KaiToM: A Benchmark for Culturally Informed Theory-of-Mind in Thai Contexts

Kun Kerdthaisong[1,†] **Pasin Buakhaw**[2,†] **Wayupuk Sommuang**[1,†]
Nutchanon Yongsatianchot[3,*]

[1]Sirindhorn International Institute of Technology, Thammasat University
[2]Department of Computer Engineering and Digital Technology,
Faculty of Engineering, Chulalongkorn University
[3]Siriraj Genomics, Faculty of Medicine Siriraj Hospital,
Mahidol University, Bangkok, Thailand

Abstract

Large Language Models (LLMs) show impressive reasoning capabilities, yet their understanding of culturally specific social dynamics remains limited. We present KaiToM, a Theory of Mind (ToM) benchmark for Thai cultural contexts, focusing on *Kreng-Jai*, a uniquely Thai concept of social reluctance and consideration. Extending TactfulToM, we construct 882 conversations (39.2K questions) spanning three Kreng-Jai dimensions: White Lies, Avoidance of Confrontation, and Fear-Based Respect. To disentangle language proficiency from cultural knowledge, the dataset includes Thai scenarios at varying cultural depths, general English scenarios, and English translations of Thai content. We evaluate Thai-specialized models(Typhoon-v2.1, SEA-LION-v3), open-weight models(Gemma-3, DeepSeek-V3.2, Kimi-K2, Qwen), and proprietary models (GPT-4o-mini, GPT-5.1, Gemini-3-Flash). Models achieve only moderate performance, struggle with Fear-Based Respect, and cannot reliably identify specific Kreng-Jai types, a failure shared by Thai-specialized models. Human evaluation shows that native speakers significantly outperform models on culturally loaded tasks. Performance also degrades in deep relative to surface-level cultural contexts, and generalist models often perform better on English translations than on the original Thai, suggesting that models capture surface cues but lack deeper cultural alignment. KaiToM supports the development of culturally aware systems for low-resource languages. Code is available at https://github.com/Augustus2011/KAI-ToM.

1 Introduction

Large Language Models (LLMs) have shown strong performance in static reasoning and knowledge retrieval tasks (Gao et al., 2025), yet their ability to reason about other people's beliefs, intentions, and knowledge a capacity known as Theory of Mind (ToM), the ability to attribute mental states to oneself and others (van Duijn et al., 2023; Wu et al., 2023) remains far less well understood.

In LLM evaluation, ToM studies have gained attention in recent years as a way to test whether LLMs can infer mental states from a given situation, leading to a growing line of benchmarks and analyses, including FANToM, Hi-ToM, ToM-SSI, and DecomposeToM (Kim et al., 2023; Wu et al., 2023; Bortoletto et al., 2025; Sarangi et al., 2025; Le et al., 2019). Although LLMs exhibit some social reasoning ability, it remains unclear whether this reflects genuine understanding or superficial pattern matching (Jin et al., 2024). Recent studies have shown that LLMs struggle with ToM reasoning across different languages. (Chan et al., 2025).

Thai society is rooted in Buddhist philosophy and a culturally sanctioned social hierarchy (Mulder, 1979), with a strong orientation toward group alignment and the maintenance of surface harmony (Komin, 1990). A capture vocab of these norms is **Kreng-Jai**, a behavior of social restraint, consideration, and the avoidance of interpersonal discomfort (Wyatt and Promkandorn, 2012). Because cultural understanding is reflected in the vocabulary people use (Wierzbicka, 1997),Kreng-Jai can be placed alongside broader Asian "harmony vocabularies" such as Burmese **Arr-nar** (Khirikoekkong et al., 2023). and Japanese **Enryo** (Takita, 2012)

Behaviorally, acting with Kreng-Jai requires an individual to estimate what another person knows, feels, or may be hurt by, and then suppress or adjust their own words and actions to protect that person's emotional comfort (Intachakra, 2012; Komin, 1990). We argue that this makes Kreng-Jai a culturally situated form of ToM: one shaped by Thai norms of hierarchy and harmony, rather than by universal social cognition alone.

[†]Equal contribution (co-first authors). [*]Corresponding author: nutjung.nutlc@gmail.com

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 2732–2755
November 6-10, 2026 ©2026 Association for Computational Linguistics

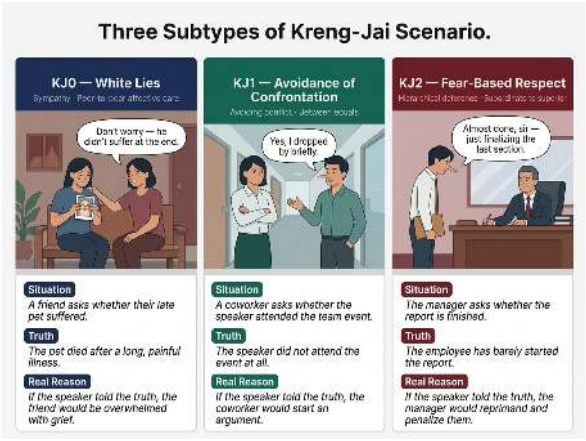


Figure 1: The three subtypes of Kreng-Jai scenarios. **Left panel (KJ0 - White Lies):** Peer-to-peer affective care where the lie protects the hearer from emotional pain. **Middle panel (KJ1 - Avoidance of Confrontation):** Lateral interactions between equals where the lie sidesteps an anticipated conflict. **Right panel (KJ2 - Fear-Based Respect):** Hierarchical deference where the lie shields the speaker from sanction by a superior. Each panel reports the situation, the lie uttered (Speech bubble), the actual motivation for lying (Real reason), and the underlying truth (Truth). The Pinocchio motif marks the liar in each scene.

One important benchmark in this area is TactfulToM (Liu et al., 2025), which operationalizes ToM through dialogue-based lie scenarios involving four social roles, and evaluates whether LLMs can reason about the beliefs and intentions behind socially motivated deception (Figure 2). We extend this framework to Thai cultural contexts by introducing three Kreng-Jai dimensions 1.White Lies 2.Avoidance of Confrontation 3.Fear-Based Respect and by varying scenarios across surface and deep cultural levels. Since prior work has shown that fluent generation in a language does not necessarily imply strong cultural knowledge (Limkonchotiwat et al., 2025), we further translate all Thai scenarios into English, allowing us to disentangle the contributions of language proficiency and cultural knowledge to LLM social reasoning.

We formulate three research questions:

RQ1: Can LLMs perform ToM reasoning in Thai cultural contexts, and how does their performance compare to native Thai speakers?

RQ2: Do LLMs understand Kreng-Jai behavior differently in Thai cultural contexts versus general Western English scenarios?

RQ3: Do LLMs require Thai language proficiency or Thai cultural knowledge to perform ToM reasoning in Thai cultural contexts? We measure this along two axes: (a) language axis original Thai versus English translations of identical content, and (b) cultural depth axis surface versus deep cultural embeddedness.

For RQ2, based on existing taxonomies of Kreng-Jai (Table 3), we select three Kreng-Jai subtypes related to ToM: **1. White Lies**, **2. Avoidance of Confrontation**, and **3. Fear-Based Respect**. We construct scenarios in two context(General English, Thai Cultural Contexts). To address RQ3, we further categorize Thai cultural scenarios into two levels of cultural embeddedness surface cultural level, where Thai elements serve as background setting, and deep cultural level, where Thai cultural beliefs fundamentally drive character behavior (Section 3.2) and additionally translate all Thai scenarios into English to create a parallel corpus that isolates language from cultural knowledge. To answer RQ1, we compare LLM predictions with human responses from native Thai speakers on a sampled subset of questions.

Our contributions:

- A dataset and generative seeds for evaluating ToM and Kreng-Jai in Thai cultural contexts, spanning three Kreng-Jai subtypes and two cultural depth levels
- An empirical study comparing LLM and native Thai human performance on ToM and Kreng-Jai identification within Thai cultural context scenarios, with separate analysis of the effects of language proficiency and cultural knowledge.

Dataset	Conv.	Qs.	Multi.	Cult.
FANToM	256	10K	–	–
XToM	900	15.1K	✓	–
TactfulToM	100	6.7K	–	–
KaiToM (Ours)	1072	39.2K	✓	✓

Table 1: Dataset comparison. Columns: number of conversations, number of questions, multilingual, cultural context. KaiToM includes TactfulToM; (Note: KaiToM includes TactfulToM)

2 Defining Kreng-Jai

In Table 3,we cluster the Kreng-Jai meanings from the 26 meanings identified by (Wyatt and Promkandorn, 2012), based on speaker motivation to act Kreng-Jai,we found that Kreng-Jai meaning can spans overlapping themes such as **Respect boundaries** and **Avoiding conflict**. Both types' goal is

Data Cate	Conv	Question
White Lies (KJ0)		
Thai-Surface	99	3,564
Thai-Deep	30	1,080
English (TactfulToM)	100	4,254
EN-Translated	129	4,644
Avoid Confront (KJ1)		
Thai-Surface	99	3,564
Thai-Deep	30	1,080
English	99	3,564
EN-Translated	129	4,644
Fear Respect (KJ2)		
Thai-Surface	99	3,564
Thai-Deep	30	1,092
English	99	3,564
EN-Translated	129	4,656

Table 2: Summary of KaiToM dataset

Theme	Meaning
Avoiding conflict	Avoid verbal confrontations to preserve relationships or save face. (Limanonda, 1995)
Respecting boundaries	Not causing trouble or inconvenience. (Komin, 1990)
Humility	Accurate self-assessment to behaving appropriately. (Phillips, 1965)
Emotional restraint	Suppressing feelings, not asking help. (Klausner, 2000)
Polite	Polite thank you, polite refusal. (Khirikoekkong et al., 2023)
Hierarchy	Fear of offending superior lead to penalty. (Sriussadaporn-Charoenngam and Jablin, 1999)
Sympathy	Feeling sympathy for another person and support their goal. (Vongvipanond, 1994)

Table 3: Definitions of Kreng-Jai's theme by Thai and Foreign Authors

avoiding unwanted situations but different speaker motivations: first is driven by a desire to respect others' boundaries, while the other simply seeks to escape the present situation. (Intachakra, 2012) describe it as a principle of "self-effacement and other-accommodation," centered on respect for another's internal emotional state. Because Kreng-Jai is specific word that representent Thai norm cultural it easily misread by outsiders as evasiveness, so we requires a clear definition of this behaivior before cross-cultural ToM evaluation. We operationalize it within lying situations for two reasons: (1) lying is inherently ToM-laden, requiring the speaker to model the hearer's knowledge and expectations; and (2) prosocial or other-protective lies (DePaulo et al., 1996) directly instantiate self-effacement and other-accommodation, while remaining interpretable to models trained on Western data. We introduce a new question type, **"KJ-ability"**, to measure whether an evaluator can infer a specific subtype of Kreng-Jai from lying scenarios.

