Universiteit Leiden
The Netherlands

Opleiding Informatica

Cultural prompting and

Theory of Mind in LLMs

Lana van Sprang
s3272192

Supervisors:
Max van Duijn & Sabijn Perdijk

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

16/07/2025

Abstract

The cognitive ability to understand the mental states of yourself and others is defined as Theory of Mind (ToM). To assess the abilities of Large Language Models (LLMs), ToM tasks are used. LLMs are biased and are affected by language. In psychology, people across different cultures learn ToM at different rates and perform differently if primed for certain cultural aspects. Moreover, prior research suggests that priming an LLM for a certain culture could affect the responses given by an LLM. With this suggestion, the question arises whether LLMs are affected in their ToM performance when culturally prompted. Hence, the goal of this study is to assess whether cultural prompting in LLMs affects their performance on ToM tasks. In this study, we chose four cultures based on the linguistic distance to English and Hofstede's cultural dimensions. Based on these cultures, LLMs are culturally prompted and perform ToM tasks in the respective language they are primed with.

In this research, we found no significant differences between the culturally prompted and the normal ToM performance, and we cannot conclude that cultural prompting affects the performance of ToM tasks in LLMs.

Contents

1 Introduction

Do Large Language Models (LLMs) understand us? Do they think like us, or do they share similar abilities? With the popularity of ChatGPT [Ope25], these questions arise. We want to focus on the social-cognitive abilities of LLMs. A method to assess whether LLMs have mental abilities is the use of Theory of Mind (ToM) tasks. Theory of Mind is the cognitive ability to understand the mental states of yourself and others [App10].

In psychology, ToM is tested in children to see how they develop the understanding that other people may have other beliefs and mental states than they do [App10]. In previous research, it has been studied how children learn ToM, or the absence of it in autistic children [App10].

Having insight into the social-cognitive abilities of LLMs is helpful in fine-tuning them or defining their limitations. Furthermore, having insight on LLM's ToM understanding may help with communicating and collaborating with them. LLMs have the ability to correctly answer ToM questions and show human-level performance on ToM tasks [vDvDK+23]. The performance of LLMs on such tasks varies depending on the task and LLM: some LLMs outperform young children [vDvDK+23], and some struggle with ToM [CWZ+24]. Moreover, previous research shows how LLM performance in ToM tasks is affected [SKN+24, TYJ+24].

However, a downside to modern LLMs is something all human-creations are sensitive to: bias. Western LLMs, such as ChatGPT, are biased towards Western cultures and values [TVBK24]. This is caused by the training data and the origin of LLMs [TVBK24]. LLMs perform better when their input is in English than in any other language. Western culture is the baseline of LLMs, even when trying to minimise bias. That is why previous research has been done on the influence of culture and language [SKN+24, LCW+24]. How an LLM performs depends on the language of the input and the context.

Now, there exists a gap in the research. Although there is research on the influence of languages and cultural context in ToM questions [SKN+24], there is no previous research on priming an LLM for a certain culture before answering ToM questions. This is called cultural prompting. Several studies suggest that cultural prompting may be a method to minimise cultural bias [TVBK24, WJH+24]. However, the main goal of this study is not to minimise cultural bias but to research the influence of cultural prompting in the social-cognitive domain. Understanding this could mitigate cultural bias in future LLMs. This research aims to answer the following research question: **Does cultural prompting in LLMs affect performance of Theory of Mind tasks?**

We expect to find a difference in ToM task performance when LLMs are culturally prompted compared to when they are not, due to language and culture bias. Current LLMs perform better with English or Chinese input compared to any other language [TVBK24]. If LLMs are culturally prompted, it forces them to access the data on the relevant culture and language, which is less compared to all data. Because the LLM is forced to access these parts, cultural prompting could have either a positive or negative effect on ToM task performance.

In this study, the performance of ToM tasks in LLMs is researched using cultural prompting with the benchmark provided by ToMBench [CWZ+24] in multiple LLMs. We chose four cultures based on Hofstede's cultural dimensions [Hof01] and the linguistic distance of the language of the respective culture compared to English [CM05].

2 Related Work

2.1 Theory of Mind in psychology

Much research on ToM has been done which indicates that ToM is influenced by culture. In the study [WFP11], the authors concluded that Chinese children learn different ToM components first compared to American children and that the children learn at different rates. In another study [TVFH+17], the authors researched how collectivism and individualism influence ToM. Students from the Netherlands, an individualism-representative country, and from Vietnam, a collectivism-representative country, were randomly assigned to a collectivism-primed, individualism-primed, or no-primed task before they performed ToM tasks. The researchers found that the Vietnamese participants performed ToM tasks more accurately if they were collectivism-primed, and if they were individualism-primed they performed worse compared to no-primed. The Vietnamese participants were also slower in reaction time than the Dutch. The researchers suggested that the reason for this difference in reaction time could be cultural or due to experience [TVFH+17]. For the Dutch participants, it did not matter which type of priming was used and which ToM task was performed.

2.2 Hofstede's cultural dimensions

Each country has its own culture, and with Hofstede's cultural scale, each culture can be described through six dimensions: power distance, uncertainty avoidance, individualism vs. collectivism, motivation towards achievement and success [CFF+24], long vs. short-term orientation [Hof01], and the most recent dimension: indulgence vs. restraint [HHM10].

Dimension 1: Power distance According to Hofstede, power distance refers to the extent people low in the hierarchy in organisations or institutions accept and expect unequally distributed power [Hof01]. For instance, if you are low on the social ladder in a student association, then you expect and accept that you do not have the same privileges as someone that is higher in the hierarchy than you. In this example, there is a high power distance.

Dimension 2: Uncertainty avoidance Uncertainty avoidance refers to how tolerant a society is with ambiguity and uncertainty [Hof01]. The more fixed rules and habits there are in a society, and the more a society is anxious about the unknown, the more uncertainty avoidant that society is. A uncertainty-avoidant society aims to know the truth at all times, instead of leaving something unknown and ambiguous. Meanwhile, uncertainty-accepting societies tolerate the uncertainty and ambiguity and have fewer fixed rules and habits.

Dimension 3: Individualism The individualism vs. collectivism dimension refers to what extent the people in a society feel independent [Hof01]. In an individualistic society, individual choices and decisions are expected. In a collectivistic society, people feel interdependent as members of a larger whole. Individual decisions are less important than the decisions of the community. The people in individualistic societies associate themselves with their independence. They also focus more on separate details than the whole picture. People in collectivistic societies associate themselves with the group they are a part of, and they look at the whole picture before they notice all the separate details.

Dimension 4: Motivation towards achievement and success The motivation towards achievement and success dimension, previously called the masculinity dimension in [Hof01], refers to the extent a society values achievements and success [CFF+24]. A high achievement motivation society expects their people to be tough. Moreover, winning is very important, and people in these societies value quantity above quality. As opposed to a low achievement motivation society, in which competition is not so endorsed, and winning is not the most important goal.

Dimension 5: Time orientation The long-term vs. short-term orientation refers to how a society deals with change [Hof01]. In a long-term-orientated society, their focus is on change. A long-term-orientated society is always preparing for the future. In a short-term-orientated society, their focus is on the state of the current world. A short-term-orientated society uses the past as a moral compass since they view the current world to be the same as the world of the past. They use the past as a guide to live a good life. However, in a long-term-orientated society, using the past as a moral compass or as a guide is not needed since they view the past world as different.

Dimension 6: Indulgence The indulgence dimension refers to a society's view on the focus of life [HHM10]. In an indulgent society, it is good to experience freedom and give in to your indulgences. Making friends is important, and life makes sense to people in an indulgent society. In a restrained society, duty is the most important thing. People in a restrained society experience a feeling that life is hard.

2.3 Linguistic distance

Linguistic distance refers to the distance between one language and another [CM05]. In this study, we will define it as the distance between the English language and other non-native English languages, as in the paper [CM05]. This research provides a methodology for measuring linguistic distance and also provides a language score for multiple non-English languages based on the approximate time needed to learn English [oLS93]. The language score ranges from 3.00 (easy to learn) to 1.25 (hard to learn) [CM05].

