Sci*, 11:e2662.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, and 57 others. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Burak Satar, Zhixin Ma, Patrick Amadeus Irawan, Wilfried Ariel Mulyawan, Jing Jiang, Ee-Peng Lim, and Chong-Wah Ngo. 2025. Seeing culture: A benchmark for visual reasoning and grounding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22227–22243, Suzhou, China. Association for Computational Linguistics.

Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428. PMID: 18839484.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Joe Stacey, Lisa Alazraki, Aran Ubhi, Beyza Ermis, Aaron Mueller, and Marek Rei. 2026. Improving the OOD performance of closed-source LLMs on NLI through strategic data selection. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5378–5404, Rabat, Morocco. Association for Computational Linguistics.

Joe Stacey and Marek Rei. 2024. Distilling robustness into natural language inference models with domain-targeted augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2239–2258, Bangkok, Thailand. Association for Computational Linguistics.

Yosephine Susanto, Adithya Venkatadri Hulagadri, Jann Railey Montalan, Jian Gang Ngui, Xianbin Yong, Wei Qi Leong, Hamsawardhini Rengarajan, Peerat Limkonchotiwat, Yifan Mai, and William Chandra Tjhi. 2025. Sea-helm: Southeast asian holistic evaluation of language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12308–12336.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Tin Van Huynh, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. 2026. A new benchmark dataset and mixture-of-experts language models for adversarial natural language inference in vietnamese. *Expert Systems with Applications*, 306:130109.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Gijs Wijnholds and Michael Moortgat. 2021. SICK-NL: A dataset for Dutch natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1474–1479, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676, Dublin, Ireland. Association for Computational Linguistics.

Hitomi Yanaka and Koji Mineshima. 2022. Compositional evaluation on Japanese textual entailment and similarity. *Transactions of the Association for Computational Linguistics*, 10:1266–1284.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xinyu Zhang, Pei Zhang, Shuang Luo, Jialong Tang, Yu Wan, Baosong Yang, and Fei Huang. 2025. CultureSynth: A hierarchical taxonomy-guided and retrieval-augmented framework for cultural question-answer synthesis. In *Findings of the Association*

for Computational Linguistics: EMNLP 2025, pages 10448–10467, Suzhou, China. Association for Computational Linguistics.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying NLI models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771, Online. Association for Computational Linguistics.

Appendix

A Data Statistics

This section summarizes the key statistics of the SEA-NLI dataset. The dataset exhibits a long-tailed distribution; while Singapore (TAM) and Indonesia (IDN) provide the highest sample volumes, the inclusion of smaller subsets such as Cambodia ensures broad regional representation. Across all cultures, the normal set consistently maintains a higher volume than the hard set, typically following a 2:1 ratio as shown in Figure 7. In addition, Figure 8 shows that the normal subset has balanced entailment labels, while the hard subset exhibits greater variance across cultures.

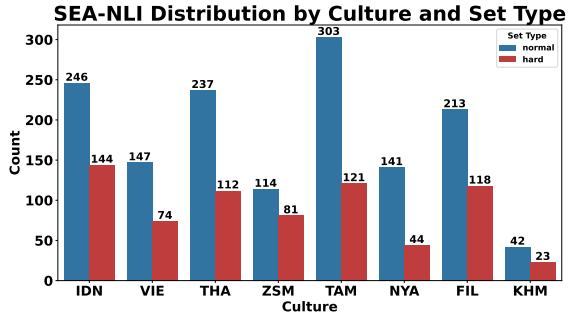


Figure 7: Distribution of SEA-NLI samples across eight cultures, partitioned into normal and hard subsets.

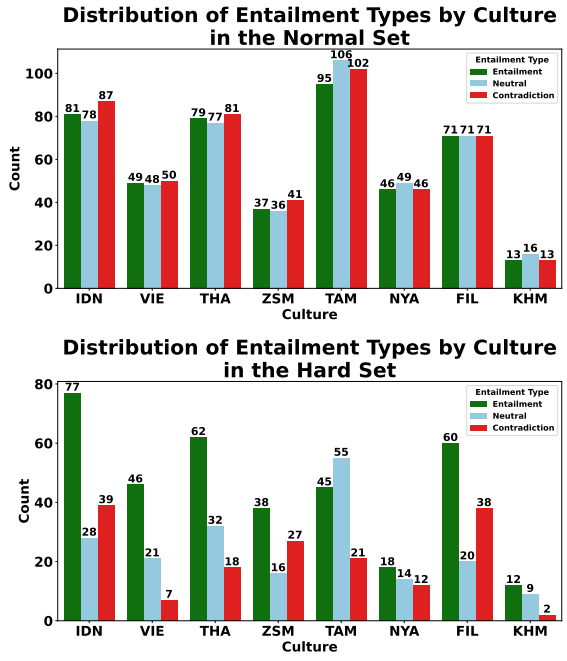


Figure 8: Comparison of Entailment Type Distributions between Normal and Hard Sets in the SEA-NLI Dataset by Country.

B Preliminary Pilot Set

To assess the quality of the LLM-generated NLI samples (GPT-5.2), we conduct a study using the premise and hypothesis generation prompts in Figures 9 and 10, respectively. Initial analysis of 2,400 premise–hypothesis pairs revealed that frequent words were spuriously associated with specific NLI labels, introducing potential annotation artifacts (Figure 11a). To mitigate these biases, we refined the prompting strategy, as detailed in section H. Moreover, we sampled 120 premise–hypothesis pairs per culture for human evaluation. Detailed pilot results regarding inter-annotator agreement, cultural relevance, and linguistic naturalness are provided in Appendices E.4.



```
system_prompt = f"""
Objective: As a Cultural Specialist and Computational Linguist, your task is to generate {n_premise} unique Natural Language Inference (NLI)
premises based on the provided cultural contexts. The goal of this task is to test LLMs' cultural knowledge in SEA contexts. You must ensure the
context reflects the "World Model" and social logic of Southeast Asia (SEA). For EACH premise you generate, you must adopt the specific 'persona'
provided within its `metadata.persona` in the input context for a nuanced cultural perspective.

Task Instructions:
You will be provided with a JSON array named `premise_contexts`. This array contains {n_premise} separate contexts. For EACH context, you must
generate a unique "Premise" along with its metadata.

Each generated premise must:
- Be exactly one sentence long.
- Describe a specific SEA cultural situation, social norm, or tradition relevant to the 'Concept title'and 'Sub Concept' in the provided context.
- Reflect localized logic (e.g., specific behaviors, etiquette, or environmental interactions unique to the 'Target Country' in the provided context).
- Use localized keywords in the 'Target Language' of the provided context (e.g., using "ทุเรียน" instead of "Durio zibethinus" for a Thai context).
- For the English premise, use only English words (no code-switching).
- Employ diverse sentence structures and contexts to ensure the premises do not follow a repetitive pattern.

For each generated premise, you must also provide the following metadata:
- Cultural Relevance: (5 = Unique/Iconic to SEA; 1 = Generic/Unrelated).
- Cultural Understanding: Your expertise on this specific topic (based on the context).
- Quality Score: Clarity, grammar, and adherence to all instructions.
- Flags: Any issues (e.g., "Linguistic Error", "Factual Inaccuracy", "Ambiguous Context"). Use ["None"] if no issues.
- Persona: You MUST copy the `persona` object exactly as it is provided in the input context for that specific premise.


Your final output MUST be a single JSON array containing {n_premise} objects. Each object must represent a generated premise and its associated
metadata, and the order of the output objects MUST correspond to the order of the `premise_contexts` provided.

The structure for EACH generated object in the output array should be:
{{
    "premise_native": "...",
    "premise_english": "...",
    "culture": "...",
    "concept_category": "...",
     "concept_sub_category": "...",
    "concept_title": "...",
    "concept_description": "...",
    "metadata": {{
            "culture_relevance_score": ...,
            "cultural_understanding_score": ...,
            "quality_score": ...,
            "flags": [<FLAG>],
            "persona": {{
                    "nationality": "...",
                    "generation": "...",
                    "age": "...",
                    "gender": "...",
                    "trait": "..."
            }}
    }}
}}

Here is the JSON array of `premise_contexts` for which to generate premises:

{
    "target_country": "thai",
    "concept_title": "Siamotyrannus",
    "concept_category": "economy",
    "concept_sub_category": "Fossils of Thailand",
    "concept_description": "Siamotyrannus (meaning "Siamese tyrant") is a genus of theropod dinosaur from the early Cretaceous of Thailand. In
1993, Somchai Traimwichanon found a partial skeleton of a large theropod at the Phu Wiang 9 site in Khon Kaen.
In 1996, Eric Buffetaut, Varavudh Suteethorn and Haiyan Tong named and described the type species Siamotyrannus isanensis. The generic name is
derived from the old Thai kingdom of Siam, and a Latinised Greek tyrannus, meaning "tyrant", in reference to a presumed membership of the
Tyrannosauridae. The specific name is derived from Thai isan, "northeastern part", referring to the provenance from northeast Thailand...",
    "target_language": "Thai",
    "metadata": {
            "persona": {
                "age": 26,
                "gender": "Non-Binary",
                "nationality": "thai",
                "generation": "Gen z",
                "trait": "pragmatic"
            }
    }
},

... Continued up to 10 inputs.
```

Figure 9: The premise generation prompt for the pilot dataset.