From the catalogued Kreng-Jai meanings, we select three subtypes satisfying two criteria: **(i)** each can be cast as a controlled ToM scenario in which the speaker pursues a distinct communicative goal (e.g., consoling the hearer or avoiding an unwanted engagement), requiring the speaker to model different hearer mental states, which makes the subtypes separable as ToM tasks; and **(ii)** each can be instantiated both in everyday interactions and in traditional practices with their own normative rules, ensuring that Thai-grounded scenarios are tied to real cultural norms rather than merely renamed English scenarios. The three retained subtypes are **White Lies (KJ0)**, targeting the hearer's *affective vulnerability* in peer settings (what would make them sad); **Avoidance of Confrontation (KJ1)**, targeting *anticipated displeasure* between equals (what would provoke conflict); and **Fear-Based Respect (KJ2)**, targeting a superior's *authority and capacity to punish* (what would invite consequences). Together, these subtypes span the social range of Kreng-Jai: affective care, lateral conflict avoidance, and hierarchical deference.

2.1 White Lies (KJ0)

White lies are a form of other-oriented lie: lies told to protect or benefit another person psychologically, or to protect their interests (DePaulo et al., 1996). In ToM terms, white lies require speakers to infer communicative intent and consider the listener's emotional state, making them a form of nonliteral communication similar to irony and sarcasm. (Beaudoin et al., 2020; Liu et al., 2025).

In the KJ0 scenario (see Figure 1, left panel), adapted from TactfulToM (Liu et al., 2025), a friend who has just lost her pet asks the speaker whether the pet suffered. The truth is that the pet died after a long and painful illness, yet the speaker replies, *"Don't worry, he didn't suffer at the end."* We define KJ0 as scenarios in which truthful disclosure would cause emotional distress and the liar chooses a false but comforting reply without gaining material benefit, aligning with the **Sympathy** dimension of Kreng-Jai (Table 3): considerate care that avoids emotional burden for others (Mekthawornwathana, 2012; Komin, 1990).

2.2 Avoidance of Confrontation (KJ1)

KJ1 concerns situations in which revealing the truth may provoke anger, dissatisfaction, or interpersonal tension. This subtype is related to the Kreng-Jai dimension of **Avoiding Conflict** and **Respecting Boundaries** (Table 3), where the speaker avoids causing trouble or discomfort for another person (Mekthawornwathana, 2012).

In the KJ1 scenario (Figure 1, middle panel), a coworker asks whether the speaker attended a team event. The speaker did not attend at all, yet replies, *"Yes, I dropped by briefly."* We define KJ1 as scenarios in which the liar expects the target to become **angry or unsatisfied** if know the truth therefore lair tells a lie to avoid confrontation, where interlocutors are casually has equal social standing and the lie is driven by avoidant conflict rather than sympathy.

2.3 Fear-Based Respect (KJ2)

KJ2 captures hierarchical situations in which a subordinate withholds the truth from a superior out of fear of that superior rather than concern for their feelings, reflecting the literal meaning of **kreng** ('respectful fear') in Kreng-Jai (Wyatt and Promkandorn, 2012). This fear-based withholding corresponds to the **Hierarchy Authority** dimension in Table 3. Thai subordinates are expected to defer upward, making candid communication socially costly (Komin, 1990), and cross-cultural evidence links high power distance to **fear of authority** and weakened upward communication (Dai et al., 2022). Crucially, KJ2 is not mere polite deference: the liar invokes a culturally sanctioned norm of hierarchical respect (Mekthawornwathana, 2012) to enable a fear-driven, self-protective act (Kish-Gephart et al., 2009). This aligns with defensive silence—withholding information to avoid personal consequences rather than to benefit the target—and contributes to organizational silence (Morrison and Milliken, 2000). Functionally, KJ2 is closer to a *self-centered lie* in DePaulo's typology (DePaulo et al., 1996) than to the altruistic white lie central to KJ0, despite its prosocial Kreng-Jai framing.

In the KJ2 scenario (Figure 1, right panel), a manager asks whether a report is finished. The truth is that the employee has barely started, yet replies, *"Almost done, sir—just finalizing the last section."* We define KJ2 as scenarios in which the liar occupies a subordinate position relative to the target and truth-telling would primarily harm the liar rather than the target. **Power asymmetry** is thus the defining condition: unlike KJ1, where interlocutors share equal standing and the lie avoids mutual discomfort, and KJ0, where the lie shields the target from emotional pain, the false statement in KJ2 protects the liar from the reprimand or punishment that honest disclosure would provoke.

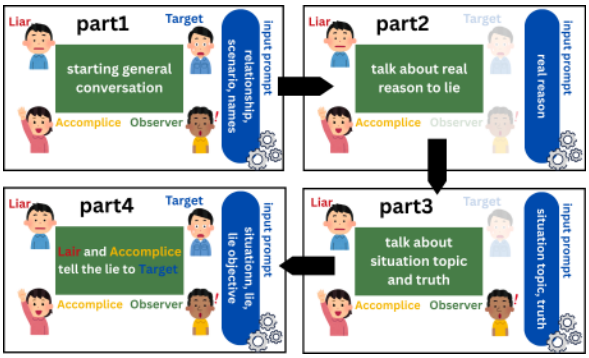


Figure 2: TactfulToM framework generates conversations in four sequential parts with four roles: the *Liar*, *Target*, *Accomplice*, and *Observer*. *(Part 1)* All four engage in a general conversation. *(Part 2)* Target and Observer leave; Liar and Accomplice privately discuss the real reason to lie. *(Part 3)* Observer returns; the conversation shifts to the situation topic and truth is revealed. *(Part 4)* Target returns; **Liar and Accomplice deliver the lie to the Target.**

3 Data Construction

Our data construction pipeline follows three sequential stages: **(1) Generative Seeds**: seed creation by human annotators; **(2) Conversation Generation** LLM-based conversation generation; and **(3) Question Generation** quality verification and question generation (Figure 4).

3.1 Generative Seeds

We define a *generative seed* as a structured set of scenario parameters manually crafted by annotators to prompt an LLM for conversation generation. Each seed contains two groups of parameters: *scenario details* (scenario topic, situation, and lie objective) and *lie details* (real reason, lie, and truth), as shown in Appendix Table 8. To maintain role-agnostic reusability, seed fields are written using role placeholders **A** (Liar), **B** (Target), **C** (Accomplice), and **D** (Observer) instead of specific character names, allowing the same seed to be instantiated with different name assignments at generation time. Three annotators collaboratively created and tested seeds across all three Kreng-Jai types (KJ0, KJ1, KJ2) for both Thai and English contexts using our



Figure 3: Illustrative examples of Thai context scenarios under **Avoidance of Confrontation (KJ1)**, contrasting two cultural depth levels(The Pinocchio motif marks the liar in each scene). **Left (Deep Cultural Level):** During a Thai Vegetarian Festival. **Right (Surface Cultural Level):**A group of government officers discuss a local election candidate over a meal at a local Thai restaurant.

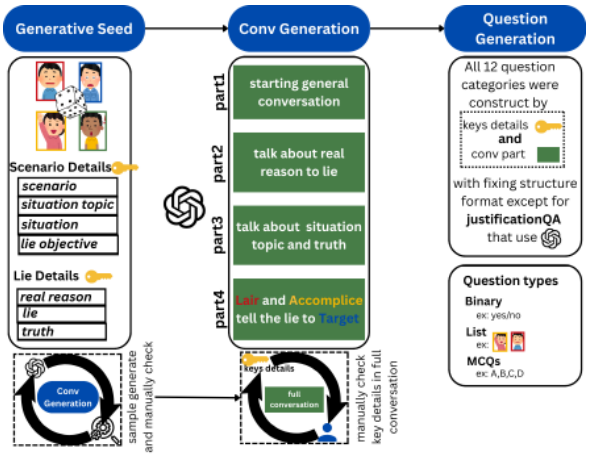


Figure 4: Data generation pipeline consisting of three stages. **Left (Generative Seed):** Annotators manually create and validate *generative seeds*. **Center (Conv Generation):** GPT-4o generates conversations following the four-part TactfulToM framework. **Right (Question Generation):** Questions are constructed from key details and conversation parts across 10 types in three formats: binary, list, and MCQs, except for JustificationQA choices, which are generated by GPT-4o.

annotation platform (Appendix B.1), iterating until each seed satisfied our quality requirements: internal consistency, character role clarity, , and, for Thai scenarios, the presence of culturally appropriate entities.