2.4 Theory of Mind, culture, and performance in LLMs

The presence of bias in LLMs has been researched and proven; LLMs pick up cultural values from their training data, with some values weaker than others [AKA23]. In the study [TVBK24], the researchers found that most of the LLM values align more with the Western values. This is caused by the training data of the LLMs, which are mostly Western and Chinese, and explains the bias in LLMs. Moreover, the authors found that ChatGPT has a gender bias due to the training data.
Additionally, the authors of the study [HRS+23] found that LLMs are not multicultural; even when prompted with other languages, most LLMs still align more with Western values. The authors conclude that multilingual LLMs do not properly learn cultural nuances, specifically in emotions. As mentioned in their research, an LLM prompted in Japanese behaves as an American who is fluent in the language but is unaware of the culture [HRS+23]. This finding is also found in the study [WJH+24].
To mitigate the cultural bias problem, the study [LCW+24] created a cultural LLM to combat this issue. In the study [SKN+24], the researchers introduce a multilingual dataset with cultural context

in their ToM tasks. They found that LLMs are less accurate when cultural nuances are introduced in the tasks.
Another way to mitigate the problem of cultural bias is by using cultural prompting, as suggested in previous studies [WJH+24, TVBK24].
The use of cultural prompts is not a new concept. One study used cultural prompting to research the underlying cultural background of ChatGPT [CZL+23]. The authors found that it had a strong alignment with American culture when culturally prompted with American context, more so than when prompted with other cultures. An important finding of their research is that prompting with English reduces variance in the models' responses and reduces the bias towards American culture.
Now, ToM has been a topic of interest with respect to LLMs. Previous research shows that LLMs can perform ToM tasks [vDvDK+23, TYJ+24, CWZ+24]. The newer ChatGPT models are even performing better than children between the ages of 7 to 10 [vDvDK+23].
In the study *ToMBench: Benchmarking Theory of Mind in Large Language Models* [CWZ+24], the researchers provide their own dataset to extensively test LLMs in ToM tasks in Chinese and English. The authors also found that Chain of Thought prompting did not change the LLM's accuracy score, which contradicts earlier work [MH23].
In the study [TYJ+24], the researchers found that persona-based prompting affects an LLM's performance on ToM tasks.

3 Methods

3.1 Cultures and Cultural prompting

As shown in Section 2, LLMs have a bias towards Western culture [TVBK24] and outperform other languages using English [SKN+24]. Moreover, LLMs give different outputs for each language for the same input [LCW+24]. To ensure that LLMs use the cultural values in the prompts, the benchmark must be in the language of the respective culture. Due to these factors, we chose to translate the benchmark into the main language of the chosen cultures. In this research, we chose the following four cultures for cultural prompting: French, German, Russian, and Thai.

Culture	Linguistic distance	Dimensions					
		PD	UA	IV	M	TO	IG
French	2.50	68	86	71	43	63	48
German	2.25	35	65	67	66	83	40
Russian	2.25	93	95	39	36	81	20
Thai	2.00	64	64	20	34	32	45

Table 1: The scores of the chosen cultures for the evaluation. Here, PD is the power distance, UA is uncertainty avoidance, IV is individualism, M is motivation towards achievement and success, TO is the time-orientation and IG is the indulgence dimension.

To evaluate the cultures and the respective languages, we use Hofstede's cultural scale [Hof01, HHM10] and linguistic distance [CM05, oLS93]. The linguistic distance has to be approximately the same for each culture, and each culture has to score differently on the cultural scales. The

scores of the evaluations are shown in Table 1. Each culture had approximately a linguistic distance of 2.25, with a range between 2.00 and 2.50.
In the cultural prompts, each culture is described by their scores on Hofstede's dimensions [Hof01, HHM10]. The cultures do not have to differ from each other in all dimensions, but should be distinguishable. For example, both Russian and Thai are collectivistic cultures, while French and German are individualistic. However, Russian and Thai are distinguishable from each other in the time-orientation dimension, and French and German are distinguishable from each other in the power distance dimension.
A cultural prompt is the combination of the 'vanilla prompt' provided in ToMBench [CWZ+24], and descriptions based on their cultural scores. As an example, the French cultural prompt is given below as a text box:

Cultural Prompt for French Evaluation

You are French, born and raised in France and still living in Paris. You are a very individualistic person. You value individualism and freedom above all else. You feel independent. You are also moderately long-time orientated. You prepare for the future, and you are less inclined to use the past as a moral compass. You accept that there is a hierarchy in your culture. There is a power distance between you and your boss and the people above you, and you mostly expect and accept this unequal distribution of power. You do not expect men to be tough. You do not openly endorse competing, and you have sympathy for the underdog. You are not openly gendered, and men and women in your culture are somewhat emotionally close. You are very anxious and distrustful of the unknown. You are very intolerant of uncertainty and ambiguity. You prefer fixed habits, rituals, and knowing the truth most of the time. You are ambivalent about indulgence. Sometimes, you believe in duty and that life is hard. Other times, you believe life is meaningful, friends are important, and it's good to follow your impulses.
Since you are French, all questions will be presented in your language. Please answer from this cultural perspective.

`[Vanilla Prompt added.]`

We chose to culturally prompt in English to minimise variance [CZL+23]. The normal prompt, or no-cultural prompt, contains only the 'vanilla' prompt. Below is the text box of the 'vanilla' prompt:

Vanilla Prompt

Below is a multiple-choice question with a story and several answer options. Based on the content of the story and the given question, please infer the most likely answer and output the answer index.
Note: (1) Please only output the most likely answer index in the format: `[[Answer Index]]`, for example, if the most likely answer option is 'A. Handbag', then output `'[[A]]'`;
(2) You must choose one of the given answer options 'A, B, C, D' as the most likely answer, regardless of whether the story provides enough information. If you think there is not enough information in the story to choose an answer, please randomly output one of `"[[A]]", "[[B]]", "[[C]]",` or `"[[D]]"`;
(3) Please only output the most likely answer index based on the given information, and do not output any other content.

The cultural prompts of the other cultures are found in Section B.

3.2 Benchmark

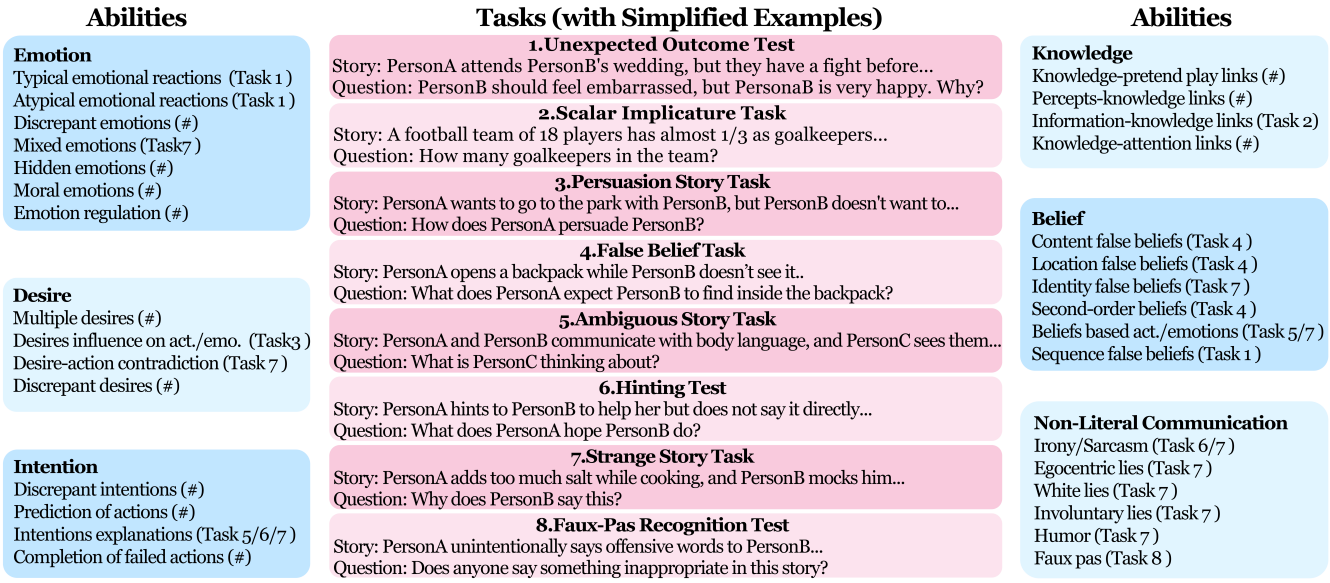

Figure 1: Figure reproduced from the research of ToMBench [CWZ+24]. The figure shows the full ToMBench inventory of abilities and tasks. The index next to the ability represents which task or file the ability is tested in. The '#' index represents a separate file for ability testing.