Figure 10: The hypothesis generation prompt for the pilot dataset.

C Iterative Refinement and Dataset Improvement

To enhance the quality of the SEA-NLI dataset, we employ an iterative refinement loop using GPT-5.2, as detailed in Section 3.3. This process is designed to mitigate annotation artifact bias.

C.1 Regeneration Prompt

Figure 12 illustrates the prompt template used for sample regeneration. The prompt accepts a premise, its corresponding hypotheses in both English and Southeast Asian languages, and specific concept metadata. The prompt also enforces rigorous quality control by anchoring all hypotheses to a single underlying fact and ensuring logical consistency. To maximize robustness, an "anti-shortcut" design is utilized; this strategy penalizes lexical overlap, necessitates deep regional cultural grounding, and mandates that all sentences remain logically plausible and concise.

C.2 What Did We Get After The Refinement Process?

The refinement process results in a measurable growth in sample length. As shown in Figure 13, the average character length of premises and hypotheses accepted at each stage increases consistently from Step 1 through Step 4 across all languages. Moreover, the length that increased also means the explanation that increased for each premise–hypothesis pair. As shown in Figure 2 in the main text, the hard sample has a longer context than the normal sample since it tries to explain the concept of each cultural topic.

Figure 14 provides a qualitative example of how a sample evolves, becoming harder to classify without deep semantic understanding. Before the filtering and refinement loop is applied, the original sample can be solved through simple lexical or keyword matching (e.g., The Sail at Marina Bay→ tall tower → tall building). Consequently, the stacked classifier used in the filtering loop, which combines Logistic Regression (LR) and DeBERTa-v3, correctly predicts the entailment relationship without requiring any cultural knowledge. The refined sample in step 4 cannot be solved through straightforward keyword matching. Correctly predicting the entailment requires the model to know that The Sail @ Marina Bay is a private condominium development in Singapore. As a result, both LR and DeBERTa-v3 fail to predict the correct entailment label. In addition to cultural knowledge requirements, the hard set intentionally includes complex or atypical phrasing. These syntactic distractors are designed to test whether LLMs can maintain accurate context tracking while applying cultural knowledge.

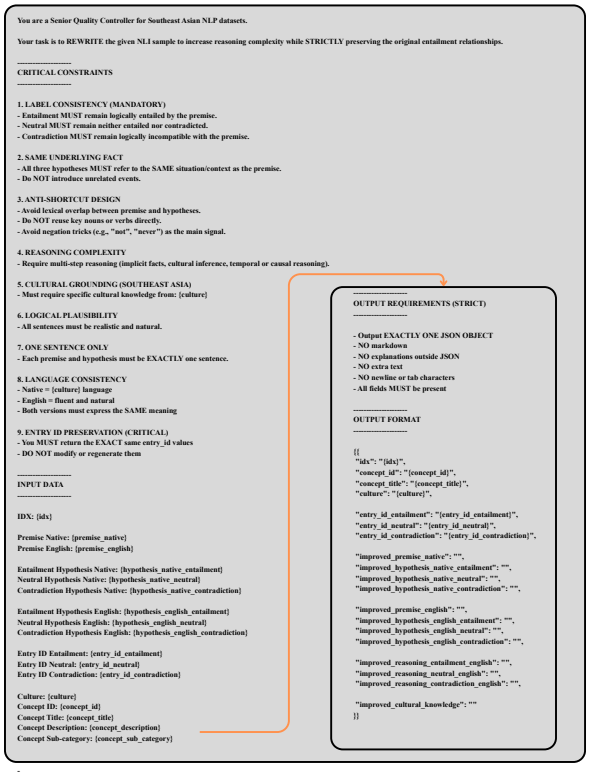
Quantitatively, this is supported by a reduction in lexical shortcuts: the loop balances word overlap across entailment types (Figure 15) and eliminates class-specific "trigger words" that models often exploit as heuristics (see Figure 11c).

D Inference Speed vs. Performance Analysis

The relationship between model performance on the SEA-NLI dataset and inference throughput is illustrated in Figure 16. The graph illustrates the efficiency-performance trade-off across different architectures on the SEA-NLI benchmark. While encoder-only models (green) offer superior throughput, they lack the cultural understanding



Figure 11: Evolution of the SEA-NLI word frequency distribution. All values are normalized to the percentage of total hypotheses per label. The sharp spikes in the pilot dataset (a) were addressed through prompt engineering (b) and automated refinement and filtering (c), resulting in a balanced and robust distribution of concrete semantic terms across all entailment categories.



```
You are a Senior Quality Controller for Southeast Asian NLP datasets.

Your task is to REWRITE the given NLI sample to increase reasoning complexity while STRICTLY preserving the original entailment relationships.

--------------------
CRITICAL CONSTRAINTS
--------------------

1. LABEL CONSISTENCY (MANDATORY)
- Entailment MUST remain logically entailed by the premise.
- Neutral MUST remain neither entailed nor contradicted.
- Contradiction MUST remain logically incompatible with the premise.

2. SAME UNDERLYING FACT
- All three hypotheses MUST refer to the SAME situation/context as the premise.
- Do NOT introduce unrelated events.

3. ANTI-SHORTCUT DESIGN
- Avoid lexical overlap between premise and hypotheses.
- Do NOT reuse key nouns or verbs directly.
- Avoid negation tricks (e.g., "not", "never") as the main signal.

4. REASONING COMPLEXITY
- Require multi-step reasoning (implicit facts, cultural inference, temporal or causal reasoning).

5. CULTURAL GROUNDING (SOUTHEAST ASIA)
- Must require specific cultural knowledge from: {culture}

6. LOGICAL PLAUSIBILITY
- All sentences must be realistic and natural.

7. ONE SENTENCE ONLY
- Each premise and hypothesis must be EXACTLY one sentence.

8. LANGUAGE CONSISTENCY
- Native = {culture} language
- English = fluent and natural
- Both versions must express the SAME meaning

9. ENTRY ID PRESERVATION (CRITICAL)
- You MUST return the EXACT same entry_id values
- DO NOT modify or regenerate them

--------------
INPUT DATA
--------------

IDX: {idx}

Premise Native: {premise_native}
Premise English: {premise_english}

Entailment Hypothesis Native: {hypothesis_native_entailment}
Neutral Hypothesis Native: {hypothesis_native_neutral}
Contradiction Hypothesis Native: {hypothesis_native_contradiction}

Entailment Hypothesis English: {hypothesis_english_entailment}
Neutral Hypothesis English: {hypothesis_english_neutral}
Contradiction Hypothesis English: {hypothesis_english_contradiction}

Entry ID Entailment: {entry_id_entailment}
Entry ID Neutral: {entry_id_neutral}
Entry ID Contradiction: {entry_id_contradiction}

Culture: {culture}
Concept ID: {concept_id}
Concept Title: {concept_title}
Concept Description: {concept_description}
Concept Sub-category: {concept_sub_category}

--------------------
OUTPUT REQUIREMENTS (STRICT)
--------------------

- Output EXACTLY ONE JSON OBJECT
- NO markdown
- NO explanations outside JSON
- NO extra text
- NO newline or tab characters
- All fields MUST be present

--------------------
OUTPUT FORMAT
--------------------

{{
"idx": "{idx}",
"concept_id": "{concept_id}",
"concept_title": "{concept_title}",
"culture": "{culture}",

"entry_id_entailment": "{entry_id_entailment}",
"entry_id_neutral": "{entry_id_neutral}",
"entry_id_contradiction": "{entry_id_contradiction}",

"improved_premise_native": "",
"improved_hypothesis_native_entailment": "",
"improved_hypothesis_native_neutral": "",
"improved_hypothesis_native_contradiction": "",

"improved_premise_english": "",
"improved_hypothesis_english_entailment": "",
"improved_hypothesis_english_neutral": "",
"improved_hypothesis_english_contradiction": "",

"improved_reasoning_entailment_english": "",
"improved_reasoning_neutral_english": "",
"improved_reasoning_contradiction_english": "",

"improved_cultural_knowledge": ""
}}
```

Figure 12: The regeneration prompt for improving NLI sample quality in each refinement loop.