3.2 Conversation Generation

Using the validated generative seeds, we prompted GPT-4o-2024-08-06 (OpenAI et al., 2024) to generate conversations following the four-part TactfulToM framework (Figure 2), as illustrated in Figure 4. Generation proceeds sequentially across four steps, where each step conditions on the accumulated conversation history from all previous steps to ensure narrative continuity. Before generation, role placeholders (A, B, C, D) embedded in each seed field are replaced with sampled character names and three leave reasons are randomly sampled to motivate natural character departures in Parts 2 and 3. Each step injects specific seed parameters into a fixed prompt template (Appendix H.2), with a temperature of 0.2 and a maximum of 1,000 output tokens (Appendix F).The variants of the generated conversation set are listed below:

Thai Context For Thai-language conversations, the system prompt instructs GPT-4o to generate dialogue in Thai while keeping character names in English (Appendix H.1). Four names are randomly sampled without replacement from a fixed pool of 20 romanized Thai names sourced from the Forebears database[1], as prompting with Thai-script names caused inconsistent personality representations across generated parts (Appendix A).

Thai Cultural Context For Thai context, annotators crafted seeds grounded in Thai cultural settings, ensuring that scenario topics, situations, and lie objectives reflect authentic Thai social dynamics across three Kreng-Jai types. To capture varying degrees of cultural embeddedness, this section is categorized into two cultural depth levels, illustrated with examples in Figure 3

- **Deep Cultural Level:** Scenarios where Thai cultural beliefs are the main driver of behavior. As shown in Figure 3 (left), a volunteer accidentally drops fried garlic into the soup pot during the Thai Vegetarian Festival. The lie *"Nothing fell in, sir"* stems from the strict prohibition of the five pungent vegetable.
- **Surface Cultural Level:** Scenarios where Thai cultural elements provide contextual setting but do not drive character behavior. As shown in Figure 3 (right), government officers privately disagree with a colleague's support for a local election candidate but lie *"We*

[1] https://forebears.io/thailand/forenames

should give him a chance" to avoid damaging the friendship. No Thai-specific rule governs this.

English Translated To investigate RQ3, we translated all Thai conversations 3.2 into English using GPT-4o, preserving all cultural content and seed keywords while removing the Thai language component. This creates a parallel corpus that isolates the effect of language from cultural knowledge.

English Context For English context conversations, we follow the TactfulToM pipeline by sampling character names from the top 20% most frequent US birth names datasets[2].

After generation, three annotators manually verified that all key seed elements were correctly embedded across the four conversation parts.

3.3 Question Generation

For each generated conversation, we construct questions across the 10 question types listed in Table 4, following the same taxonomy as TactfulToM with the addition of **KJ-ability (KJ)**. KJ-ability measured across Thai and English scenarios serves as the primary instrument for addressing RQ2. Each question is constructed by filling fixed templates with key details extracted from the conversation and its seed such as character names, the lie, the truth, and the real reason as shown in Table 7. Exception for the **JustificationQA (JQ)**, whose multiple-choice distractors are generated by GPT-4o rather than derived from fixed templates, producing three semantically plausible but incorrect options per question (Appendix H.3). Questions are formatted in three answer types: **Binary** (yes/no exact matching), **List** (set-based evaluation), and **MCQs** (single correct option), with evaluation metrics defined accordingly (Appendix E).

4 Experiments

4.1 Models

We evaluate diverse sets of large language models to assess their Theory of Mind capabilities in Kreng-Jai contexts:

Thai-specialized models: We test models specifically trained on Thai language data, including Typhoon-v2.1-12b-instruct (Pipatanakul et al., 2024) and rich cultural Southeast Asian data SEA-LION-v3-9B (Ong and Limkonchotiwat, 2023), to assess their cultural and linguistic understanding.

Open-source models: We include open models such as the Gemma-3 series(gemma-3-12B, gemma-3-27B) (Gemma Team et al., 2025), DeepSeek-V3.2 (DeepSeek-AI et al., 2025), Kimi-K2-Instruct (Kimi Team et al., 2026), and the Qwen series(Qwen2.5-72B, Qwen3-30B-A3B,Qwen2.5-7B) (Qwen Team et al., 2025a,b).

Proprietary models: We evaluate commercial models including GPT-4o-mini (OpenAI, 2024), GPT-5.1-2025-11-13 with low reasoning effort (OpenAI, 2025) and Gemini-3-flash-preview (Doshi, 2025).

Question Type	To Test
Comprehension	Detect false statements
Justification	Reasoning about character motivations
Fact_Reason	Who discussed the real reason
Fact_Truth	Distinguishing factual truth from stated information
Belief	Inferring character beliefs
InfoAccessibility	Who has access to which information
Answerability	Determining what characters can answer based on their knowledge
LieDetectability	Predicting whether lies will be detected
LieAbility	Assessing characters' ability to lie convincingly
KJ-ability	Identifying the type of Kreng-Jai motivating a lie

Table 4: Question types used in the KaiToM evaluation

4.2 Evaluation Setup

To evaluate models, we using Zero-Shot on **KaiToM dataset** consist of three types of answers: multiple-choice questions (MCQs), list generation, and binary (yes/no) responses. Additionally, we test chain-of-thought (CoT) reasoning (Wei et al., 2022) by appending "Let's think step by step" to the prompts.

4.3 Human Evaluation

To establish a human performance baseline for addressing RQ1, we hired more 9 annotators age between 22 to 28 years old and live in central of thailand who met our minimum qualification requirements(Appendix D). Each annotator evaluated 10 scenario sets, stratified as follows: 3 surface-level Thai cultural scenarios, 3 deep-level Thai cultural scenarios, and 4 English scenarios. The scenarios were balanced across the three types of Kreng-Jai.

[2] https://www.kaggle.com/datasets/ryanburnsworth/popular-names-by-birth-year-1880-2022

the form[3].

inter-annotator agreement (IAA) For scenario validation (binary: Does the character exhibit Kreng-Jai?), the agreement by Gwet's AC1 (Gwet, 2008) is 0.73, for ToM question correctness the pooled agreement is Gwet's AC1 = 0.32.

5 Results

After converting all outputs to binary scores (Appendix E), we find that CoT prompting yields a slight improvement on social reasoning tasks (Figure 5). Notably, in Thai scenarios, performance degrades in **deep cultural contexts** compared to **surface-level contexts**. This suggests that while models can handle superficial cultural cues, they struggle with the subtle, high-context nuances inherent in deep Thai cultural interactions.

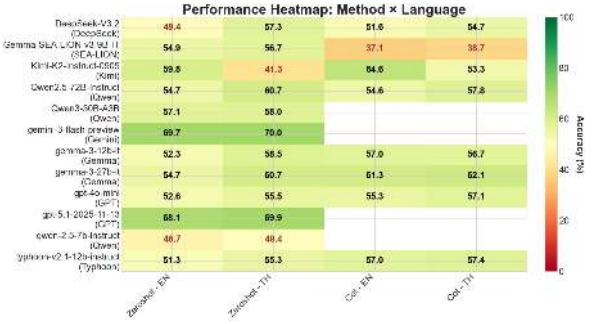


Figure 5: Performace summary average Accuracy score from MCQs/Binary/List across all KJ0,KJ1,KJ2; Heatmap promting method x language where qwen2.5-7b instruct, qwen3-30B-A3B and GPT-5.1 are no result in CoT prompting; (note: en-traslated not included)

RQ1: Capability in Thai Cultural Contexts. Our findings indicate that LLMs possess a baseline capability to perform Theory of Mind (ToM) social reasoning tasks in Thai cultural contexts; however, overall performance is moderate, with accuracy ranging from 41.3% to 70.0%. Compared with general English scenarios, the top-performing models reach proficiency equivalent to, and in some cases exceeding, their English baselines (Figure 5).

Despite these capabilities, a substantial gap remains between machine and human social cognition. As shown in Figure 6, humans consistently outperform LLMs, including the top-performing Gemini-3-Flash-preview, particularly on complex tasks such as *liedetectabilityQAs* and *lieabilityQAs*.

[3] https://docs.google.com/forms/d/e/1FAIpQLSf9xceW74aRO1FUl4-nd3G_InSehGEEWMwuCx9qHgbMaxo6Bw/viewform

This human advantage on culturally loaded categories is highly significant after Holm correction (Holm, 1979): lie detectability (list) (0.90 vs. 0.34, $p_{\text{Holm}} = 7.5 \times 10^{-23}$), KJ ability (0.80 vs. 0.35, $p_{\text{Holm}} = 1.4 \times 10^{-16}$), and lieability (0.70 vs. 0.38, $p_{\text{Holm}} = 1.2 \times 10^{-8}$). In contrast, models significantly outperform humans only on mechanical set-tracking or template-like categories (answerability-list, beliefQAs, and fact_truthQA).

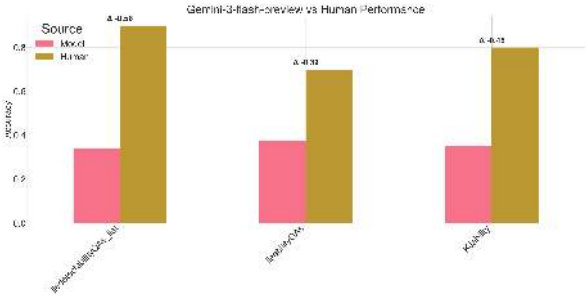


Figure 6: Average score from human vs Gemini-3-flash-preview, Across 10 scenarios

RQ2: Understanding Kreng-Jai (Thai vs. Western Scenarios). We investigate whether LLMs can distinguish and reason about *Kreng-Jai* behaviors across cultures. The results (Table 6) show that models generally achieve higher scores in English scenarios (EN) than in Thai scenarios (TH) on comprehension (CQ) and belief (BQ) questions.