To test the performance of ToM tasks in LLMs, we use the ToMBench benchmark [CWZ+24]. The benchmark and its code to run the benchmark are obtained from the official ToMBench repository on GitHub [CWZ+24].
ToMBench is a Chinese benchmark that is specifically made for testing ToM in LLMs. Although the benchmark is originally in Chinese, they also provide English translations. The benchmark contains 20 `JSONL` files with ToM questions. Eight files represent the classic ToM tasks: the *'Unexpected Outcome'* test, the *'Ambiguous Story'* task, the *'False Belief'* task, the *'Faux Pas Recognition'* test,

the *'Persuasion Story'* task, the *'Scalar Implicature'* test, the *'Hinting Task'* test and the *'Strange Story'* task. The other files are meant to assess 31 ToM abilities, as well as the task files. All 31 abilities fall under the following 6 root abilities: *'Intention'*, *'Belief'*, *'Desire'*, *'Emotion'*, *'Non-literal Communication'*, and *'Knowledge'*. See Figure 1, reproduced from the paper of ToMBench [CWZ+24], for an overview of the ToM tasks and abilities.
The performance of these tasks and abilities are assessed using stories. Each story describes a fictional situation, and to test the LLM's perspective and understanding of these stories, questions are asked. These are either multiple-choice or true-false questions. The benchmark also provides the correct answers to these questions and uses accuracy as a metric [CWZ+24].
ToMBench provided two prompting methods: 'vanilla' and Chain of Thought (CoT) [CWZ+24]. With the 'vanilla' method, the LLM is asked to answer the right choice directly. The CoT method forces the LLM to provide step-by-step reasoning before answering the correct option. However, in this study, we focus solely on the 'vanilla' prompting method, as the authors of ToMBench found that there was no significant difference between the accuracy of the LLMs using CoT and using 'vanilla' with their benchmark [CWZ+24].
Additionally, the authors of ToMBench [CWZ+24] ran each question 5 times per LLM and selected the most frequently answered choice as the LLM's final answer to the question. The order of the multiple-choice is randomly shuffled each time to reduce bias. In this study, we chose to run the benchmark once, as ToMBench has an extensive number of ToM questions. If an LLM answers a question randomly or due to bias, these answers will be averaged out by the sheer number of questions.
In this research, we used the English version of the benchmark, and we compared our results with the English results of ToMBench.

3.3 Translating agent

To translate the benchmark, we created a translating agent. This agent is based on the translating agent in the study [SKN+24]. Our translating agent consists of three agents: the first agent prunes the dataset to English fields only and translates only the specified English texts in ToMBench [CWZ+24] to the target language. The second agent reviews the translation of the first agent with the original file. The third agent refines the given translation based on the original input, the translated input, and the feedback given by the second agent. Additionally, it prunes the review columns of the second agent. Now, the full translating agent gives the translated `JSONL` files in the target language based on the original texts of ToMBench. Each agent is created with GPT-3.5-turbo-0125 [Ope23]. However, while running the first translations with ChatGPT-3.5-turbo, we found that it often silently fails when translating non-Western languages, such as Thai and Russian. To combat this, ChatGPT-4o-2024-08-06 [OAA+24] is used for the retries in translating agent 1. The pipeline and the exact prompts of the translating agents are found in Section A.

3.4 Testing

We run the translating agent once for each culture and for each file. Then, the translated benchmark is run once for each prompting type: cultural and normal. These prompts are provided in the system messages of the LLMs. All code is run with Python version 3.10.12. We test with various LLMs, corresponding to the LLMs used in ToMBench [CWZ+24]. These LLMs are: GPT-4-0613

[OAA+24], Mistral-7B-v0.3 [AI23], Llama-2-13B-Chat [TMS+23], and Qwen-14B-Chat [BBC+23]. Due to time and cost constraints, GPT-4 is only tested in French.

3.5 Statistical tests

To test the significance of cultural prompting, we use the linear mixed-effects model [PB00] due to the randomness of tasks and the hierarchical structure of the data. The model predicts accuracy based on the prompting type, the language, and their interaction. The ToM tasks and abilities are selected as the random intercept. The following model assumptions are assessed: the residuals follow a normal distribution and the residuals are independent. These are assessed using model plots, such as a histogram, a Q-Q plot, and a residual vs. fitted plot. To test whether the use of random intercepts is justified, the likelihood ratio test is used to compare the mixed-effects model with a simple model. If the model meets all assumptions, the robustness is tested by refitting the model after removing 10 potential outliers with high accuracy. Finally, the final model will be compared with the model that predicts accuracy based only on language. Here, we calculate the p-value of the cultural prompting effect using the likelihood ratio test. For further analysis, the effect size of cultural prompting is calculated by fitting a scaled model such that the coefficients are standardised. These coefficients simulate the Cohen's d [Coh88] or effect size. Additionally, the model will also be used on a task-only dataset and an ability-only set.
Although not a statistical test, the coherent test is used as in ToMBench [CWZ+24]. The coherent test shows whether LLMs understand a story or if they are making an educated guess. If all questions in a story are answered correctly, then the LLM shows a complete understanding of that story [KSZ+23]. The coherent test counts all correctly answered stories in the tasks. If one question in a story is answered incorrectly, then the coherent accuracy of that story is 0%. The average coherent score in a task will drop significantly if an LLM takes multiple educated guesses. Vice versa, the average coherent score will be approximately the same as the original score if an LLM has a complete understanding of the stories.

4 Results

4.1 General prompting effect

The overall performance among various LLMs is shown in Figure 2. This average is calculated over all languages (French, German, Russian, and Thai) and over all 8 ToM tasks and the 6 main abilities. The results show that the translated benchmark performed around 10% worse compared to the original English benchmark from ToMBench [CWZ+24], excluding GPT-4. Moreover, the results from the translated benchmark had a smaller standard deviation than ToMBench. Within the LLMs, Qwen performed the best and Llama2 the worst. As stated in Section 3.4, GPT-4 was run only in French and not with all languages. For a fair comparison, the GPT-4 results are excluded from the main comparisons. Comparing only the French results, GPT-4 achieved the highest performance for both prompting types. However, the difference between cultural and normal prompting is consistent with the other LLMs. Furthermore, the French results of GPT-4 are consistent with the English GPT-4 results from ToMBench [CWZ+24]. For precise accuracy scores for each ToM task and ability, see Section C.

Turning to the statistical analysis, we tested the assumptions of the mixed-effects model and found that it met all the assumptions. For further details of the results of the assumptions, see Section D. We tested the justification of the model using the likelihood ratio test on the mixed model and the simple model, which yielded a $p < 0.05$. The use of the mixed-effects model with random intercepts is justified. To test the robustness, we refitted the model without potential outliers, and the results did not change. For detailed output of the model, see Section E.

The likelihood ratio test showed that the effect of cultural prompting alone had a $p = 0.767$. Furthermore, the model predicted that the use of the German, Russian, and Thai languages results in lower accuracy. Moreover, no significant interaction was found between the language and cultural prompting for each language.

The average performance between cultures is shown in Figure 3. In general, French performed the best with an average accuracy of 50%, and Thai the worst with an average accuracy of around 30%. In addition, French and Thai performed worse with cultural prompting compared to normal prompting, and German and Russian performed slightly better. The mixed-effects model revealed a $p < 0.001$ for the effect of language on ToM tasks and abilities compared to French. The effect size of cultural prompting was < 0.1 for all languages.

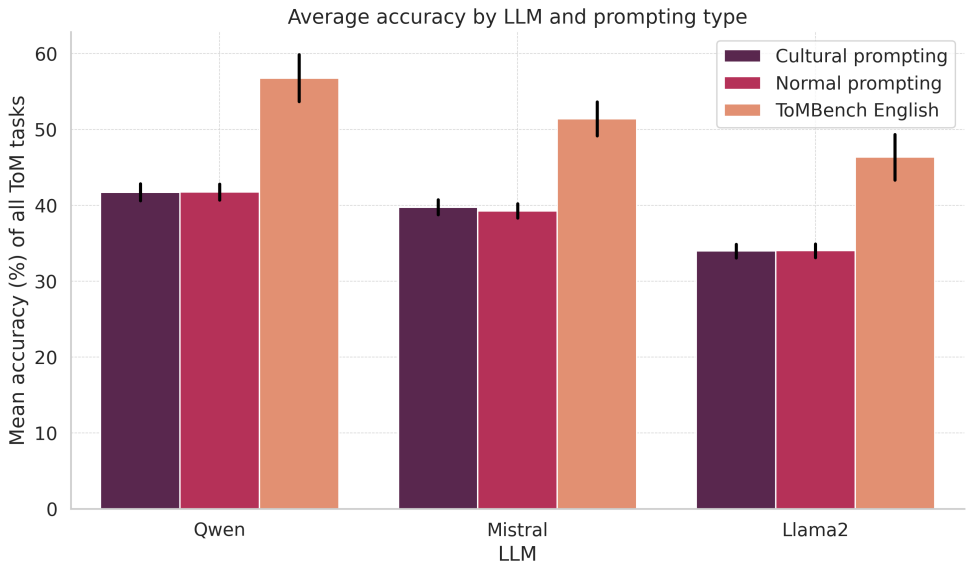

Figure 2: The average accuracy of the 8 ToM tasks and 6 main abilities per LLM compared with the results of ToMBench [CWZ+24]. For a fair comparison, the French GPT-4 results are excluded.