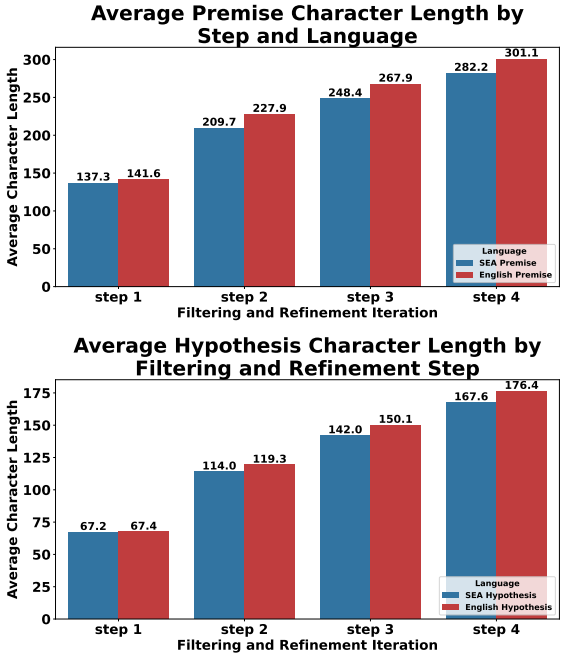


Figure 13: Average character length of premises (top) and hypotheses (bottom) across LLM filtering and refinement step (Steps 1-4) for both SEA and English language samples.

found in larger LLMs. A key finding is the "cultural collapse" of general models like Llama-3.1-8B-Instruct, which performs well in English but fails in the SEA context. Conversely, SEA-specific LLMs (purple) demonstrate high cultural robustness, maintaining top-tier F1 scores in regional contexts despite lower inference speeds. This underscores that for regional tasks, localized training is more critical for robustness than sheer inference speed or general-purpose scaling.

Step1 - No improvement

English
Premise: After exiting Downtown MRT, I started drawing The Sail at Marina Bay for my school art project because the tower looked tall.
hypothesis: After leaving the MRT, I drew a tall building.

native
Premise:டவுன்டவுன் MRT-ல இருந்து வெளிய வந்ததும் மெரினா பேயில் இருக்கும் தி சேல் உயரமா தெரிந்ததால், பள்ளி கலைப்பணிக்காக நான் அந்தக் கட்டிடத்தை வரைய ஆரம்பிச்சேன்.
hypothesis:MRT-ல இருந்து வெளிய வந்ததும் நான் ஒரு உயரமான கட்டிடத்தை வரைந்தேன்.

Step2 - First improvement

English
Premise: After alighting at Downtown MRT and walking towards the Marina Bay promenade, the gleaming Sail on the skyline caught my eye, so I began a sketch of it for my school art assignment.
hypothesis: After coming out of the station, I sketched a tall skyline building as part of a school assignment.

native
Premise:டவுன்டவுன் MRT-யில் வெளியேறி மெரினா பே ப்ரொமனேட்க்கு நடக்கும்போது skyline-ல பளிச்சென்று தெரிந்த The Sail-ஐ பார்த்ததும், பள்ளி art assignment-க்காக நான் அதோட ஸ்கெட்ச் போட ஆரம்பிச்சேன்.
hypothesis:ரெயில் நிலையத்திலிருந்து வெளிய வந்த பிறகு, பள்ளி பணிக்காக நான் skyline-ல இருக்கிற ஒரு உயரமான கட்டிடத்தின் ஸ்கெட்ச் போட்டேன்.

Step3 - Second improvement

English
Premise: After getting off at Downtown station on the Downtown Line and heading towards the Marina Bay waterfront walkway, I spotted the glittering Sail condo on the CBD-facing skyline and immediately started drafting its outline in my notebook as reference for my school art coursework.
hypothesis: Once I was out of the train station, I began a quick drawing of a tall residential tower seen on the city skyline to fulfil a school visual-arts task.

native
Premise:Downtown line-ல Downtown station-ல இறங்கி Marina Bay waterfront boardwalk பக்கம் நடந்தப்போ CBD side-ல glitter ஆக standing இருந்த The Sail condo என் கண்ணுக்கு பட்டதால, பள்ளி art coursework-க்கு reference ஆக உடனே notebook-ல அதன் outline sketch போட ஆரம்பிச்சேன்
hypothesis:ரெயில் நிலையத்திலிருந்து வெளிய வந்த பிறகு, பள்ளிக்கான visual-arts coursework முடிக்க CBD skyline-ல தெரிந்த ஒரு உயரமான residential tower-ஐ நான் quick-ஆ draw பண்ண ஆரம்பிச்சேன்.

Step4 - Third improvement

English
Premise: After alighting at Downtown MRT on the Downtown Line and heading toward the Marina Bay promenade, I noticed The Sail @ Marina Bay sparkling on the CBD-facing skyline over the reclaimed area, and I immediately began capturing its silhouette in my notebook as a reference for my school art coursework.
hypothesis: Once I got out from Downtown MRT on the DTL, I promptly sketched a rough outline of a private condominium tower visible along the Marina Bay skyline to support a school art assignment.

native
Premise:Downtown MRT (DTL)-ல வெளியேறி Marina Bay promenade பக்கம் போனப்போ, reclaimed land பக்கம் CBD-க்கு எதிரா மின்னி தெரியுற The Sail @ Marina Bay-ஐ பார்த்ததும், O-Level/A-Level art coursework reference-க்காக notebook-ல அதன் silhouette-ஐ அப்போதே கையால பிடிச்சு வைக்க தொடங்கினேன்.
hypothesis:DTL-ல Downtown MRT-லிருந்து வெளிய வந்ததும், Marina Bay பக்கத்து skyline-ல தெரிந்த ஒரு private condo tower-ஐ பள்ளி art assignment-க்கு support ஆக உடனே என் notes புத்தகத்துல rough outline-ஆ போட்டேன்.

Figure 14: Example of SEA-NLI in each regeneration step.

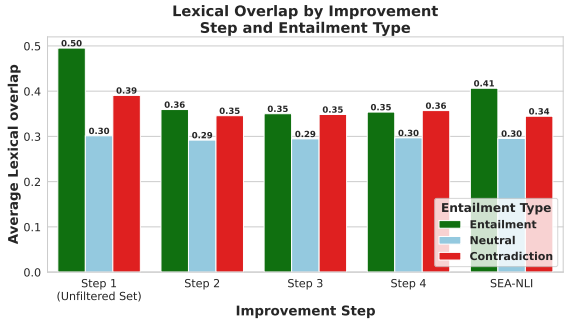


Figure 15: Lexical overlap between premise and hypothesis in each regeneration step.

E Annotator Agreement and Details

To ensure linguistic and cultural authenticity, we established a rigorous quality control framework involving human experts. This section outlines the selection criteria for Southeast Asian language specialists, the metrics employed to evaluate LLM outputs, and the methodologies for measuring inter-annotator agreement. By combining expert oversight with standardized scoring for cultural nuance and grammatical precision, we measured inter-annotator agreement specifically for the NLI labels to confirm logical consistency. Furthermore, we assessed cultural relevance, cultural understanding, and grammatical precision through expert-assigned quality scores.