This cultural context effect is statistically robust: 9 of the 12 models show a significant Thai-vs.-Western difference after Holm correction ($p_{\text{Holm}} < 0.05$), with per-model gaps of up to 0.23 (Kimi-K2). The direction of the effect, however, varies across models, which itself suggests that models process the two cultural contexts (Thai and Western) differently rather than applying a single, uniform social reasoning ability.

For **KJ1** and **KJ2** specifically, models such as GPT-4o-mini and DeepSeek-V3.2 show a marked performance drop in Thai contexts compared with English. This discrepancy suggests that while models can reason about social friction (a universal concept) in English, they fail to map it effectively onto the specific Thai context.

RQ3: Language Proficiency vs. Cultural Knowledge. To determine if social reasoning requires native Thai language proficiency or intrinsic cultural knowledge, we analyzed the performance of generalist models against Thai-optimized models (e.g., Typhoon-v2.1-12b, SEA-LION-v3-9B) using

Model	L	KJ0				KJ1				KJ2			
		KJ		FQA		KJ		FQA		KJ		FQA	
		D	S	D	S	D	S	D	S	D	S	D	S
SEA-LION-v3-9B	TH	.50	.82	.40	.55	**.20**	**.39**	.45	.41	.50	.33	**.43**	.47
GPT-4o-mini	TH	.37	.56	.40	**.58**	.00	.04	**.65**	**.44**	**.93**	**.92**	.38	**.67**
Typhoon-12B	TH	**.67**	**.84**	**.50**	.45	.13	.08	.60	.23	.36	.37	**.43**	.44
SEA-LION-v3-9B	EN-T	.60	.77	.60	**.47**	**.43**	**.70**	.40	.23	.10	.05	.43	.42
GPT-4o-mini	EN-T	**.83**	**.99**	**.66**	.41	.23	.27	**.50**	**.32**	.27	.26	**.48**	**.58**
Typhoon-12B	EN-T	**.83**	.90	.60	.42	.23	.25	.45	**.32**	**.33**	**.31**	.29	.56

Table 5: Model Performance results by D=Deep and S=Surface cultural level on thai scenarios and english-traslated scenarios at Kreng-Jai ability(KJ) and fact_truthQA(FQA), focus on thai-specialized models like Typhoon-v2.1-12b-instruct, SEA-LION-v3-9B

both original Thai and English-translated counterparts (Table 5).

We observe two key trends: **Translation Effectiveness:** Generalist models (e.g., GPT-4o-mini) often perform better on Thai cultural scenarios when they are translated into English (EN-T) rather than processed in Thai (TH). For instance, GPT-4o-mini's score on KJ0 (Deep) jumps from 0.37 (TH) to 0.83 (EN-T). This implies the model possesses the *reasoning* capacity but lacks the *linguistic-cultural mapping* in Thai.

Limitation of Local Models: Surprisingly, even Thai-specialized models (Typhoon, SEA-LION) struggle to infer the specific mental states related to **KJability**. While Typhoon-12B performs well on surface-level KJ0, it fails to generalize to KJ1, while rich Southeast Asian cultural knowledge like SEA-LION performs well on KJ1 but drastically struggle to KJ2

This indicates that current Thai proficiency in LLMs does not automatically equate to Thai *cultural* proficiency. Models can process the language but lack the "Deep" cultural context required to identify the specific nuance of *Kreng-Jai* behind a lie.

6 Conclusion

In this work, we introduced a novel dataset to evaluate Theory of Mind (ToM) reasoning within the specific cultural framework of Thai *Kreng-Jai*. Our experiments reveal that while many LLMs exhibit emerging capabilities in Thai social reasoning, they still lag significantly behind human intuition, particularly in culturally motivated lies. We found that performance drops notably in "Deep" cultural contexts compared to surface-level. Furthermore, our cross-lingual analysis demonstrates that translating cultural scenarios into English often improves social reasoning performance for generalist models, suggesting that current limitations stem from a lack of alignment between the Thai language and cultural-level social concepts. Crucially, even models explicitly trained on Thai data struggle to categorize complex social dynamics like *Fear-Based Respect*, highlighting the need for future research to move beyond linguistic fluency toward cultural alignment in LLM development.

Model	L	Avoid Confront (KJ1)					Fear Respect (KJ2)				
		BQ	CQ	JQ	LA	LD	BQ	CQ	JQ	LA	LD
GPT-5.1	TH	.44	.43	.53	.37	.57	.41	.43	.50	.33	.60
	EN	.66	.45	.56	.33	.56	.50	.56	.46	.39	.55
GPT-4o-mini	TH	.35	.17	.43	.35	.47	.29	.12	.50	.32	.45
	EN	.45	.51	.39	.28	.57	.34	.61	.34	.48	.48
Gemini-3	TH	**.57**	.48	.74	.40	.59	**.51**	.44	**.78**	.35	**.62**
	EN	.71	.66	**.83**	**.48**	**.73**	.59	.72	**.77**	.43	**.61**
DeepSeek-V3.2	TH	.37	.51	.63	.31	.51	.35	.49	.61	.27	.51
	EN	.39	**.93**	.61	.28	.45	.31	**.95**	.52	.38	.43
Kimi-K2	TH	.49	**.57**	**.75**	.28	.50	.14	.12	.19	.10	.15
	EN	**.75**	.70	.78	.30	.51	**.66**	.81	.70	.47	.43
Qwen2.5-72B	TH	.34	.28	.47	.37	.56	.29	.26	.36	.35	.54
	EN	.52	.46	.46	.33	.53	.36	.63	.29	.35	.43
Qwen3-30B-A3B	TH	.39	.48	.57	.20	.48	.37	**.61**	.56	.23	.46
	EN	.51	.52	.53	.25	.50	.37	.44	.32	.35	.49
Gemma-3-27B	TH	.31	.28	.54	**.47**	.56	.31	.27	.62	.40	.55
	EN	.40	.66	.65	.36	.49	.31	.71	.46	**.50**	.43
Gemma-3-12B	TH	.25	.23	.47	.31	**.63**	.26	.17	.46	.31	.58
	EN	.35	.46	.35	.21	.53	.28	.61	.31	.28	.45
Typhoon-12B	TH	.27	.23	.46	.36	.57	.25	.15	.45	.29	.51
	EN	.34	.46	.32	.21	.50	.29	.58	.30	.26	.43
SEA-LION-v3-9B	TH	.26	.31	.57	.43	.44	.23	.21	.53	**.42**	.41
	EN	.37	.38	.49	.41	.60	.27	.55	.39	.47	.51
Qwen-2.5-7B	TH	.23	.53	.74	.37	.56	.20	.52	.52	.39	.51
	EN	.30	.60	.55	.42	.55	.22	.82	.36	.45	.48

Table 6: Model performance comparison. L: language, Thai scenarios (TH) vs. English scenarios (EN). Question categories: BQ (belief QA), CQ (comprehension QA), JQ (justification QA), LA (lieability QA), and LD (lie detectability). Scores are averaged over binary and list question types.

Ethical Statement

We hired 9 additional annotators to establish a human performance baseline for our evaluation. Annotators were compensated at a rate exceeding the local minimum wage for approximately one hour of work, which involved completing evaluations through a Google Form questionnaire. **Use of AI**: We utilized Claude Code for reproducing the TactfulToM pipeline.

Limitations

This work has several limitations. First, we had insufficient API compute budget to conduct rigorous statistical testing. Second, computational resource constraints prevented us from exploring fine-tuning techniques or better approach to solve cultural adaptation of theory of mind in LLM.

Acknowledgments

This research has received funding support from the National Science, Research and Innovation Fund (NSRF) via the Program Management Unit for Human Resources & Institutional Development, Research and Innovation [grant number B13F680099].

The first author acknowledges the scholarship under the Thailand Advanced Institute of Science and Technology and Tokyo Institute of Technology (TAIST-Tokyo Tech) Program, awarded by the National Research Council of Thailand, National Science and Technology Development Agency, and Sirindhorn International Institute of Technology, Thammasat University.

References

Cindy Beaudoin, Élizabel Leblanc, Charlotte Gagner, and Miriam H. Beauchamp. 2020. Systematic review and inventory of theory of mind measures for young children. *Frontiers in Psychology*, Volume 10 - 2019.

Matteo Bortoletto, Constantin Ruhdorfer, and Andreas Bulling. 2025. ToM-SSI: Evaluating theory of mind in situated social interactions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32252–32277, Suzhou, China. Association for Computational Linguistics.

Chunkit Chan, Yauwai Yim, Hongchuan Zeng, Zhiying Zou, Xinyuan Cheng, Zhifan Sun, Zheye Deng, Kawai Chung, Yuzhuo Ao, Yixiang Fan, Cheng Jiayang, Ercong Nie, Ginny Y. Wong, Helmut Schmid, Hinrich Schütze, Simon See, and Yangqiu Song. 2025. Xtom: Exploring the multilingual theory of mind for large language models. *Preprint*, arXiv:2506.02461.

Yuwan Dai, Hao Li, Wenting Xie, and Tianyi Deng. 2022. Power distance belief and workplace communication: The mediating role of fear of authority. *International Journal of Environmental Research and Public Health*, 19(5).