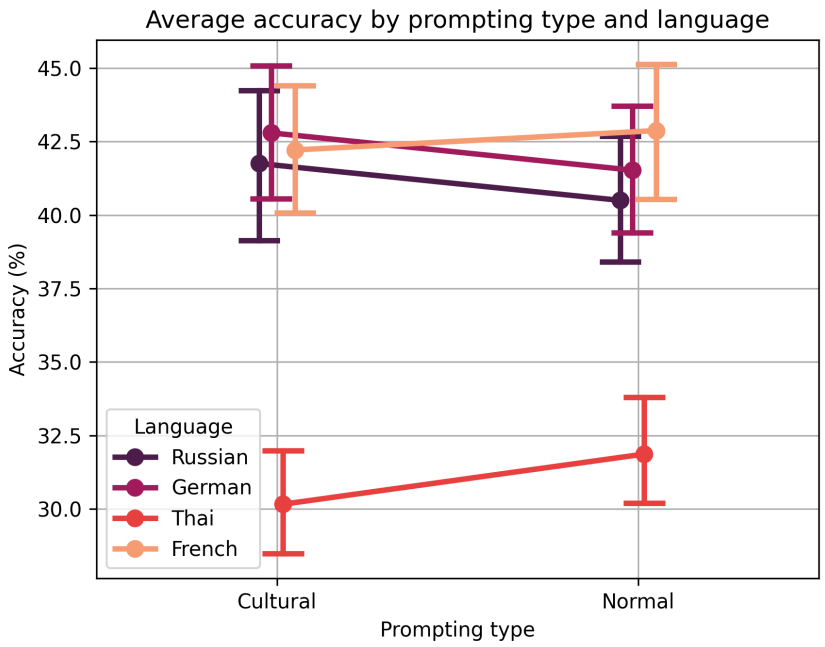

Figure 3: The average accuracy of ToM tasks and abilities for each prompting type across languages. For a fair comparison, the French GPT-4 results are excluded.

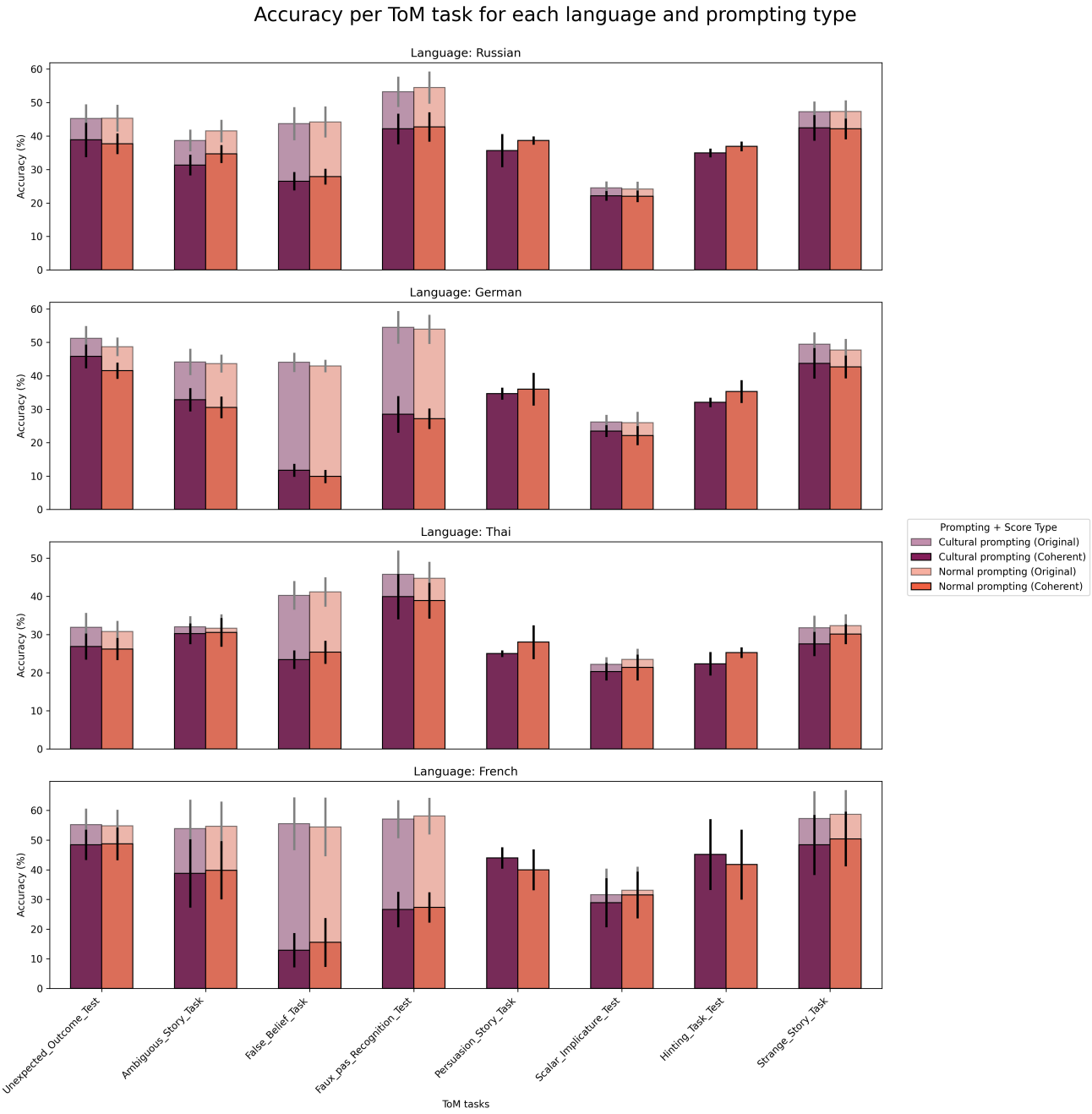

Figure 4: The average accuracy per ToM task for normal prompting and cultural prompting per language, including the standard deviation. This figure also shows the results of the coherent test.