E.1 Annotator Selection

To select annotators for the study, we recruit native speakers and language experts from eight Southeast Asian countries, where we give two tasks to annotators. The first task is entailment type classification, which includes twelve multiple-choice questions where annotators identify if a relationship is an Entailment, Neutral, or Contradiction (12 examples). The second task is hypothesis generation, which includes three questions where annotators are given a premise and must write their own entailment, neutral, and contradiction hypotheses (3 examples). We select only the annotator that pass the tasks with a score of more than 80%. This process confirms that every annotator fully understands the logic required for the project.

E.2 Annotator for Cultural Filtering

For the cultural filtering process in Section 3.1, we asked three annotators for each country to filter unrelated cultures. In particular, we asked annotators to check whether they were familiar with the concept title from scraped Wikipedia or not, where we removed only the topics that all three annotators were unfamiliar with. We also found that the rejection rate is only $\sim$5% from around $\sim$24,000 culture topics. The pattern we found is that the culture topics that do not have much information on Wikipedia pages would be removed.

Then, we filter the topic that is the least popular in Wikipedia using n_link, where the number indicates the number of citations in Wikipedia. We removed the page that has the least 70% citation in Wikipedia, where the number was also calculated by our annotators, increasing or decreasing the number might include the least popular or not well-known topics. Using these topics might be incorrect in the label verification step since most annotators might not fully understand them.

E.3 Human Evaluation Metrics

To evaluate the quality of LLM-generated NLI samples in Section 3.4, we utilize the following metrics:

1. **Cultural Relevance Score:** Adapted from Cahyawijaya et al. (2025), this metric assesses how effectively the generated content aligns with the intended Southeast Asian (SEA) cultural context.
    - **Score 5 (Unique to SEA):** The premise describes traditions, objects, or landmarks

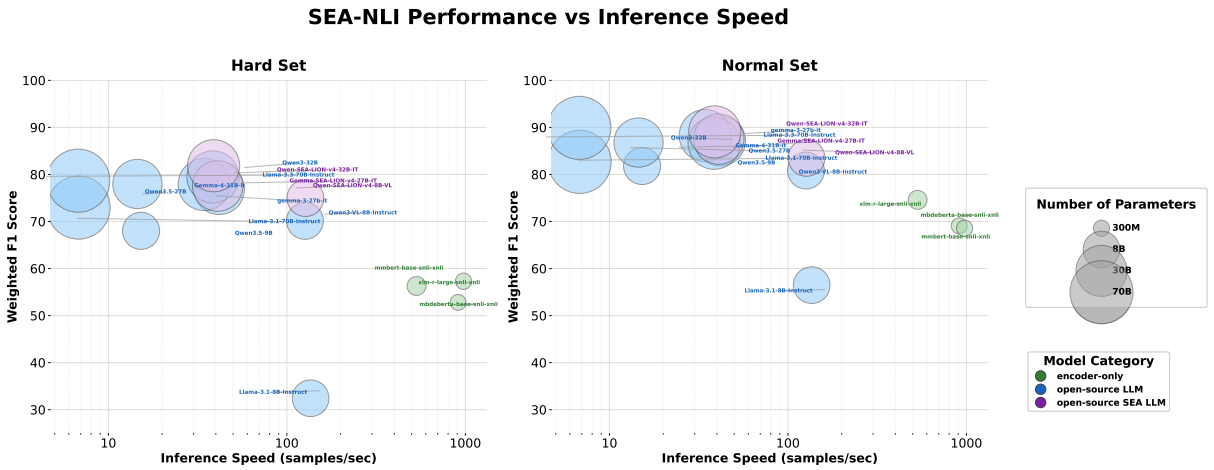


Figure 16: Inference speed vs. F1 score on the SEA-NLI benchmark. Results are shown for SEA performance (left) and English performance (right). Bubble sizes scale with parameter count, illustrating the trade-off between model scale, speed, and F1-score.

that originate in SEA and are considered iconic, such as Pad Thai, Batik, Songkran, or the Petronas Towers.

- **Score 3 (Ubiquitous):** The content features concepts that are not exclusive to SEA but are common in daily life or possess regional nuances, such as the Chinese New Year in Singapore, local mosques, or specific tropical fruit flavors.
- **Score 1 (Unrelated):** The content involves concepts unrelated to SEA, such as the Super Bowl, the Statue of Liberty, or generic international brands like IKEA.

Intermediate scores (2 or 4) should be assigned if the context falls between these primary categories.

2. **Cultural Understanding Score:** This score quantifies the annotator's personal familiarity with the specific cultural context of the sample.
   - **Score 5 (Native / Expert):** The annotator identifies with the culture and understands specific nuances, local slang, and social traditions instinctively.
   - **Score 4 (High Familiarity):** The annotator has lived in the region or possesses strong personal experience and comfort with the context.
   - **Score 3 (Moderate Awareness):** The annotator recognizes the topic and its general meaning but lacks deep personal experience or specific detail.
   - **Score 2 (Limited Exposure):** The annotator has heard of the tradition but has minimal knowledge of its actual practices.
   - **Score 1 (No Prior Knowledge):** The culture is foreign to the annotator, who must rely entirely on the provided metadata to complete the task.
3. **Quality Score:** This metric evaluates the linguistic clarity and contextual accuracy of the premise and its associated metadata.
   - **Score 5 (Excellent):** The sentence is natural, grammatically perfect, and provides a clear cultural context.
   - **Score 4 (Good):** The sentence is clear and usable, with only minor stylistic awkwardness.
   - **Score 3 (Fair):** The sentence is understandable, but the phrasing is unnatural or the context is slightly vague.
   - **Score 2 (Poor):** Significant grammatical issues or confusing context make logical deduction difficult.
   - **Score 1 (Broken):** The sample is nonsensical, contains major factual errors, or is unusable.
4. **Flagging Issues:** Annotators identify specific qualitative concerns that may impact the reliability of the sample. These include:
   - **Linguistic Error:** Significant grammar, spelling, or translation issues.
   - **Factual Inaccuracy:** The premise contains incorrect information regarding the culture or location.
   - **Ambiguous Context:** The statement is too vague to determine a definitive logical relationship.
   - **Culturally Offensive:** The content is

disrespectful or promotes harmful stereotypes.

- **Poor Formatting:** Issues with character encoding (e.g., broken local scripts) or punctuation.
- **Irrelevant to Topic:** The sample doesn't actually relate to the specific cultural topic or metadata provided.
- **Translation Mismatch:** The English translation and the native language sentence do not convey the same meaning.

E.4 Annotator Metrics

E.4.1 Annotator Agreement

Inter-annotator agreement was assessed using the Intraclass Correlation Coefficient (ICC) to validate the reliability and consistency of the human judgments (Shrout and Fleiss, 1979). As shown in Table 4, we calculated the ICC for each cultural category individually to account for regional variances in perception. To derive a unified reliability metric across the dataset, these individual coefficients were aggregated using Fisher's Z-transformation (Fisher, 1921). This transformation normalizes the distribution of the correlation coefficients, allowing for a statistically valid averaging process that accurately reflects the overall stability of the annotations across the diverse Southeast Asian languages studied.

Culture	Pilot Set	Non-filter Set	SEA-NLI
Cambodian	0.78	0.99	0.93
Filipino	0.99	0.98	N/A
Indonesian	0.98	0.99	0.97
Malaysian	0.96	1.00	0.90
Myanmar	0.75	1.00	0.82
Singaporean	N/A	N/A	0.93
Thai	0.96	1.00	0.99
Vietnamese	0.94	0.99	0.98
Fisher's Z ICC	**0.95**	**0.99**	**0.96**

Table 4: Inter-annotator agreement (ICC(C,k)) for Entailment Type Classification across three refinement rounds.

E.4.2 Cultural Relevant

Annotators assessed regional authenticity using the Cultural Relevance Score (CRS). As shown in Table 5, SEA-NLI achieved a high weighted average of 4.493, indicating that the samples are strongly grounded in Southeast Asian regional contexts.

E.4.3 Naturalness

The linguistic naturalness of the SEA-NLI dataset was evaluated by human annotators using the 'Quality Score' metric (Section E.3), which measures grammatical correctness, cultural authenticity, and stylistic fluency. As shown in Table 6, from these sets, the overall weighted average improved from 4.151 to 4.651. While scores increased across the board, the relative rankings of individual cultures shifted significantly in each round, reflecting varying rates of progression. These results indicate a high level of quality and strong linguistic reliability, with final scores across all regional subsets approaching a perfect rating.