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Bella M DePaulo, Deborah A Kashy, Susan E Kirkendol, Melissa M Wyer, and Jennifer A Epstein. 1996. Lying in everyday life. *Journal of personality and social psychology*, 70(5):979.

Tulsee Doshi. 2025. Gemini 3 Flash: frontier intelligence built for speed. Google Blog. On behalf of the Gemini team.

Yunfan Gao, Yun Xiong, Yijie Zhong, Yuxi Bi, Ming Xue, and Haofen Wang. 2025. Synergizing rag and reasoning: A systematic review. *Preprint*, arXiv:2504.15909.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, and 1 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70.

Songthama Intachakra. 2012. Politeness motivated by the 'heart' and 'binary rationality' in thai culture. *Journal of Pragmatics*, 44(5):619–635. Towards an Emancipatory Pragmatics Part Two.

Chuanyang Jin, Yutong Wu, Jing Cao, Jiannan Xiang, Yen-Ling Kuo, Zhiting Hu, Tomer Ullman, Antonio Torralba, Joshua Tenenbaum, and Tianmin Shu. 2024. MMToM-QA: Multimodal theory of mind question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16077–16102, Bangkok, Thailand. Association for Computational Linguistics.

Napat Khirikoekkong, Supa-at Asarath, Suphak Nosten, Borimas Hanboonkunupakarn, Nattapat Jatuporn-pimol, Jennifer Roest, Michael Parker, Francois Nosten, Rose McGready, Phaik Yeong Cheah, and Maureen Kelley. 2023. Culturally responsive research ethics: How the socio-ethical norms of arr-nar/kreng-jai inform research participation at the thai-myanmar border. *PLOS Global Public Health*, 3(5):1–25.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore. Association for Computational Linguistics.

Kimi Team, Yifan Bai, Yiping Bao, Y. Charles, Cheng Chen, Guanduo Chen, Haiting Chen, Huarong Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, and 181 others. 2026. Kimi k2: Open agentic intelligence. *Preprint*, arXiv:2507.20534.

Jennifer J. Kish-Gephart, James R. Detert, Linda Klebe Treviño, and Amy C. Edmondson. 2009. Silenced by fear:: The nature, sources, and consequences of fear at work. *Research in Organizational Behavior*, 29:163–193.

William J. Klausner. 2000. *Reflections on Thai Culture: Collected Writings of William J. Klausner*, 5th edition. The Siam Society, Bangkok.

Suntaree Komin. 1990. Psychology of the thai people : values and behavioral patterns.

KBTG Labs, :, Anuruth Lertpiya, Danupat Khamnuansin, Kantapong Sucharitpongpan, Pornchanan Balee, Tawunrat Chalothorn, Thadpong Pongthawornkamol, and Monchai Lertsutthiwong. 2026. Thalle-thaillm: Domain-specialized small llms for finance and thai – technical report. *Preprint*, arXiv:2601.04597.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. 2019. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 5872–5877, Hong Kong, China. Association for Computational Linguistics.

Bhassorn Limanonda. 1995. Families in thailand: Beliefs and realities. *Journal of Comparative Family Studies*, 26(1):67–82.

Peerat Limkonchotiwat, Pume Tuchinda, Lalita Lowphansirikul, Surapon Nonesung, Panuthep Tasawong, Alham Fikri Aji, Can Udomcharoenchaikit, and Sarana Nutanong. 2025. WangchanThaiInstruct: An instruction-following dataset for culture-aware, multitask, and multi-domain evaluation in Thai. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3535–3558, Suzhou, China. Association for Computational Linguistics.

Yiwei Liu, Emma Jane Pretty, Jiahao Huang, and Saku Sugawara. 2025. TactfulToM: Do LLMs have the theory of mind ability to understand white lies? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 25043–25061, Suzhou, China. Association for Computational Linguistics.

Thasanee Mekthawornwathana. 2012. The comprehensive meanings of krengchai and the important aspects of thai culture. *NIDA Development Journal*, 52(1):235–263.

Elizabeth Wolfe Morrison and Frances J. Milliken. 2000. Organizational silence: A barrier to change and development in a pluralistic world. *The Academy of Management Review*, 25(4):706–725.

N. Mulder. 1979. *Everyday Life in Thailand: An Interpretation*, duang kamol edition. Cambridge University Press.

David Ong and Peerat Limkonchotiwat. 2023. SEA-LION (Southeast Asian languages in one network): A family of Southeast Asian language models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 245–245, Singapore. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, and et al. Alan Hayes. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. Accessed: 2026-01-06.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/. Accessed: 2026-01-18.

OpenThaiGPT and ThaiLLM Team. 2025. Openthaigpt-thaillm-8b-v7.2-research-preview: A thai knowledge-focused language model. https://huggingface.co/openthaigpt/openthaigpt-thaillm-8b-instruct-v0.7.2-research-preview.

Herbert P. Phillips. 1965. *Thai Peasant Personality: The Patterning of Interpersonal Behavior in the Village of Bang Chan*, 1st edition. University of California Press, Berkeley.

Kunat Pipatanakul, Potsawee Manakul, Natapong Nitarach, Warit Sirichotedumrong, Surapon Nonesung, Teetouch Jaknamon, Parinthapat Pengpun, Pittawat Taveekitworachai, Adisai Na-Thalang, Sittipong Sripaisarnmongkol, Krisanapong Jirayoot, and Kasima Tharnpipitchai. 2024. Typhoon 2: A family of open text and multimodal thai large language models. *Preprint*, arXiv:2412.13702.

Kunat Pipatanakul and Pittawat Taveekitworachai. 2026. Typhoon-s: Minimal open post-training for sovereign large language models. *Preprint*, arXiv:2601.18129.

Qwen Team, 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, and 42 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang Qwen Team, Baosong Yang, Beichen Zhang, Binyuan Hui, and et al. Bo Zheng. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Sneheel Sarangi, Maha Elgarf, and Hanan Salam. 2025. Decompose-ToM: Enhancing theory of mind reasoning in large language models through simulation and task decomposition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10228–10241, Abu Dhabi, UAE. Association for Computational Linguistics.

Nongluck Sriussadaporn-Charoenngam and Fredric M Jablin. 1999. An exploratory study of communication competence in thai organizations. *The Journal of Business Communication (1973)*, 36(4):382–418.

Fuyuko Takita. 2012. Reconsidering the concept of negative politeness 'enryo' in japan. *Hiroshima Studies in Language and Language Education*, 15:189–196.

Max van Duijn, Bram van Dijk, Tom Kouwenhoven, Werner de Valk, Marco Spruit, and Peter van der Putten. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 389–402, Singapore. Association for Computational Linguistics.

Peansiri Vongvipanond. 1994. Linguistic perspectives of thai culture. In *Workshop of Teachers of Social Science*. University of New Orleans. Retrieved from https://web.archive.org/web/20121120/http://thaiarc.tu.ac.th/thai/peansiri.htm.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Anna Wierzbicka. 1997. *Understanding Cultures through Their Key Words: English, Russian, Polish, German, and Japanese*. Oxford University Press.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-ToM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706, Singapore. Association for Computational Linguistics.

Brett Wyatt and Suthinee Promkandorn. 2012. A discourse analysis of the Thai experience of "being krengjai". *Intercultural Pragmatics*, 9(3):361–383.

Appendix

A Personality in conversation between Thai name vs romanised Thai name

To testing how **GPT-4o-2024-08-06** generate natural thai conversation with thai character name are differ from the generated with romanised name in personality domain, we test by select only four famous thai name and fixed their role to be *Somchai(liar), Somchit(target), Sombun(accomplice)* and *Prasoet(observer)* and focus scenario at "white-lie **KJ0**", we then use **GPT-4o-mini-2024-07-18** to translate the whole conversation to english and use **ocean-personality-distilbert**[4] to predict their personality, we compare OCEAN/Big Five personality score between thai name and romanized name, we found that in average personality score of all scenarios see no different as Figure 7 but in (Figures 8, 9,10) have personality shift.

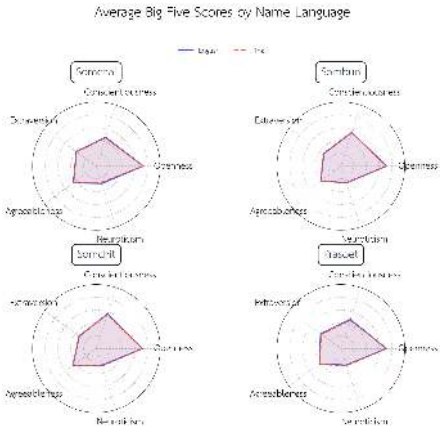


Figure 7: Average Big Five Score of prompting with thai letter vs romanized thai name across 43 scenarios

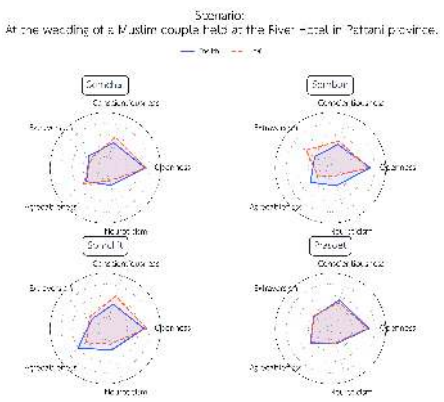


Figure 8: At the wedding in pattani, Somchai, Sombun and Somchit have personality shift.