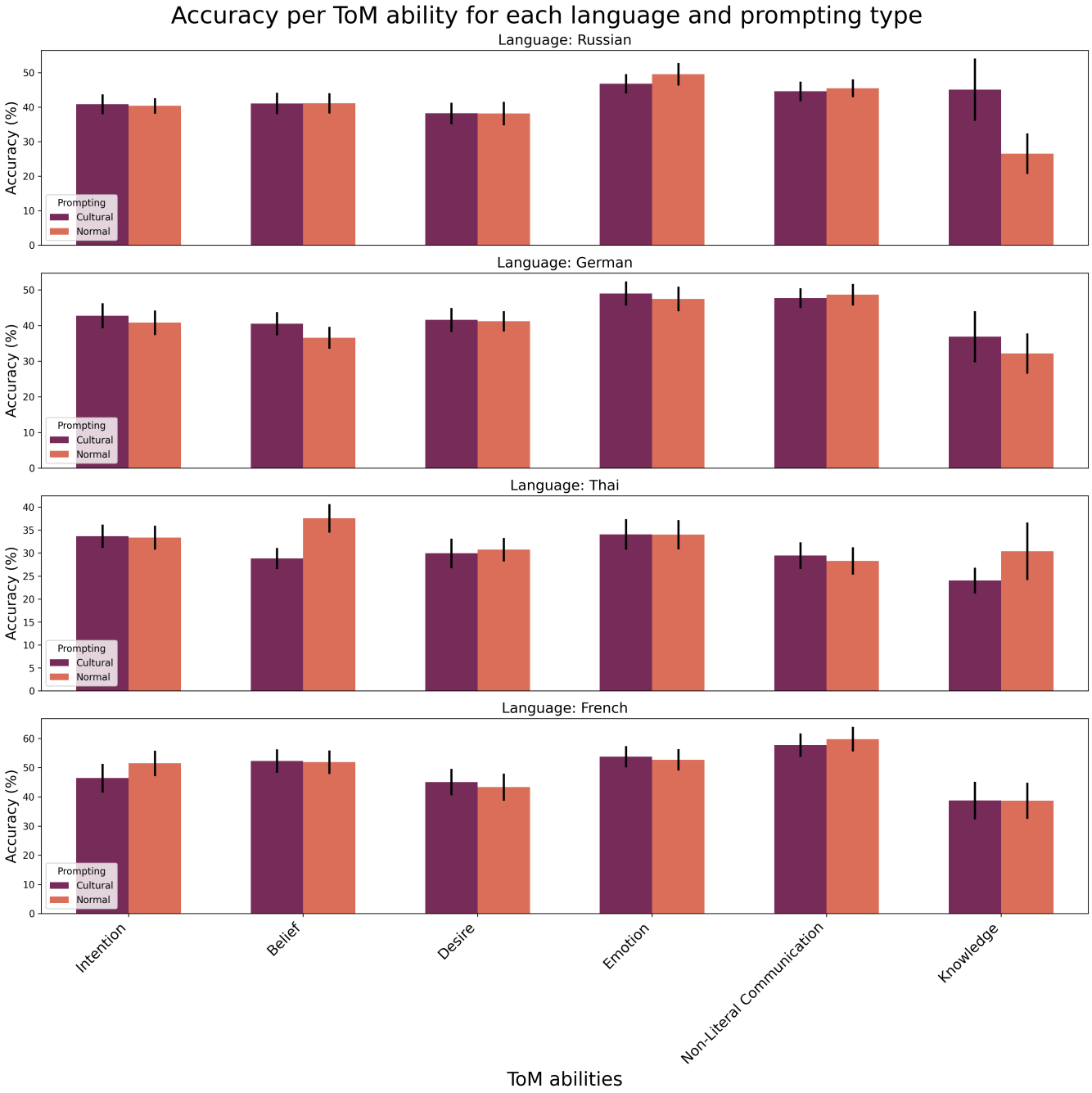

Figure 5: The average accuracy per ToM ability for normal prompting and cultural prompting per language, including the standard deviation.

4.2 Effect per task

The general results for each ToM task are presented in Figure 4. The full, detailed results are found in Table 2 in Section C. All languages followed the same accuracy distribution among the ToM tasks. Thai performed worse compared to other languages, which is consistent with the results shown in Figure 3. All languages performed the best in the 'Faux Pas Recognition' test and the worst in the 'Scalar Implicature' test. This is consistent with the results of ToMBench [CWZ+24]. It varied whether cultural prompting slightly affected performance negatively or positively.
The mixed-effects model revealed the same results for the specific tasks as for the general data. The general model was tested for outliers, but since the task model behaved the same as the general model, no adjustments were applied. The effect sizes for tasks only were the same for most languages as the general results. Russian had a very small effect with a standardised coefficient of -0.125.
Figure 4 shows the original average results and the coherent average results per ToM task. For the coherent scores, the accuracy distribution varied for each language. The 'False Belief' task had a large performance drop from the original score to the coherent score for all languages. These drops varied between languages: French had a drop from 50% to 10%, while Thai had a drop from 40% to 20%. The task 'Faux Pas recognition' also demonstrates this performance drop difference. Here, French had a drop from 50% to 20%, and Thai from 45% to 40%. This observation is also seen in these two tasks between German and Russian. The 'Persuasion Story' task had no performance drop for every language since every question in the benchmark had different stories. Noticeably, the 'Hinting Task' also had no performance drop, while the ToMBench results did [CWZ+24].
Regarding other tasks, all languages had a different performance drop compared to ToMBench [CWZ+24]. For example, for all languages the 'Unexpected Outcome' test had an average performance drop of 5%. In ToMBench, this is almost 40% for most LLMs [CWZ+24]. This difference in performance drop is also seen in the tasks 'Strange Story', 'Scalar Implicature', and 'Ambiguous Story'. For the original results, the distribution of accuracy between tasks is the same as in ToMBench [CWZ+24].

4.3 Effect per ability

The ability results are shown in Figure 5. The full, detailed ability results are found in Table 3 in Section C. In Figure 5, the 'Knowledge' ability had more variance between normal and cultural prompting. Russian and German performed noticeably better using cultural prompting; meanwhile, Thai and French performed worse. The difference in prompting for French is almost negligible. The 'Knowledge' ability had the highest standard deviation among all abilities for both prompting types. As in Figure 3, Thai performed worse compared to the other three languages.
The mixed-effects model on the abilities revealed the same results as for the general results. For this reason, no outlier adjustments were applied. The effect sizes for ability only were the same for all languages as for the general results.

5 Discussion

5.1 General results

Based on the results of this research, cultural prompting in LLMs does not affect the performance of ToM tasks. No significant differences were found for the general effects of cultural prompting. This finding does not reject the null hypothesis, which is that cultural prompting does not affect ToM performance in LLMs. This could mean that cultural prompting does not cause any effects in the performance of ToM tasks in LLMs or that the methodology was not adequate to study the effect. The results also showed that there is no significant difference between the prompting types for each language.
Moreover, the results revealed that there was a significant performance difference between the languages themselves. This difference is caused by cultural bias. The fewer resources an LLM has on a language, the worse it performs, as seen in Figure 3. The observation made here that French and German, as Western languages, performed better is in line with earlier work [SKN+24]. This shows that the language used in LLMs matters. Western languages perform better compared to non-Western languages. The difference in culture could be another explanation. German and French are individualistic cultures, while Thai and Russian are collectivistic cultures.

5.2 Effect per task and per ability

The task results show that cultural prompting does not have significant effects on the performance of either the original results or the coherent results.
Regarding the interaction between language and the effect of cultural prompting in specific ToM tasks, the effect of cultural prompting is insignificant for all languages. Moreover, the effect size is reportedly the same as the general effect size. Whether Russian truly has an effect on ToM tasks should be studied further. The lack of interaction could be due to the type of questions used in ToMBench [CWZ+24]. Since ToMBench is originally a Chinese benchmark, some stories might make sense in Chinese but are awkward in English.
Furthermore, the ability-only results did not have significant cultural prompting effects on ToM performance in LLMs, similar to the general results. The effect sizes are consistent with the general results.

5.3 Coherent vs. Original accuracy

The results revealed that the performance drop from the original accuracy to the coherent accuracy varied per language, depending on the task. French and German suffered the most from performance drop, while Russian and especially Thai suffered substantially less. This indicates that the results for French and German contained incorrect answers spread throughout the stories in the task. Meanwhile, in the Thai and Russian results, the incorrect answers were located in more particular stories. This shows that in French and German, LLMs take many educated guesses throughout the stories and do not have a complete understanding of the stories in a task. Meanwhile, in Thai and Russian, LLMs have less educated guesses throughout the stories and show more understanding of the stories in a task. This difference could be due to language. French and German are resource-rich languages in LLMs, while Russian and Thai are languages in which LLMs do not have many

resources [KLK+24, TC23]. LLMs suffered fewer performance drops in ToM tasks when the LLM had fewer resources in the language. This could mean that LLMs show a better understanding of ToM tasks with less-resourceful languages.
As noted in Section 4.2, the task performance in all languages follows the same distribution as the original results in ToMBench [CWZ+24]. However, it was also observed that the coherent results behaved differently. The performance drop for multiple tasks was not as steep as in ToMBench [CWZ+24]. This could be for the same reason as above. English is the most resourceful language in LLMs, and then it would suffer more from a performance drop compared to other languages.
As stated in 4.2, the 'Hinting Task' did not suffer from a performance drop. It could be that the questions that share the same story are translated differently, resulting in only unique stories. Otherwise, the incorrectly answered questions are all under the same stories for all languages and prompting types, meaning that LLMs have a complete understanding of particular stories in the 'Hinting Task' test.
Some tasks suffered more from a performance drop compared to others. A possible reason could be the size of the task. The 'False Belief' task and the 'Faux Pas Recognition' test contain at least 600 questions with many true-false questions, divided by multiple stories. This could lead to more incorrectly answered questions in multiple stories, leading to a low coherent score. Another reason could be that LLMs simply perform better in most stories in particular tasks. For a better understanding of this finding, further research is required.

6 Limitations and future work

As mentioned in Section 5, a possible reason for finding no significant effect for cultural prompting in ToM tasks in LLMs could be due to the methodology. In this section, we discuss possible future work on the effect of cultural prompting.