	Pilot Set	Unfiltered Set	SEA-NLI	
Culture	($n = 120$)	($n = 300$)	n	**CRS**
Cambodian	4.100	4.445	65	4.277
Filipino	4.267	4.463	331	4.979
Indonesian	4.505	4.458	390	4.599
Malaysian	3.822	4.608	195	4.192
Myanmar	4.558	4.825	185	4.881
Singaporean	3.875	3.017	424	3.692
Thai	4.475	4.596	349	4.704
Vietnamese	3.602	4.498	221	4.607
Weighted Avg.	**4.151**	**4.364**	**2160**	**4.493**

Table 5: Cultural Relevant Score Progression Across Evaluation Rounds

	Pilot Set	Unfiltered Set	SEA-NLI	
Culture	($n = 120$)	($n = 300$)	n	**Quality**
Cambodian	4.079	4.477	65	3.623
Filipino	4.719	4.982	331	4.983
Indonesian	4.757	4.796	390	4.824
Malaysian	4.200	4.661	195	4.403
Myanmar	4.119	4.680	185	4.832
Singaporean	4.142	4.117	424	4.272
Thai	4.467	4.582	349	4.772
Vietnamese	4.373	4.651	221	4.756
Weighted Avg.	**4.357**	**4.618**	**2160**	**4.651**

Table 6: Quality Score Progression Across Evaluation Rounds

E.4.4 Flags During Quality Control

Samples flagged by annotators were adjudicated via majority vote to confirm the presence of anomalies. The proportions of anomalous samples are 8.75%, 4.12%, and 1.48%, respectively. The distribution of raw flags prior to the majority voting process of SEA-NLI is presented in Figure 17. The most frequent flag in this dataset is the linguistic error, which exhibits the highest overall density across the majority of cultures, followed by ambiguity.

E.5 Annotator Details

The annotators are native speakers of the target Southeast Asian languages and originate from eight Southeast Asian countries. The number of annotators assigned to each cultural group is detailed in

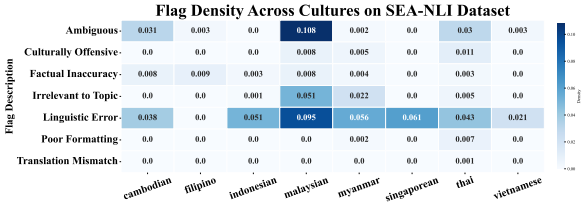


Figure 17: Distribution of data quality flags by culture on the SEA-NLI dataset. Higher values indicate a greater prevalence of specific issues, such as linguistic errors or ambiguity, within a given cultural subset.

Table 7.

Language	Country of Origin	Count		
		Pilot Set	Unfiltered Set	SEA-NLI
Indonesian	Indonesia	5	6	5
Khmer	Cambodia	2	2	2
Malay	Malaysia	3	2	2
Myanmar	Myanmar	3	2	3
Filipino	Philippines	3	2	1
Tamil	Singapore	1	1	2
Thai	Thailand	2	8	8
Vietnamese	Vietnam	4	4	4
Total		**23**	**27**	**27**

Table 7: Distribution of native-speaking annotators by language and country of origin.

F Error Analysis Full Results

The full error analysis results from all models are shown in Table 8.

Model	Both ✓	En ✓ / SEA ×	En × / SEA ✓	Both ×
Encoder Models				
mmBERT-base	51.76	11.99	12.59	23.66
mDeBERTa-v3-base	48.89	14.17	11.39	25.56
XLM-RoBERTa-large	54.17	7.18	13.56	25.09
Decoder Models				
Llama-3.1-8B-Instruct	44.72	30.56	6.16	18.56
Qwen3-VL-8B-Instruct	68.56	9.31	8.24	13.89
Qwen-3.5-9B	69.95	9.95	6.85	13.24
Llama-3.1-70B-Instruct	76.94	9.95	3.24	9.86
Qwen3-32B	80.09	6.06	4.49	9.35
Qwen3.5-27B	77.50	5.28	6.06	11.16
Gemma-3-27B-IT	79.44	4.35	4.77	11.44
gemma-4-31b-IT	78.95	5.56	3.89	11.62
Llama-3.3-70B-Instruct	81.90	3.38	6.30	8.43
Qwen-SEA-LION-v4-8B-VL	73.94	8.33	6.30	11.44
Gemma-SEA-LION-v4-27B-IT	79.54	4.26	5.09	11.11
Qwen-SEA-LION-v4-32B-IT	81.71	4.21	4.91	9.17
GPT-5.4	80.32	4.63	5.69	9.35
Gemini-3-flash-preview	81.25	3.15	6.25	9.35

Table 8: Model performance grouped by architecture (%). Both ✓ (Accuracy), ENG ✓/SEA × (Language Gap), ENG ×/SEA ✓ (Alignment Error), Both × (Task Failure).

G Category Results on English Set

The performance by category for the English subset of SEA-NLI is presented in Figure 18. Consistent with the results for SEA languages, all models continue to struggle with the "Language" and "Science and Technology" topics.

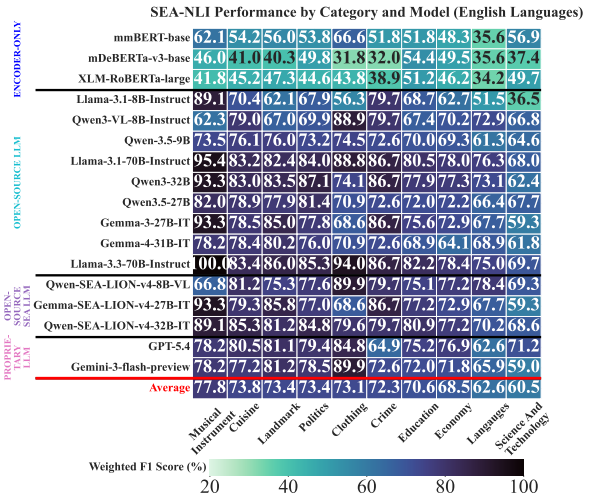


Figure 18: English Weighted F1 Heatmap for the Hard Set, showing per-category performance for each evaluated model.

H Prompts for Generating Premises and Hypotheses

Building upon the pilot dataset generation phase described in Appendix B, we identified several limitations that required correction. To address these, we designed a revised hypothesis generation prompt by adding new rules to enforce linguistic neutrality and precision. Specifically, we prohibited the use of hedging or speculative language in 'Neutral' hypotheses, as we found that words such as 'might,' 'may,' 'perhaps,' and 'likely' created spurious correlations. The revised strategy requires the model to describe specific actions or states that lack evidence in the premise, rather than using vague wording to signal neutrality (see Figure 15b). This ensures that the 'Neutral' status arises strictly from the logical relationship between the premise and hypothesis. The prompts used to generate the premises and hypotheses are shown in Figures 9 (same as premise generation prompt for pilot dataset) and 19, respectively.

I Cultural-aware Prompting

Building on the base prompt shown in Figure 20a, we employ culturally aware prompting to elicit the model's latent SEA knowledge. This methodology integrates (i) target-culture metadata (`Base+Cult.`), (ii) target-culture-topic metadata (`Base+Cult.+Topic.`), (iii) target-culture-concept-summary metadata (`Base+Cult.+Sum.`), (iv) target-culture metadata with CoT (`CoT+Cult.`), (v) target-culture-concept-summary metadata with CoT (`CoT+Cult.+Sum.`), and (vi) target-culture metadata with CoT and in-context learning



Figure 19: The hypothesis generation prompt.

(CoT+ICL+Cult.), as illustrated in Figures 20b to 20g, respectively. We hypothesize that such targeted prompting is essential to activate the model's internal cultural representations for accurate reasoning.

J Lower Performance in CoT Prompt

We evaluated the SEA-NLI dataset in various prompt types in subsection 6.2. Notably, the reasoning prompt (CoT+Cult.) underperformed compared to the answer-only prompt (Base+Cult.). To investigate this, we conducted an error analysis on the reasoning results. As shown in Table 9, the primary error mode of both prompts involves the model predicting a "Neutral" label instead of "Entailment" or "Contradiction."