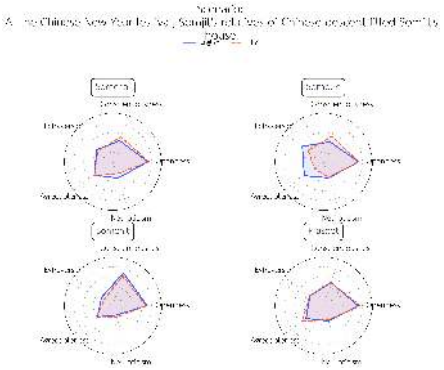


Figure 9: At Chinese New Year, Somchai and Sombun have personality shift.

[4] https://huggingface.co/ppp57420/ocean-personality-distilbert

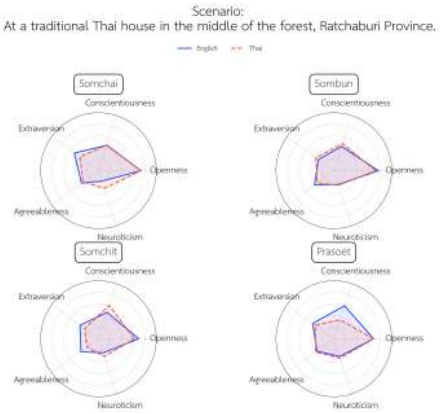


Figure 10: At traditional thai house, Somchai and Prasoet have personality shift.

B Annotation

B.1 Annotation Platform

We developed a web-based scenario generation platform using Streamlit to facilitate the systematic creation of Theory of Mind scenarios See in (Figures 11, Figure 12). The platform provides an intuitive interface for generating multi-party dialogues involving four characters (A: Liar, B: Target, C: Accomplice, D: Observer) across various social scenarios. The platform supports bilingual generation, allowing annotators to create scenarios in both English and Thai modes, with appropriate name and language-specific prompts for each mode. All generated scenarios follow a four-part structure: (1) initial scenario with all participants, (2) B and D leave while A and C discuss sensitive information, (3) D returns and learns the truth before leaving again, and (4) B returns and receives a lie from A and C.

The platform requires annotators to input comprehensive scenario parameters including metadata (set ID, lie ID, conversation ID), lie characteristics (type, emotion, falsification status), relationship dynamics (family, friends, colleagues), character roles, and the core deception elements (real reason, truth, and lie). Character names can be manually specified or randomly generated from culturally appropriate name databases. The system uses GPT-4o API to generate natural dialogues based on defined templates. Generated scenarios are automatically saved to JSON format with full metadata and can be previewed, downloaded, or loaded for further review through the platform's interface.

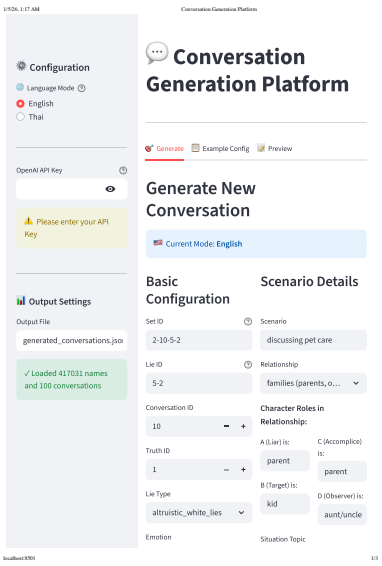


Figure 11: Annotation Platform in English mode

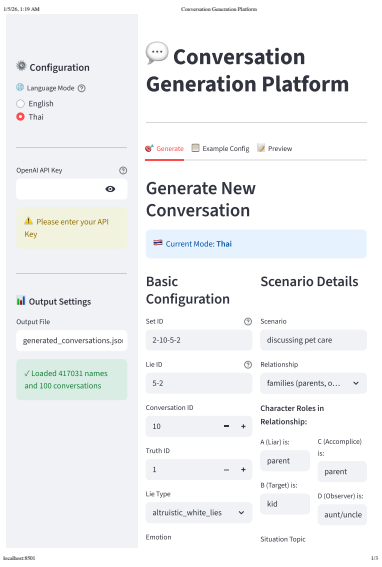


Figure 12: Annotation Platform in Thai mode

Question Type	Template
Comprehension(**CQ**)	"Was what {liar} told {target}, {lie}′, true?"
Justification(**JQ**)	"Why did {liar} and {accomplice} tell {target} {lie}′?"
Lie Ability(**LA**)	Please choose the correct answer from the options below: Correct: "{liar} tell {target} that {lie}′ because {liar} believes that {target} is unaware that {real reason}." Wrong 1: "{liar} tell {target} that {lie}′ because {liar} believes that {target} believes {real reason}." Wrong 2: "{liar} tell {target} that {lie}′ because {liar} is unaware that {target} is unaware that {real reason}." Wrong 3: "{liar} tell {target} that {lie}′ because {liar} is unaware that {target} believes {real reason}."
Lie Detectability(**LD**)	"Was what {liar} told {target}, {lie}′, true?" "List all the characters who know the precise correct answer to this question."
Fact(**FQA**)	"Who discussed {real reason}′?" "Who discussed {truth}′?"
Belief(**BQ**)	"What does character believe about {truth question} or {real reason question}?" "What does {liar} believe about {target}'s belief regarding {truth question} or {real reason question}′?"
Info_Accessibility(**IA**) (List)	"List all the characters who know this information: {real reason}."
Info_Accessibility(**IA**) (Binary)	"Does {character} know this information: {real reason}?"
Answerability(**A**) (List)	"Who discussed {real reason}′?" "List all the characters who know the precise correct answer to this question."
Answerability(**A**) (Binary)	"Who discussed {real reason}′?" "Does {character} know the precise correct answer to this question?"
KJability(**KJ**)	"what {character} believes about their own type of consideration" "what {character_target} believes about {character_liar}'s type of consideration"

Table 7: Question generation template is the same as TactfulToM.

Set 1-1-1-0-1 (Thai)		Set 1-1-42-1 (Thai)	
Attribute	**Value**	**Attribute**	**Value**
Language	Thai (th)	Language	Thai (th)
Lie Type	Avoid Confront	Lie Type	Avoid Confront
Emotion	Angry	Emotion	Angry
Liar	Sombun	Liar	Wichian
Target	Sawat	Target	Sombat
Accomplice	Prasoet	Accomplice	Somphon
Observer	Thawi	Observer	Sawat
Scenario	Wedding in Chiang Rai during Visakha Bucha	Scenario	Vegetarian festival(Jay Festival) charity kitchen
Situation Topic	Bragging about money envelope	Situation Topic	5 forbidden vegetables in vegetarian festival
Lie Objective	Save face and avoid confrontation	Lie Objective	Cover mistake, prevent food waste
Truth	Sawat gave empty envelope	Truth	Dropped fried garlic in pot
Lie	Agreed it's true	Lie	Said nothing was added
Real Reason	Sawat would be embarrassed	Real Reason	Sombat would be angry about waste
Entity Distribution			
Place	1.0	Place	0
Tradition	1.0	Tradition	1.0
Food	0	Food	2.0
Time	1.0	Time	0

Table 8: Comprehensive comparison of two seed scenario(Surface-cultural level vs Deep-cultural level) sets under KJ1 with character roles, key details, and cultural entities; (note1#: all possible cultural entity are PLACE, PERSON, ORGANISM, ARTIFACT/OBJECT, TRADITION/EVENT, ORGANIZATION, FOOD, TIME); (note2#: the original generative seed for thai context is the thai letter)

C Dataset

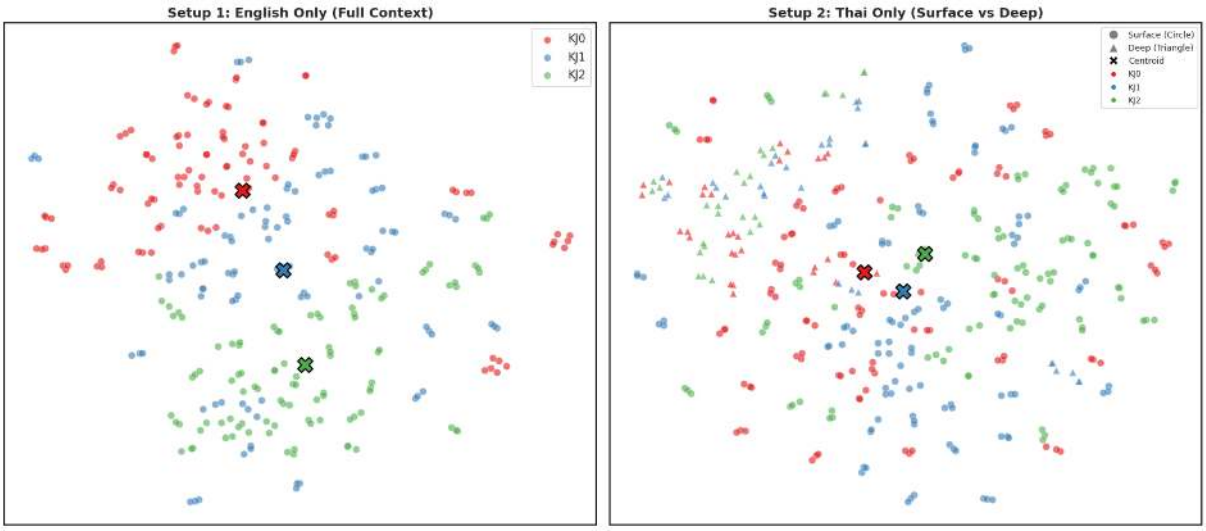


Figure 13: t-SNE visualization of conversation embeddings comparing English versus Thai scenarios of our KAI-ToM dataset using multilingual-e5-large-instruct (https://huggingface.co/intfloat/multilingual-e5-large-instruct). The left panel demonstrates well-separated clusters for Western English contexts across three Kreng-Jai subtypes **(KJ0, KJ1, KJ2)**, whereas the right panel shows less structured clustering for Thai contexts.