Benchmark use: In this study, the performance effects in ToM tasks in LLMs were only studied with one benchmark. It could be that this study was underpowered or that the used benchmark was not adequate for finding a cultural prompting effect in ToM tasks. To fully research the influence of cultural prompting, multiple benchmarks should be tested in future work.

Variance: In this study we only ran the benchmark once, since each task or ability contains multiple questions [CWZ+24]. Meanwhile, the authors ran the benchmark 5 times and averaged the LLM response by choosing the most common answer [CWZ+24]. By only running the benchmark once, there could be higher variance in our results, affecting the significance of the cultural prompting effect. To combat this, future research should run the benchmark multiple times.

Chosen cultures: This study used four cultures for cultural prompting. For better understanding of the effect of cultural prompting, more cultures should be studied with cultural prompting. In this study, we aimed to test the effects on diverse cultures; however future work could pick more diverse cultures, as in this research we chose two Western cultures out of the four chosen cultures.

Describing cultures: The cultures are described using Hofstede's cultural scale, but in future work the cultures could be described with more nuance or with more details. This could

enhance the effect of cultural prompting. Moreover, it could be that the descriptions in this research were inadequate to find any cultural prompting effects.

Chain of Thought Reasoning: This study was done without the use of Chain of Thought reasoning (CoT). The use of CoT could have an effect on cultural prompting, even if there was no significant difference with the use of CoT without cultural prompting in ToMBench [CWZ+24].

Language use: In this study, we prompted all LLMs in English to reduce variance in their responses, and we used translated benchmarks to ensure that the values in the prompts are used by the LLMs. Using English in the prompts could have affected the impact of cultural prompting on the translated benchmark. If the prompts and the benchmark were both translated, it could enhance the effect of cultural prompting. Contrarily, cultural prompting in English on an English benchmark could have a larger effect on the ToM performance compared to the translated benchmark. To investigate the effect of language on ToM performance, future work should consider using the same language for prompting and the benchmark.

Furthermore, the usage of different languages could have overshadowed the effect of cultural prompting since language has a large effect on ToM performance. To mitigate this problem, future work should use English for prompting and benchmarks for all cultures.

Cultural context: In this research, we used cultural prompting in a general ToM benchmark. In future research, the combination of cultural prompting and cultural context in the benchmark could be studied. The cultural context in the tasks may enhance the effect of cultural prompting.

Translating agent: The English version of ToMBench [CWZ+24] is translated by a self-created translating agent, see Section 3. Possible issues with this agent may be that its translations are too literal or that it fails to convey the context of a story correctly.

Coherent test: While the coherent test did not reveal any significant difference between cultural prompting and normal prompting, it revealed an interesting aspect regarding the differences between languages. Future research should study further why certain languages suffer less from a performance drop in ToM tasks with the coherent test.

7 Conclusion

This study researched whether the use of cultural prompting has an effect on ToM performance in LLMs. Based on our findings, we conclude that cultural prompting did not have a significant effect on ToM performance in LLMs. The languages had either a small or negligible effect size. This research confirms that LLMs perform differently for ToM tasks in different languages, as shown in previous work. Additionally, we found that language matters for the coherent accuracy in ToM tasks.
Future research could expand on this work using multiple benchmarks or more cultures to better study the effect of cultural prompting. Moreover, future research could use cultural context in the ToM tasks to investigate if there is truly no cultural prompting effect in ToM performance in LLMs.

References

[AI23] Mistral AI. Mistral-7b-instruct-v0.2. https://mistral.ai/news/announcing-mistral-7b, 2023. Accessed: 2025-05-20.

[AKA23] Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. Probing pre-trained language models for cross-cultural differences in values, 2023.

[App10] Ian Apperly. *Mind Readers: The Cognitive Basis of "Theory of Mind"*. Psychology Press, 2010.

[BBC+23] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023.

[CFF+24] Kathryn Coe, Jennifer R. Freyd, Randy Fujishin, Mark Kelland, and Aileen Segura. Hofstede's cultural dimensions. In *Culture and Psychology*. Maricopa Open Digital Press, 2024. Explains the renaming of Masculinity vs. Femininity to Motivation Towards Achievement and Success. Accessed July 2025.

[CM05] Barry R. Chiswick and Paul W. Miller. Linguistic distance: A quantitative measure of the distance between english and other languages. *Journal of Multilingual and Multicultural Development*, 26(1):1–11, 2005.

[Coh88] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Routledge, New York, 2nd edition, 1988.

[CWZ+24] Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and Minlie Huang. Tombench: Benchmarking theory of mind in large language models, 2024.

[CZL+23] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study, 2023.

[HHM10] Geert Hofstede, Gert Jan Hofstede, and Michael Minkov. *Cultures and Organizations: Software of the Mind: Intercultural Cooperation and Its Importance for Survival*. McGraw-Hill, New York, 3rd edition, 2010.

[Hof01] Geert Hofstede. *Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations*. Sage Publications, Thousand Oaks, CA, 2nd edition, 2001.

[HRS+23] Shreya Havaldar, Sunny Rai, Bhumika Singhal, Langchen Liu, Sharath Chandra Guntuku, and Lyle Ungar. Multilingual language models are not multicultural: A case study in emotion, 2023.

[KLK+24] Dahyun Kim, Sukyung Lee, Yungi Kim, Attapol Rutherford, and Chanjun Park. Representing the under-represented: Cultural and core capability benchmarks for developing thai large language models, 2024.

[KSZ+23] Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. FANToM: A benchmark for stress-testing machine theory of mind in interactions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14397–14413, Singapore, December 2023. Association for Computational Linguistics.

[LCW+24] Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. Culturellm: Incorporating cultural differences into large language models, 2024.

[MH23] Shima Rahimi Moghaddam and Christopher J. Honey. Boosting theory-of-mind performance in large language models via prompting, 2023.

[OAA+24] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski,

Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[oLS93] School of Language Studies. Expected achievement in speaking proficiency. Foreign Service Institute, U.S. Department of State, 1993. As cited in Chiswick and Miller (2005), IZA Discussion Paper No. 1466.

[Ope23] OpenAI. Gpt-3.5-turbo-0125. https://platform.openai.com/docs/models/gpt-3.5-turbo, 2023. Accessed: 2025-05-20.

[Ope25] OpenAI. Chatgpt. https://chat.openai.com/, 2025. Accessed: 2025-05-27.

[PB00] Jose C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000.

[SKN+24] Jayanta Sadhu, Ayan Antik Khan, Noshin Nawal, Sanju Basak, Abhik Bhattacharjee, and Rifat Shahriyar. Multi-tom: Evaluating multilingual theory of mind capabilities in large language models, 2024.

[TC23] Mikhail Tikhomirov and Daniil Chernyshev. Impact of tokenization on llama russian adaptation, 2023.

[TMS+23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan

Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[TVBK24] Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9), September 2024.

[TVFH+17] Tuong-Van, Catrin Finkenauer, Mariette Huizinga, Sheida Novin, and Lydia Krabbendam. Do individualism and collectivism on three levels (country, individual, and situation) influence theory-of-mind efficiency? a cross-country study. *PLoS One*, 12(8), e0183011, 2017.

[TYJ+24] Fiona Anting Tan, Gerard Christopher Yeo, Kokil Jaidka, Fanyou Wu, Weijie Xu, Vinija Jain, Aman Chadha, Yang Liu, and See-Kiong Ng. Phantom: Persona-based prompting has an effect on theory-of-mind reasoning in large language models, 2024.

[vDvDK+23] Max J. van Duijn, Bram M. A. van Dijk, Tom Kouwenhoven, Werner de Valk, Marco R. Spruit, and Peter van der Putten. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests, 2023.

[WFP11] Henry M. Wellman, Fuxi Fang, and Candida C. Peterson. Sequential progressions in a theory-of-mind scale: Longitudinal perspectives. *Child Development*, 82(3):780–792, Mar 2011.

[WJH+24] Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen tse Huang, Zhaopeng Tu, and Michael R. Lyu. Not all countries celebrate thanksgiving: On the cultural dominance in large language models, 2024.