For qualitative analysis, we observed cases of over-constrained neutral prediction, where the reasoning model failed to recognize entailment/contradiction due to subtle paraphrasing, as shown in Figure 21. In these instances, the model's reasoning chain tends to over-analyze minor lexical or syntactic variations, incorrectly interpreting them as evidence of missing information. Rather

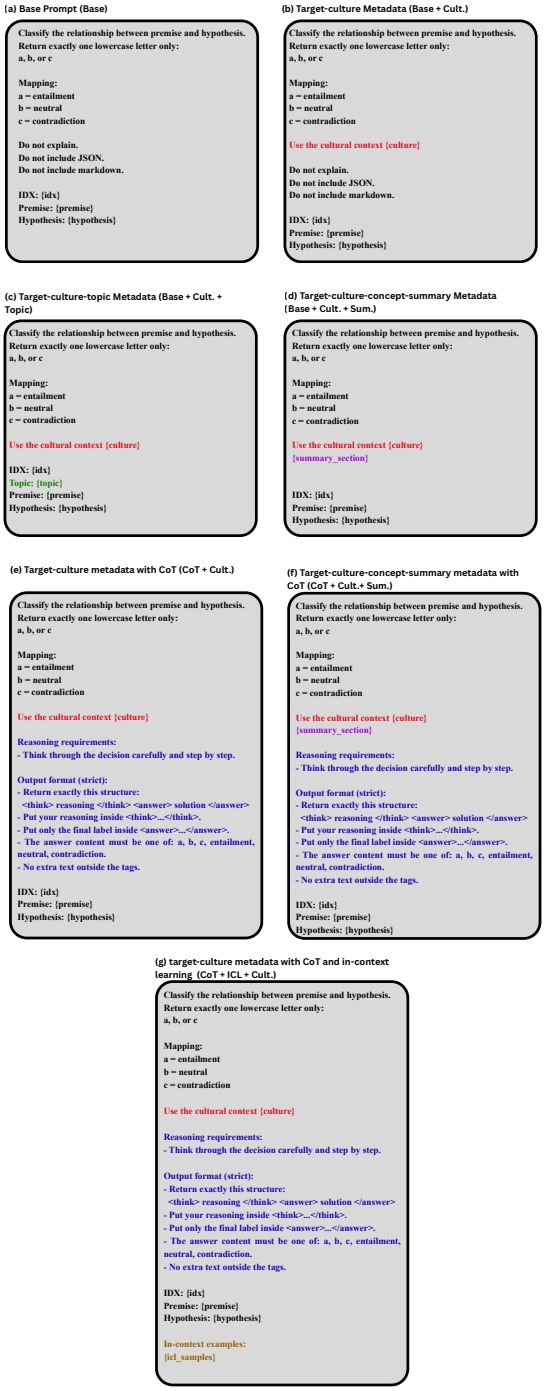


Figure 20: Evolution of prompt templates for culturally-aware reasoning.

than recognizing semantic equivalence, the model adopts an overly literal stance, concluding that the hypothesis cannot be strictly inferred from the premise because the exact wording differs. This suggests that while explicit reasoning can improve transparency, it may also introduce a "hallucination of divergence," where the model perceives a lack of logical connection due to a focus on surface-level differences rather than underlying semantic meaning. From the given example in Figure 21. ,

Model	Lang.	Type	Gold	Pred.	Freq.
Gemma-SEA-LION-v4-27B-IT	english	Base+Cult.	Ent.	Neut.	23
			Ent.	Cont.	4
			Cont.	Neut.	145
			Cont.	Ent.	7
			Neut.	Ent.	51
			Neut.	Cont.	36
		CoT+Cult.	Ent.	Neut.	44
			Ent.	Cont.	4
			Cont.	Neut.	214
			Cont.	Ent.	8
			Neut.	Ent.	59
			Neut.	Cont.	27
	native	Base+Cult.	Ent.	Neut.	21
			Ent.	Cont.	4
			Cont.	Neut.	146
			Cont.	Ent.	12
			Neut.	Ent.	61
			Neut.	Cont.	39
		CoT+Cult.	Ent.	Neut.	37
			Ent.	Cont.	3
			Cont.	Neut.	229
			Cont.	Ent.	12
			Neut.	Ent.	54
			Neut.	Cont.	20

Table 9: Comprehensive Error Analysis for Gemma-SEA-LION-v4-27B-IT: Updated frequency of misclassification types grouped by Language across Base and CoT prompts.

the model fails to account for Thai cultural norms: since wearing Scout or Girl Guide uniforms to a Sunday ordination ceremony is highly unusual and contextually improbable, the relationship represents a situational contradiction that the model dismisses as a neutral, unrelated event."

Premise: เช้าวันพฤหัสลูกชายต้องใส่ชุดลูกเสือคอผ้าพันคอ ส่วนลูกสาวใส่ชุดเนตรนารีสีเขียวเข้ม (translation: On Thursday mornings my son wears the khaki scout uniform with a neckerchief, while my daughter wears the dark green guide uniform.)

Hypothesis: ชุดนี้ใส่ไปงานบวชวันอาทิตย์ที่วัด และไม่ได้เกี่ยวกับตารางกิจกรรมของโรงเรียน
(translation: Those uniforms are worn for Sunday temple ordination events, unrelated to school activity schedules.
)

Label: Contradiction

Reasoning Model Thinking: The premise states that the son must wear a scout uniform with a neckerchief and the daughter must wear a dark green Girl Guide uniform on Thursday mornings. The hypothesis states that this outfit is worn to a Buddhist ordination ceremony on Sunday at a temple and is not related to the school schedule. The premise describes a school uniform requirement on a specific day of the week (Thursday). The hypothesis describes the same uniform being worn at a different event (ordination ceremony on Sunday) and explicitly states it's *not* related to school. This means the hypothesis provides information *outside* of what the premise states, and doesn't contradict it. It's possible to wear the uniform outside of school hours for other events. Therefore, the relationship is neutral.

Model Prediction: Neutral

Figure 21: Error analysis of a cultural knowledge gap. The model fails to recognize the entailment relationship because it lacks the specific grounding to link school-mandated scout uniforms to their broader social and religious usage in the Southeast Asian context.

K Bias Analysis

To understand the annotation bias in NLI that might be generated during the dataset generation, we conduct three studies to demonstrate that SEA-NLI mitigates these problems. Firstly, we conduct a word overlap guessing using Jaccard similarity as the indicator for NLI prediction. Secondly, we conduct a hypothesis length prediction since we found that previous NLI datasets can guess the class by using the length for making predictions, where the entailment class has the highest length, and the neutral is the second-highest word length. Lastly, we conduct a hypothesis-only prediction to study the possible of a model guessing a correct answer using only hypotheses.

K.1 Word Overlap

The lexical overlap for each entailment type is illustrated in Figure 15. The results indicate that the sample filtering method proposed in section 3.3 successfully reduces the discrepancy in lexical overlap across all classes. Despite this reduction, the entailment class continues to exhibit the highest word overlap (0.41 after filtering), followed by contradiction (0.34) and neutral (0.30). To assess whether models could exploit these lexical cues as a shortcut, we conducted a heuristic-based experiment. This baseline predicts the label based on overlap thresholds, assuming the highest overlap correlates with entailment (H:E), the middle with contradiction (M:C), and the lowest with neutral (L:N). The evaluation results across different filtering processes are presented in Table 10. For the unfiltered set, the heuristic achieves a weighted F1-score of 56.95%. However, after applying the data filtering and refinement process, the heuristic's performance drops to 54.09% on the SEA-NLI (normal) set, and drops to 39.66% on the SEA-NLI (hard) set. This shows that our filtering and refinement process can mitigate word overlap artifacts to a certain extent.

Dataset	Weighted Avg. F1 (ENG)	Prediction Rule
Unfiltered Set	56.95%	Word Overlap (H:E, M:C, L:N)
SEA-NLI (normal)	54.09%	Word Overlap (H:E, M:C, L:N)
SEA-NLI (hard)	39.66%	Word Overlap (H:E, M:C, L:N)

Table 10: Impact of filtering on word-overlap prediction performance. The notation (H:E, M:C, L:N) indicates that highest, medium, and lowest overlap levels are mapped to Entailment, Contradiction, and Neutral, respectively.