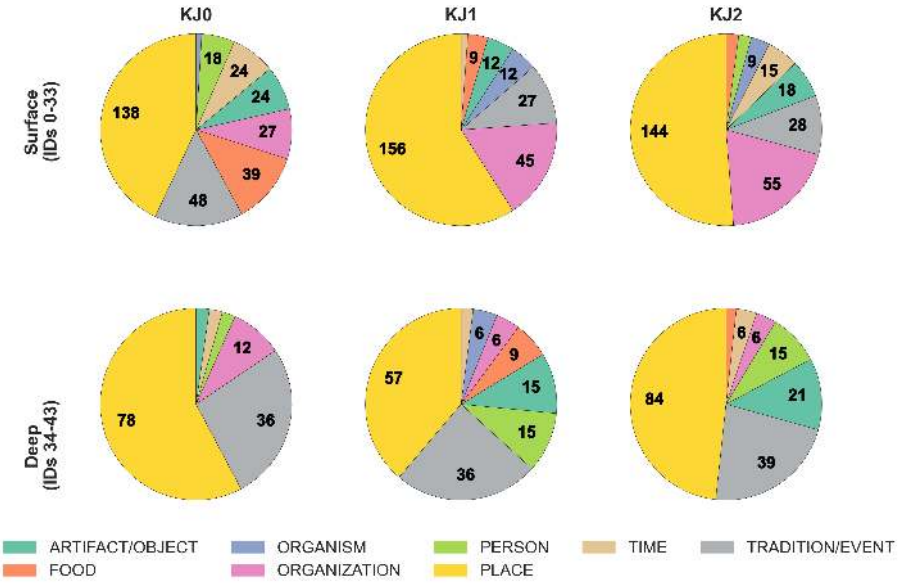


Figure 14: Cultural entity in each type of Kreng-Jai(KJ0,KJ1,KJ2)

D Human Recruitment

To be candidate annotators, there are two minimum requirements 1.be thai native and 2.english test score as Table 10

english test	score
IELTS	$\geq$ 5.0
TOEFL ibt	$\geq$ 60
TOEIC	$\geq$ 600
TU-GET cbt	$\geq$60
TU-GET pbt	$\geq$600
CU-TEP	$\geq$60
CEFR	$\geq$B1

Table 10: minimum requirements about english test score

D.1 Form

D.2 Form Question Types

Theory of Mind Questions (7 per scenario):

1. *JustificationQA(JQ)*
2. *Fact_TruthQA(FQA)*
3. *BeliefQAs(BQ)*
4. *AnswerabilityQA_list(A)*
5. *LieDetectabilityQAs_list(LD)*
6. *LieAbilityQAs(LA)*
7. *KJAbility(KJ)*

Opinion Questions (3-4 per scenario):

1. Does the character exhibit Kreng-Jai behavior?
2. What level of Kreng-Jai does the character exhibit?
3. What is the Thai cultural resonance level? (Thai scenarios only)
4. What is the participant's comprehension level after reading the entire conversation?

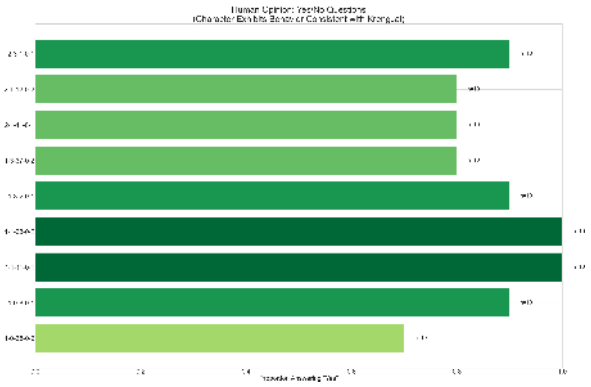


Figure 15: Result of opinion Yes/no question, question about does character in scenario exhibit Kreng-Jai behavior?; tester (n=9)

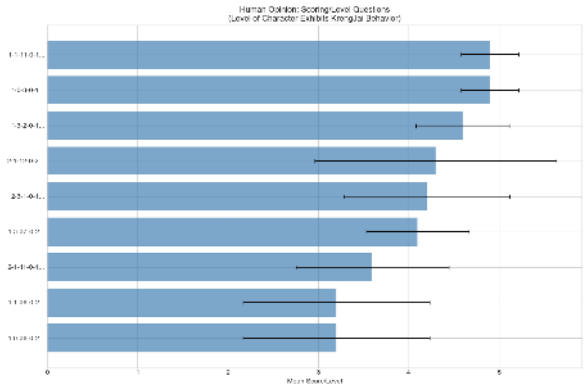


Figure 16: Result of opinion scoring question, question about does character in scenario exhibit Kreng-Jai behavior?; tester (n=9)

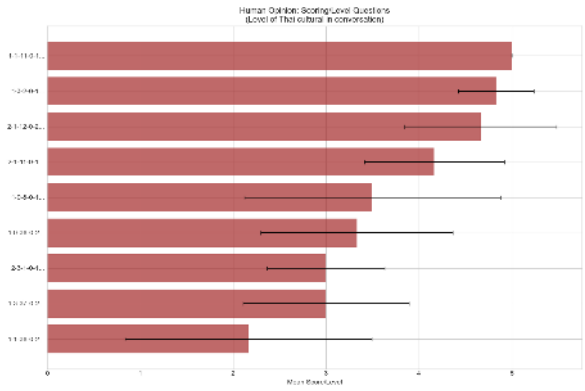


Figure 17: Result of opinion scoring question, cultural level; tester (n=9)

E Metrics

Performance is scored by question type:

Binary Questions

For binary questions, performance is evaluated using exact yes/no matching

Multiple Choice Questions (MCQ)

For MCQ questions, performance is binary based on whether the answer is correct

List Questions

For list questions, performance is evaluated using set comparison. Let C be the set of predicted items and A be the set of correct items. The evaluation computes:

$$I = C \cap A$$

$$\text{precision} = \frac{|I|}{|A|}, \quad \text{recall} = \frac{|I|}{|C|}$$

$$\text{performance} = \begin{cases} 1 & \text{if precision} = 1 \wedge \text{recall} = 1 \\ 0 & \text{otherwise} \end{cases}$$

F Inference Configuration

Parameter	Value
temperature	0.2
max_tokens	1000

Table 11: inference configuration for conversation generation

Model Used	Reasoning	Temperature	Max Tokens
gpt-5.1-2025-11-13	yes(low effort)	–	2000
gpt-4o-mini	no	0.2	8192
Kimi-K2-Instruct-0905	no	0.2	8192
DeepSeek-V3.2	no	0.2	8192
Gemini-3-flash-preview	no	0.2	8192
thai-llm-series	no	0.2	8192
SEA-LION-v3-9B	no	0.2	8192
Typhoon-v2.1-12b-instruc	no	0.2	8192
Qwen-series(except Qwen3-30B-A3B)	no	0.2	8192
Qwen3-30B-A3B	yes	0.2	8192
thai-llm-openthai	yes	0.2	8192
thai-llm-typhoon	no	0.2	8192
thai-llm-kbtg	no	0.2	8192

Table 12: inference configuration for getting the answer

G Additional Result

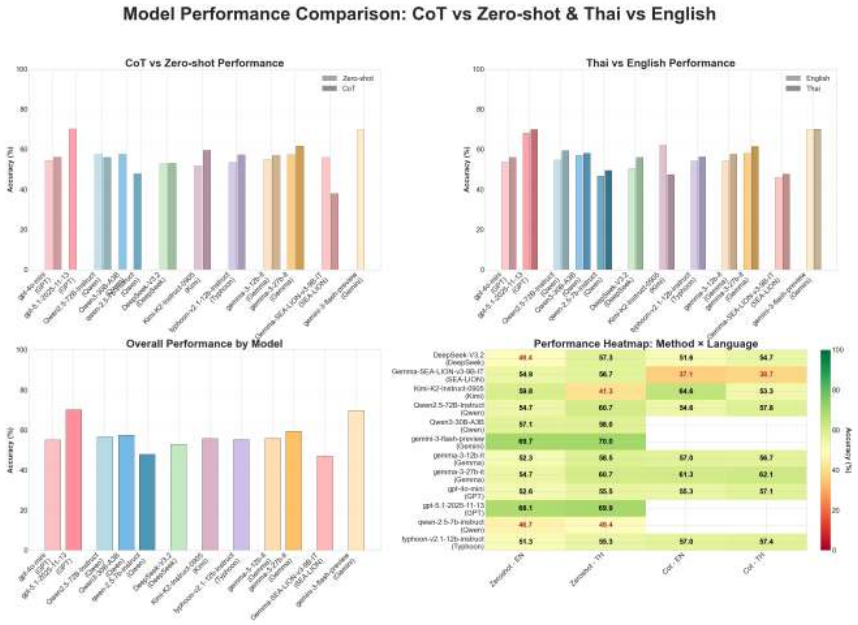


Figure 18: Result of summary; the top-leftimage is score comparing with zero-shot and Chain of Thought; the top-right image is score comparing with english vs thai scenarios data; the bottom-left is overall score of each model; the bottom-right is performance heatmap: method x language

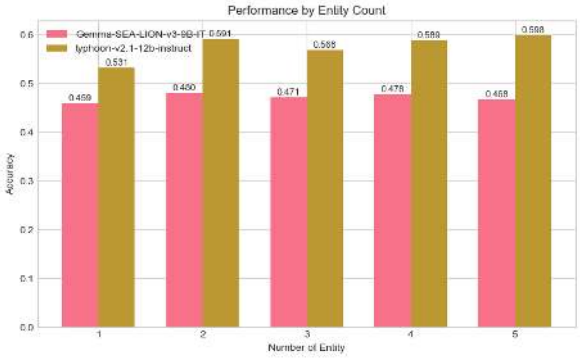


Figure 19: Results of Thai scenario data across Typhoon-v2.1-12B, Gemma Sea Lion-v3-9B, and the number of cultural entities in the generation seed.