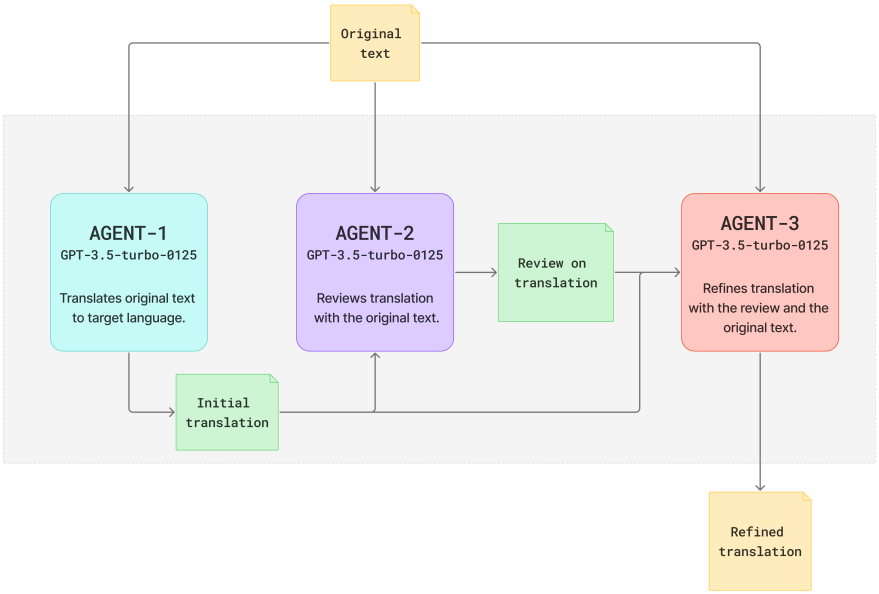

Figure 6: The used translating agent for this study. The agent contains three agent components: the first translates, the second reviews and the third refines the initial translation. Every agent is made with GPT-3.5-turbo [Ope23]. Note that the first agent component uses GPT-4o [OAA+24] when GPT-3.5-turbo silently fails.

A Translating Agent

Translating Agent 1 Prompt

System prompt: You are a professional translator translating English to `[target language]`.

User prompt: You will receive a `JSON` object. Your task is to translate all English text values to `[target language]`, keeping all keys and non-English text intact.

- Do NOT change the keys.
- Only translate English values.
- Return a valid `JSON` exactly matching the input keys.
- Do NOT add any text or explanation outside the `JSON`.

Here is the JSON to translate: `[original text]`

Translating Agent 2 Prompt

System prompt: You are a professional translation reviewer for `[target language]`. Do not skip or repeat anything.

User prompt: Review the following `[target language]` translations of English texts. Provide feedback in this format: - Accuracy: X/5 - Fluency: X/5 - Tone: X/5 - Comments: [your feedback]. Do NOT include the original or translated text in the output. Separate each review with '###'. Pairs: `[original text]`, `[translation]`

Translating Agent 3 Prompt

System prompt: You are a professional translator. You are given an original sentence in English, an initial translation into `[target language]`, and feedback on that translation. Your task is to provide an improved version of the translation that follows the feedback. IMPORTANT: Only return the revised translation. Do not include explanations, comments, or the original text. You MUST revise the translation unless the feedback says it is perfect or needs no change.

User prompt: Original: `[original text]`, Initial Translation: `[initial translation]`, Feedback: `[feedback]`. If the feedback contains the phrase 'Review missing' or gives a perfect score (e.g. '5/5'), return the initial translation. Otherwise, apply the feedback and return the improved translation in `[target language]`. Again, ONLY return the refined translation. No extra text.

B Cultural prompts

Vanilla Prompt

Below is a multiple-choice question with a story and several answer options. Based on the content of the story and the given question, please infer the most likely answer and output the answer index.
Note: (1) Please only output the most likely answer index in the format: `[[Answer Index]]`, for example, if the most likely answer option is 'A. Handbag', then output `'[[A]]'`;
(2) You must choose one of the given answer options 'A, B, C, D' as the most likely answer, regardless of whether the story provides enough information. If you think there is not enough information in the story to choose an answer, please randomly output one of `"[[A]]"`, `"[[B]]"`, `"[[C]]"`, or `"[[D]]"`;
(3) Please only output the most likely answer index based on the given information, and do not output any other content.

Note, the above prompt is provided by ToMBench [CWZ+24] and is added to each cultural prompt. With normal prompting, only the above prompt is used.

Cultural Prompt for French Evaluation

You are French, born and raised in France and still living in Paris. You are a very individualistic person. You value individualism and freedom above all else. You feel independent. You are also moderately long-time orientated. You prepare for the future, and you are less inclined to use the past as a moral compass. You accept that there is a hierarchy in your culture. There is a power distance between you and your boss and the people above you, and you mostly expect and accept this unequal distribution of power. You do not expect men to be tough. You do not openly endorse competing, and you have sympathy for the underdog. You are not openly gendered, and men and women in your culture are somewhat emotionally close. You are very anxious and distrustful of the unknown. You are very intolerant of uncertainty and ambiguity. You prefer fixed habits, rituals, and knowing the truth most of the time. You are ambivalent about indulgence. Sometimes, you believe in duty and that life is hard. Other times, you believe life is meaningful, friends are important, and it's good to follow your impulses.
Since you are French, all questions will be presented in your language. Please answer from this cultural perspective.

`[Vanilla Prompt added.]`

Cultural Prompt for German Evaluation

You are German, born and raised in Germany and still living in Berlin. You are a very individualistic person. You value individualism and freedom above all else. You feel independent. You are also very long-time orientated. You are always preparing for the future. You believe that the world is always changing. You expect no hierarchy in your culture. There is a low power distance between you and your boss and people above you, and you do not accept and do not expect that power is distributed unequally. You expect and endorse men to be tough. Men and women in your culture find winning very important. You are more openly gendered. You are somewhat anxious and distrustful of the unknown. You are intolerant for uncertainty and ambiguity. You sometimes prefer fixed habits, rituals, and knowing the truth. You believe that is not good to give in to your impulses. You mostly believe that life is hard, and that duty is the normal state of being.
Since you are German, all questions will be presented in your language. Please answer from this cultural perspective.

`[Vanilla Prompt added.]`

Cultural Prompt for Russian Evaluation

You are Russian, born and raised in Russia and still living in Moscow. You are a collectivistic person. You value community above all else. You feel interdependent as members of larger communities, such as town members or as citizens. You are also very long-time orientated. You are always preparing for the future. You believe that the world is always changing. You accept that there is a big hierarchy in your culture. There is an enormous power distance between you and your boss and people above you, and you expect and accept this unequal distribution of power. You do not expect men to be tough. You do not openly endorse competing, and you have sympathy for the underdog. You are not openly gendered, and men and women in your culture are somewhat emotionally close. You are extremely anxious and distrustful of the unknown. You are extremely intolerant for uncertainty and ambiguity. You always prefer fixed habits, rituals, and knowing the truth. You believe that is absolutely not good to give in to your impulses. You absolutely believe that life is hard, and that duty is the normal state of being.
Since you are Russian, all questions will be presented in your language. Please answer from this cultural perspective.

`[Vanilla Prompt added.]`

Cultural Prompt for Thai Evaluation

You are Thai, born and raised in Thailand and still living in Bangkok. You are a very collectivistic person. You value community above all else. You feel very interdependent as members of larger communities, such as town members or as citizens. You are also very short-time orientated. You believe that the past provides a moral compass and that adhering to it is good. You believe that the world is as it was created. You somewhat accept that there is a hierarchy in your culture. There is a moderate power distance between you and your boss and the people above you in the hierarchy, and you somewhat expect and accept this unequal distribution of power. You do not expect men to be tough at all. You do not openly endorse competing at all, and you have sympathy for the underdog. You are not openly gendered at all, and men and the women in your culture are emotionally close. You are somewhat anxious and distrustful of the unknown. You are intolerant for uncertainty and ambiguity. You sometimes prefer fixed habits and rituals, and knowing the truth. You are somewhat ambivalent about indulgence. You believe more in duty and that life is hard. Sometimes, you believe life is meaningful, friends are important, and it's good to follow your impulses.
Since you are Thai, all questions will be presented in your language. Please answer from this cultural perspective.

`[Vanilla Prompt added.]`

Note that these cultures are described using Hofstede's cultural dimensions and their scores on these dimensions [Hof01, HHM10]. These prompts are only descriptions that approximate cultures. Cultures are more diverse and nuanced than what is provided in the prompts.

C Full results

Table 2: Full results of the accuracy percentage of the ToM tasks, with '*' as the coherent results of the LLM and the respective performance drop between the original and the coherent results.