K.2 Hypothesis Length

We also investigated potential statistical biases related to hypothesis length. As illustrated in Figure 22, entailment hypotheses in our dataset tend to be longer, while neutral hypotheses are generally shorter. To test the robustness of the dataset against length-based shortcuts, we implemented a heuristic baseline that predicts labels based on length rankings: the longest hypotheses were classified as entailment (H:E), followed by contradiction (medium, M:C), and neutral (shortest, L:N). Table 11 presents the evaluation results for the hypothesis length heuristic. This heuristic achieves a weighted F1-score of 39.02% on the SEA-NLI normal set, a notable reduction from the unfiltered baseline of 51.02%. In contrast, the hard set exhibits a higher F1-score of 54.04%, suggesting an amplified length bias. This increased reliance on length in the hard set can be attributed to a label imbalance skewed toward the Entailment class. However, the overall performance of the hard set still challenging. These results indicate that hypothesis length alone is an insufficient predictor of the entailment label, suggesting that SEA-NLI requires deeper semantic understanding.

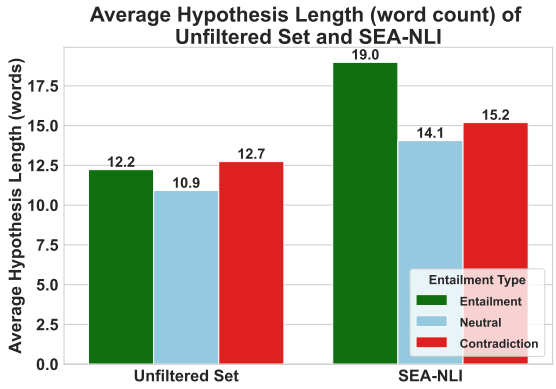


Figure 22: Average hypothesis length before and after the filtering step, categorized by entailment label.

K.3 Hypothesis-only Prediction

To investigate the presence of potential annotation artifacts in SEA-NLI, we conducted a hypothesis-only experiment. If a model can easily predict the label without the premise, it suggests the hypothesis contains linguistic cues characteristic of specific classes (e.g., the word 'not' appearing frequently in contradictions). The results are presented in Table 12. From the results, the F1-score drops significantly for both the hard and normal sets, decreasing by approximately 64 and 49 points, respectively. Crucially, the hypothesis-only performance (ranging between 21% and 29%) is below the random baseline (33.3%) for a three-class problem, indicating that the SEA-NLI dataset is robust against simple annotation artifacts and requires the model to reason over the premise-hypothesis pair.

Dataset	Weighted Avg. F1 (ENG)	Prediction Rule
Unfiltered Set	51.02%	hypothesis length (H:E, M:C, L:N)
SEA-NLI (normal)	39.02%	hypothesis length (H:E, M:C, L:N)
SEA-NLI (hard)	54.04%	hypothesis length (H:E, M:C, L:N)

Table 11: Evaluation results of hypothesis length heuristics across dataset splits. The notation (H:E, M:C, L:N) indicates that highest, medium, and lowest hypothesis length levels are mapped to Entailment, Contradiction, and Neutral, respectively.

Model	Setting	Only Hypo.	SEA F1	ENG F1
Gemma-SEA-LION-v4-27B-IT	Normal	False	87.51	86.69
		True	23.66	21.04
	Hard	False	77.53	76.90
		True	28.96	24.46
Gemma-3-27B-IT	Normal	False	87.16	86.81
		True	23.59	20.89
	Hard	False	76.85	76.53
		True	27.91	23.26

Table 12: Comparison between using hypothesis-premise pair and using only hypothesis to predict entailment type on SEA-NLI dataset.

L Free Form Prompt vs Enforce Prompt

Due to the fact that we format the prompt as a multiple-choice question-answering format, e.g., asking LLMs to answer with A (Entailment), B (Neutral), and C (Contradiction), this might limit the performance of LLMs for thinking or reasoning skills. Thus, we evaluated the reasoning capabilities of the LLM by comparing a free-form prompt (Figure 23) with an enforced reasoning prompt (Figure 20b). The results show only marginal performance differences across both the English and native language subsets. These findings indicate that performance limitations stem from deficits in cultural knowledge rather than constraints imposed by the prompt format.

Free Form Generation

Classify the relationship between premise and hypothesis. The format is [ANSWER]. You need to replace ANSWER with "a" ,"b", or "c"

Mapping:
a = entailment
b = neutral
c = contradiction

Use the cultural context {culture}

IDX: {idx}
Premise: {premise}
Hypothesis: {hypothesis}

Figure 23: Average hypothesis length before and after the filtering step, categorized by entailment label.

Model: Qwen-SEA-LION-v4-32B-IT			
Category	**Prompt Type**	**F1 native**	**F1 english**
Normal	free-form	85.03	87.78
	enforce	86.14	87.94
Hard	free-form	78.45	76.68
	enforce	78.34	79.97

Table 13: Model Performance Comparison

M Full Results of SEA-NLI

The performance of entailment classification on the SEA-NLI dataset, segmented by cultural context, is presented in Table 15. We also report human baseline performance in Appendix O.

N Cross-model Variation Analysis

To assess variation in performance across models, we analyze the results in Table 15. First, we calculate the mean score across models for each culture, difficulty level, and language condition:

$$\bar{x}_c^{(d,\ell)} = \frac{1}{N} \sum_{m=1}^{N} x_{m,c}^{(d,\ell)}. \quad (1)$$

Here, N denotes the number of models, and $x_{m,c}^{(d,\ell)}$ denotes the score of model m for culture c, difficulty level $d \in \{\text{Normal}, \text{Hard}\}$, and language condition $\ell \in \{\text{SEA}, \text{English}\}$. The quantity $\bar{x}_c^{(d,\ell)}$ is the mean score across all N models for a fixed culture, difficulty level, and language condition. Next, we calculate the sample variance to quantify the dispersion of model scores around the mean:

$$\left(s_c^{(d,\ell)}\right)^2 = \frac{1}{N-1} \sum_{m=1}^{N} \left(x_{m,c}^{(d,\ell)} - \bar{x}_c^{(d,\ell)}\right)^2. \quad (2)$$

Here, $\left(s_c^{(d,\ell)}\right)^2$ denotes the sample variance across the N models for a fixed culture, difficulty level, and language condition. A larger variance indicates greater performance differences across models, whereas a smaller variance reflects more consistent scores. As shown in Table 14, smaller subsets like Khmer (KHM) demonstrate stability comparable to larger subsets (e.g., KHM hard set variance is 153.09, compared to 294.35 for IDN in the hard SEA version). If data quality were compromised, we would instead observe erratic, spiking variances.

	Normal		**Hard**	
Culture	**SEA**	**English**	**SEA**	**English**
IDN	79.54	31.16	294.35	205.98
VIE	77.43	101.50	243.95	414.25
THA	65.10	54.25	219.34	250.74
ZSM	87.96	36.70	186.36	89.43
TAM	153.76	58.86	91.37	107.35
MYA	183.39	56.84	276.25	137.46
FIL	112.25	68.41	214.32	89.86
KHM	150.65	51.65	153.09	100.96

Table 14: Sample variance across models by culture, difficulty level, and language condition.

O Human baselines

To calculate human baseline, we first establish the ground-truth label using a weighted vote based on annotators' Cultural Understanding Scores (CUS). For each label, we sum up the CUS of the annotators who voted for it. The option with the highest cumulative score becomes the ground-truth label. Example: If a sample is reviewed by 5 annotators:

- Annotator 1: Neutral (CUS: 5).
- Annotator 2: Neutral (CUS: 4).
- Annotator 3: Neutral (CUS: 5).
- Annotator 4: Neutral (CUS: 5).
- Annotator 5: Entailment (CUS: 4).