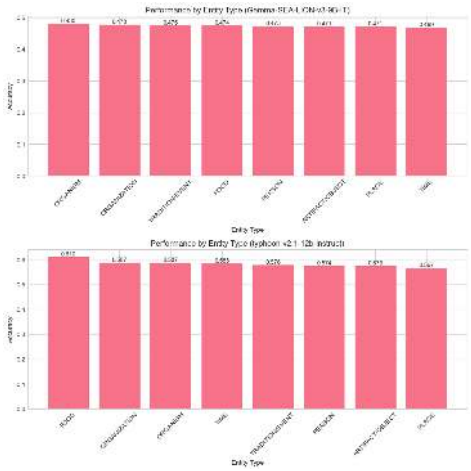


Figure 20: Results of Thai scenario data across Typhoon-v2.1-12B, Gemma Sea Lion-v3-9B, and different cultural entity types in the generation seed

Table 13: Zero-shot Thai-LLM series([5]https://huggingface.co/ThaiLLM/ThaiLLM-8B) compared GPT5.1, Typhoon-v2.1 and Gemma-SEA-LION with Results Across Models and Question types on Thai scenarios.

Model	Conv	CQ	FRA	FQA	BQ	IA.List	IA.Binary	A.List	A.Bianry	LA	LD.List	LD.Binary	KJ	JQ
GPT-5.1-2025-11-13	conv0(KJ0)	50.39	94.57	94.19	45.18	77.67	92.79	81.40	95.12	23.72	9.30	66.86	81.78	68.22
	conv1(KJ1)	43.41	96.12	83.72	43.74	80.47	92.09	81.40	95.47	36.74	3.88	69.96	41.47	52.71
	conv2(KJ2)	43.41	97.67	82.76	41.09	79.63	91.44	78.70	92.36	32.87	8.53	72.29	32.17	50.39
Thai-LLM-KBTG (Labs et al., 2026)	conv0(KJ0)	15.50	72.09	59.30	27.80	0.00	75.23	0.00	78.95	28.37	0.00	59.50	42.25	22.48
	conv1(KJ1)	24.81	73.64	40.70	29.35	0.00	77.67	0.00	75.70	31.63	0.00	59.88	37.60	31.01
	conv2(KJ2)	22.48	75.19	45.98	28.13	0.00	70.14	0.00	73.73	23.61	0.00	59.88	39.15	28.68
Thai-LLM-OpenThai (OpenThaiGPT and Team, 2025)	conv0(KJ0)	19.38	27.91	18.60	28.13	0.00	73.60	0.00	70.23	46.98	0.00	54.65	31.01	60.47
	conv1(KJ1)	19.38	32.56	27.91	27.80	0.00	72.09	0.00	66.98	38.60	0.00	61.05	36.82	54.26
	conv2(KJ2)	15.50	31.01	17.24	24.58	0.00	69.56	0.00	64.81	35.19	0.00	61.63	33.33	45.74
Thai-LLM-Typhoon (Pipatanakul and Taveekitworachai, 2026)	conv0(KJ0)	8.53	58.91	51.16	14.51	0.00	67.67	9.77	66.28	35.81	0.00	58.53	42.25	44.19
	conv1(KJ1)	15.50	62.79	40.70	16.61	0.00	71.28	10.23	65.70	28.37	0.00	61.24	24.03	52.71
	conv2(KJ2)	13.18	62.79	55.17	14.51	0.00	65.16	15.28	64.00	31.02	0.00	57.75	70.54	35.66
Typhoon-v2.1-12b-Instruct	conv0(KJ0)	20.16	100.00	46.51	26.80	19.07	80.93	67.91	81.16	33.49	44.96	65.70	80.62	49.61
	conv1(KJ1)	23.26	96.12	31.40	26.80	21.86	80.12	70.23	77.91	36.28	33.33	63.37	9.30	45.74
	conv2(KJ2)	14.73	94.57	43.68	25.36	19.91	78.47	69.91	73.84	29.17	10.85	61.63	39.53	44.96
Gemma-SEA-LION-v3-9B-IT	conv0(KJ0)	24.81	97.67	51.16	28.13	76.28	78.84	80.93	67.91	44.19	29.46	54.65	69.77	51.16
	conv1(KJ1)	31.01	97.67	41.86	26.47	80.00	80.00	81.40	67.56	43.26	13.18	52.33	37.21	57.36
	conv2(KJ2)	21.71	98.45	45.98	23.70	74.07	76.74	80.56	66.67	42.13	1.55	51.16	46.90	53.49

H Prompts

H.1 prompts for define the linguistic characteristics of conversation.

System prompt thai scenario

You are a creative writer generating natural, realistic conversations in Thai language. Write dialogue in a clear format with character names followed by their speech. Keep the names in English but the conversation in Thai.Use natural Thai conversation style.

System prompt for english scenario

You are a creative writer generating natural, realistic conversations. Write dialogue in a clear format with character names followed by their speech.

H.2 prompts for conversation generation from TactfulToM framework

Conversation generation prompt: Step1

In this conversation, four characters are participating: **A_name** , **B_name** , **C_name** , and **D_name**. Start the conversation with all participants **scenario**. Relationship **relationship_text**. Generate a natural conversation between all four characters. Use their names and format the dialogue clearly.

Conversation generation prompt: Step2

Now, **B_name** leaves the conversation because of the reason: **leave_reason_B**, while **D_name** leaves the conversation because of the reason: **leave_reason_D_1**. With **B_name** and **D_name** absent, **A_name** and **C_name** continue their conversation, shifting the topic to **'real_reason_c_filled'**. Ensure the dialogue flows naturally without directly repeating these exact words. Remember to indicate that **B_name** and **D_name** are leaving the conversation.

Conversation generation prompt: Step3

Now, only **D_name** returns and rejoins the conversation after previously leaving. Note that **B_name** has not returned to the conversation yet. Then the conversation continues and naturally shifts to **situation_topic**. Do NOT catch up or recap details from the previous conversation. And then **'truth_c_filled'** is naturally revealed in the conversation. After engaging briefly, **D_name** leaves the conversation again because of the reason: **leave_reason_D_2**.
IMPORTANT:
- Clearly indicate through dialogue only (e.g., "I'm back," or "I have to leave again now.") that **D_name** first returns and later departs again.
- Do NOT use third-person narrative or stage directions.
- Avoid mentioning details from previous conversations.

Conversation generation prompt: Step4

Now **B_name** returns to the conversation after leaving the conversation. First have **B_name** explicitly indicate the return through dialogue. Do NOT catch up or recap details from the previous conversation. And then **situation** naturally unfolds. Make sure the dialogue flows naturally without directly repeating these exact words. In response, **A_name** and **C_name** tells **B_name** that **'lie_c_filled'** to **lie_objective**.
IMPORTANT:
- Do NOT use third-person narrative or stage directions.
- Avoid mentioning details from previous conversations.

H.3 prompts for question generation

justificationQA generation

You are generating a justification question for a white-lie scenario. CONTEXT: - Liar: **liar** - Target: **target** - Accomplice: **accomplice** - What **liar** said (lie): **l_c** - Real reason for lying: **real_reason_c** - Truth: **truth_c**
Generate a justification question asking why **liar** told **target** the lie.
Return ONLY a JSON object with this exact structure: {{ "question": "Why did **liar** tell **target** '≥shortened version of the lie≤'?", "correct_answer": "≥The genuine psychological reason - typically about sparing feelings, maintaining harmony, or avoiding awkwardness. Max 30 words.≤", "wrong_answers": ["≥Type-1: Literal Reason - restates the spoken lie as if it were true≤", "≥Type-2: Negative Feeling - blames dislike, annoyance, frustration, etc.≤", "≥Type-3: Random Excuse - plausible but unrelated reason involving **accomplice** or other circumstances≤"] }}
RULES: 1. The question should be clear and concise
2. correct_answer must reflect the real_reason_c (sparing feelings, maintaining harmony)
3. Each wrong answer must be 30 words
4. Wrong answers must be distinct and follow their type descriptions
5. Use third person perspective
6. Return ONLY valid JSON, no markdown wrappers

H.4 prompts for an LLM to answer the question

Answer Type	Prompt Template
Binary	You are an expert in social reasoning. Answer the following question with 'Yes' or 'No'. Remember: Your answer should ONLY include 'Yes' or 'No' with nothing else. # Context: {} # Question: {} (Let's think step by step:)
MCQs	You are an expert in social reasoning. Answer the following question with the option number of the most appropriate answer. Remember: Your answer should ONLY include the option number with nothing else. # Context: {} # Question: {} # Options: {} (Let's think step by step:)
List	You are an expert in social reasoning. List the required items and split them with commas. Remember: Your answer should ONLY include the required items split by commas with nothing else. # Context: {} # Question: {} (Let's think step by step:)

Table 14: Prompt templates for different answer types; the "Let's think step by step" was added when to test CoT.