Result: The total score for Entailment is 4, Neutral is 19 ($5+4+5+5$), and Contradiction is 0. Therefore, the final ground-truth label is Neutral. We then compute each annotator's accuracy with respect to the final CUS-weighted ground-truth labels. The accuracies are aggregated using the number of

Model	IDN	VIE	THA	ZSM	TAM	MYA	FIL	KHM	Avg
Normal Set									
mDeBERTa-v3-base	77.36/75.96	69.09/60.37	70.13/65.70	69.75/66.31	65.48/70.25	69.63/64.93	64.21/73.83	58.90/70.73	69.08/69.32
mmBERT-base	74.68/78.73	72.97/67.72	65.00/68.87	68.95/71.55	66.61/71.66	63.03/64.42	67.15/64.53	69.19/73.37	68.56/70.50
XLM-RoBERTa-large	82.21/82.97	74.56/64.85	73.26/68.26	75.71/71.95	71.13/70.15	74.16/66.77	73.06/67.22	66.23/70.09	74.60/71.07
Qwen3-VL-8B-Instruct	87.29/88.23	84.15/79.48	78.83/77.05	81.44/80.20	79.36/82.93	78.32/78.84	78.01/80.84	71.30/70.63	80.83/81.32
Qwen-SEA-LION-v4-8B-VL	90.27/89.48	87.65/85.70	82.10/82.49	84.24/82.40	81.43/86.40	78.43/82.55	82.87/84.66	68.99/84.98	83.52/85.28
Llama-3.1-8B-Instruct	61.64/82.50	77.06/86.59	69.52/80.97	52.68/78.67	38.42/78.00	30.53/80.02	55.68/79.08	40.20/76.71	56.48/80.55
Qwen-3.5-9B	88.73/89.63	87.88/85.24	82.70/83.12	81.09/77.29	78.93/86.33	73.43/77.86	79.42/86.22	73.83/85.68	81.80/84.76
Qwen-3.5-27B	91.17/91.23	91.86/88.56	85.96/84.36	81.57/78.89	85.01/89.22	80.57/82.54	88.94/89.39	82.62/80.71	86.78/87.06
Gemma-3-27B-IT	93.08/91.54	91.12/88.01	87.13/84.66	82.63/81.07	84.77/87.47	82.77/81.89	88.36/89.78	76.21/78.72	87.16/86.81
Gemma-SEA-LION-v4-27B-IT	93.08/91.95	91.12/88.05	87.94/83.91	81.80/81.07	84.72/87.14	82.77/81.89	89.76/89.31	78.71/81.03	87.51/86.69
Gemma-4-31B-IT	93.14/91.97	93.21/89.82	87.20/84.73	83.32/81.51	87.77/**91.75**	81.03/84.56	88.41/89.79	87.83/87.92	88.26/88.56
Qwen3-32B	93.05/93.13	**97.27**/91.26	87.46/86.72	85.37/87.02	79.58/89.07	83.34/84.49	85.35/87.82	78.72/**92.92**	86.64/88.92
Qwen-SEA-LION-v4-32B-IT	93.10/93.15	95.22/90.54	88.82/86.41	86.97/85.21	85.40/90.76	86.03/83.36	89.72/88.33	83.51/85.63	89.04/88.78
Llama-3.1-70B-Instruct	89.79/93.50	90.17/90.46	86.01/89.45	82.31/87.81	73.64/90.00	75.46/84.97	86.21/88.30	65.48/85.86	82.68/89.52
Llama-3.3-70B-Instruct	**94.32**/93.58	93.83/**93.19**	**90.80**/**89.92**	**92.96**/**88.72**	82.06/91.70	85.22/**88.75**	88.75/89.67	73.55/88.23	88.70/**90.95**
GPT-5.4	93.19/**93.61**	95.28/90.48	89.23/85.06	86.92/82.72	87.78/91.46	86.01/84.53	90.17/**90.65**	**90.38**/87.92	89.90/89.11
Gemini-3-flash-preview	94.30/93.19	95.27/91.23	88.73/86.77	88.62/85.03	**89.08**/90.77	**88.73**/88.03	**90.68**/90.22	83.00/87.93	**90.54**/89.72
Hard Set									
mDeBERTa-v3-base	52.11/43.62	50.67/34.08	55.27/47.33	57.59/52.57	52.68/43.27	55.84/52.94	50.35/46.49	60.29/37.91	52.81/44.17
mmBERT-base	62.27/53.13	54.43/52.57	52.63/54.16	59.60/54.08	61.03/43.39	42.30/57.26	57.95/60.09	61.02/45.64	57.30/52.65
XLM-RoBERTa-large	56.52/46.85	58.75/30.05	53.67/43.09	59.96/48.93	60.21/45.75	52.48/42.49	54.80/50.74	56.52/47.10	56.29/44.65
Qwen3-VL-8B-Instruct	81.51/78.38	74.18/79.04	69.30/73.49	68.02/74.54	59.18/62.61	75.53/53.25	66.91/79.08	67.96/42.59	70.12/71.67
Qwen-SEA-LION-v4-8B-VL	82.41/84.91	83.52/82.17	80.39/77.96	68.53/76.78	65.48/71.78	80.06/69.57	69.40/**79.50**	**73.29**/48.28	74.98/77.18
Llama-3.1-8B-Instruct	23.69/64.82	56.12/73.95	54.78/66.17	19.18/57.45	34.25/68.56	26.73/77.46	14.53/54.75	22.01/49.87	32.43/64.02
Qwen-3.5-9B	76.04/76.38	74.96/80.75	76.72/75.38	66.97/69.70	67.32/69.11	56.72/68.83	55.35/69.25	45.65/50.23	68.02/72.08
Qwen-3.5-27B	82.62/81.69	90.88/82.89	83.80/85.47	73.17/71.83	73.24/66.19	76.63/75.54	71.40/69.88	61.17/57.12	77.98/75.56
Gemma-3-27B-IT	83.40/81.51	90.22/94.27	90.97/89.17	65.25/69.02	70.13/68.95	80.53/75.58	63.96/63.13	60.56/67.66	76.85/76.53
Gemma-SEA-LION-v4-27B-IT	83.40/83.68	90.22/94.27	**91.81**/89.17	68.89/69.02	71.27/68.95	80.53/73.70	63.96/63.13	60.56/67.66	77.53/76.90
Gemma-4-31B-IT	82.92/79.22	94.71/83.27	83.61/78.70	72.41/67.81	71.01/65.49	73.61/67.41	71.49/70.96	65.36/46.25	77.87/72.78
Qwen-3-32B	85.35/85.48	91.72/89.94	91.75/90.22	76.47/78.75	67.62/73.54	75.55/79.36	72.24/69.40	67.66/**68.95**	79.48/80.23
Qwen-SEA-LION-v4-32B-IT	86.28/84.61	93.31/94.71	89.50/90.46	**82.40**/**80.43**	**73.27**/72.96	79.24/75.80	74.24/71.71	72.77/61.17	**81.85**/80.61
Llama-3.1-70B-Instruct	78.89/85.51	84.65/91.88	86.00/91.88	65.64/74.06	60.25/**74.04**	73.62/83.02	68.14/74.00	49.44/62.15	72.96/81.08
Llama-3.3-70B-Instruct	83.33/**88.37**	86.59/90.49	90.05/**93.83**	72.70/75.79	67.39/73.51	83.90/**84.21**	71.02/76.24	55.23/67.66	77.96/**82.45**
GPT-5.4	87.34/85.14	90.80/**95.98**	87.93/85.43	71.85/72.32	68.63/66.77	79.83/78.00	70.68/69.04	61.17/61.17	78.68/77.51
Gemini-3-flash-preview	**89.78**/82.59	**96.00**/84.35	87.05/82.08	74.42/68.45	68.32/66.06	**86.53**/73.58	**77.38**/72.22	72.77/48.46	81.75/74.88

Table 15: F1 score on the SEA NLI benchmark across languages. We format the results in the SEA/English result format, where the column represents the culture of the country.

evaluated samples as weights to obtain the overall human performance. The weighted average human accuracy is 95.74% on the normal subset, 88.83% on the hard subset, and 93.44% overall. These results show that the hard subset is substantially more challenging even for native human annotators, providing additional context for interpreting model performance.

P Data Licensing

This dataset is released under the permissive MIT License. Please note that the seed data originates from Wikipedia under a CC-BY-SA license, and the synthetic extensions were generated via OpenAI's GPT-5.2 model (with output ownership fully assigned to us per OpenAI's terms).