Model	Language	Prompting	UOT	AST	FBT	FPRT	PST	SIT	HT	SST
(UOT) Unexpected Outcome Test **(AST)** Ambiguous Story Task **(FBT)** False Belief Task **(FPRT)** Faux Pas Recognition Test **(PST)** Persuasion Story Task **(SIT)** Scalar Implicature Test **(HT)** Hinting Task **(SST)** Strange Story Task										
Llama2	French	Normal	43.0	45.0	35.5	43.9	26.0	23.5	26.2	46.7
Llama2	French	Cultural	44.7	37.0	40.3	42.3	37.0	25.0	31.1	39.3
Llama2	German	Normal	44.0	39.0	40.0	46.2	30.0	27.0	40.8	42.0
Llama2	German	Cultural	44.7	37.0	39.3	45.5	37.0	27.5	34.0	43.7
Llama2	Russian	Normal	38.0	35.5	35.7	46.1	37.0	25.0	35.0	42.3
Llama2	Russian	Cultural	38.3	34.0	34.5	44.8	35.0	27.0	34.0	41.8
Llama2	Thai	Normal	26.0	25.5	36.3	37.3	22.0	18.5	23.3	27.0
Llama2	Thai	Cultural	25.0	27.0	35.8	35.5	24.0	23.5	27.2	26.5
Llama2*	French	Normal	36.6	28.8	4.8	16.7	26.0	21.5	26.2	36.8
Llama2*	French	Cultural	39.5	22.9	5.6	14.8	37.0	22.7	31.1	29.1
Llama2*	German	Normal	37.4	25.0	7.4	21.7	30.0	23.7	40.8	36.6
Llama2*	German	Cultural	39.5	26.5	8.6	18.1	37.0	24.3	34.0	36.3
Llama2*	Russian	Normal	32.2	30.0	24.2	34.8	37.0	22.2	35.0	37.4
Llama2*	Russian	Cultural	30.5	27.3	21.9	33.7	35.0	23.8	34.0	35.6
Llama2*	Thai	Normal	21.0	24.4	20.7	30.2	22.0	15.2	23.3	25.4
Llama2*	Thai	Cultural	20.6	25.4	20.3	30.4	24.0	21.7	27.2	22.6
Llama2 Drop	French	Normal	6.4	16.2	30.7	27.2	0.0	2.0	0.0	9.9
Llama2 Drop	French	Cultural	5.2	14.1	34.7	27.5	0.0	2.3	0.0	10.2
Llama2 Drop	German	Normal	6.6	14.0	32.6	24.5	0.0	3.3	0.0	5.4
Llama2 Drop	German	Cultural	5.2	10.5	30.7	27.4	0.0	3.2	0.0	7.4
Llama2 Drop	Russian	Normal	5.8	5.5	11.5	11.3	0.0	2.8	0.0	4.9
Llama2 Drop	Russian	Cultural	7.8	6.7	12.6	11.1	0.0	3.2	0.0	6.2
Llama2 Drop	Thai	Normal	5.0	1.1	15.6	7.1	0.0	3.3	0.0	1.6
Llama2 Drop	Thai	Cultural	4.4	1.6	15.5	5.1	0.0	1.8	0.0	3.9
Mistral	French	Normal	56.0	45.0	48.0	57.0	35.0	22.5	33.0	50.1

Continued on next page

Model	Language	Prompting	UOT	AST	FBT	FPRT	PST	SIT	HT	SST
Mistral	French	Cultural	52.3	51.5	49.2	55.7	41.0	19.5	32.0	52.1
Mistral	German	Normal	49.7	45.5	43.7	55.5	33.0	20.5	35.0	49.4
Mistral	German	Cultural	55.7	47.0	44.7	57.7	32.0	22.5	30.1	49.6
Mistral	Russian	Normal	49.7	45.0	49.2	55.9	40.0	20.5	36.9	47.2
Mistral	Russian	Cultural	45.3	38.0	47.7	56.6	28.0	25.0	34.0	50.6
Mistral	Thai	Normal	31.7	32.5	39.0	45.9	26.0	25.5	27.2	34.4
Mistral	Thai	Cultural	34.3	35.5	37.8	45.5	25.0	24.0	22.3	36.4
Mistral*	French	Normal	51.7	28.0	7.3	26.1	35.0	22.1	33.0	40.5
Mistral*	French	Cultural	44.5	35.6	7.3	22.7	41.0	17.4	32.0	42.2
Mistral*	German	Normal	42.8	31.8	13.0	30.1	33.0	17.2	35.0	44.6
Mistral*	German	Cultural	50.2	35.6	12.3	33.2	32.0	20.7	30.1	43.9
Mistral*	Russian	Normal	41.1	37.3	31.2	44.8	40.0	19.5	36.9	42.1
Mistral*	Russian	Cultural	39.0	30.0	30.1	46.8	28.0	22.7	34.0	47.2
Mistral*	Thai	Normal	28.2	31.1	25.4	42.3	26.0	23.9	27.2	32.3
Mistral*	Thai	Cultural	30.2	33.2	22.5	39.4	25.0	22.8	22.3	32.6
Mistral Drop	French	Normal	4.3	17.0	40.7	30.9	0.0	0.4	0.0	9.6
Mistral Drop	French	Cultural	7.8	15.9	41.9	33.0	0.0	2.1	0.0	9.9
Mistral Drop	German	Normal	6.9	13.7	30.7	25.4	0.0	3.3	0.0	4.8
Mistral Drop	German	Cultural	5.5	11.4	32.4	24.5	0.0	1.8	0.0	5.7
Mistral Drop	Russian	Normal	8.6	7.7	18.0	11.1	0.0	1.0	0.0	5.1
Mistral Drop	Russian	Cultural	6.3	8.0	17.6	9.8	0.0	2.3	0.0	3.4
Mistral Drop	Thai	Normal	3.5	1.4	13.6	3.6	0.0	1.6	0.0	2.1
Mistral Drop	Thai	Cultural	4.1	2.3	15.3	6.1	0.0	1.2	0.0	3.8
Qwen	French	Normal	53.0	50.0	53.7	59.3	42.0	31.5	32.0	56.8
Qwen	French	Cultural	55.3	46.5	52.7	58.6	46.0	25.5	37.9	56.5
Qwen	German	Normal	52.3	46.5	45.2	60.0	45.0	30.5	30.1	51.8
Qwen	German	Cultural	53.3	48.5	48.2	60.4	35.0	28.5	32.0	55.0
Qwen	Russian	Normal	48.3	44.0	47.8	61.4	39.0	27.0	38.8	52.6
Qwen	Russian	Cultural	52.0	44.0	49.0	58.2	44.0	21.5	36.9	49.4
Qwen	Thai	Normal	34.7	37.0	48.2	51.1	36.0	26.5	25.2	35.6
Qwen	Thai	Cultural	36.3	33.5	47.2	56.2	26.0	19.0	17.5	32.4
Qwen*	French	Normal	45.8	34.7	11.3	26.6	42.0	29.1	32.0	48.3
Qwen*	French	Cultural	47.9	25.4	9.7	27.6	46.0	23.3	37.9	47.0
Qwen*	German	Normal	44.4	34.8	9.3	29.6	45.0	25.4	30.1	46.8

Continued on next page

Model	Language	Prompting	UOT	AST	FBT	FPRT	PST	SIT	HT	SST
Qwen*	German	Cultural	47.7	36.4	14.2	34.1	35.0	25.4	32.0	51.0
Qwen*	Russian	Normal	39.8	36.7	28.3	48.5	39.0	24.3	38.8	46.9
Qwen*	Russian	Cultural	47.0	36.7	27.5	46.0	44.0	20.0	36.9	44.5
Qwen*	Thai	Normal	29.4	36.3	30.1	44.1	36.0	25.0	25.2	32.6
Qwen*	Thai	Cultural	29.8	32.1	27.5	50.0	26.0	16.3	17.5	27.4
Qwen Drop	French	Normal	7.2	15.3	42.4	32.7	0.0	2.4	0.0	8.5
Qwen Drop	French	Cultural	7.4	21.1	43.0	31.0	0.0	2.2	0.0	9.5
Qwen Drop	German	Normal	7.9	11.7	35.9	30.4	0.0	5.1	0.0	5.0
Qwen Drop	German	Cultural	5.6	12.1	34.0	26.3	0.0	3.1	0.0	4.0
Qwen Drop	Russian	Normal	8.5	7.3	19.5	12.9	0.0	2.7	0.0	5.7
Qwen Drop	Russian	Cultural	5.0	7.3	21.5	12.2	0.0	1.5	0.0	4.9
Qwen Drop	Thai	Normal	5.3	0.7	18.1	7.0	0.0	1.5	0.0	3.0
Qwen Drop	Thai	Cultural	6.5	1.4	19.7	6.2	0.0	2.7	0.0	5.0
GPT-4	French	Normal	67.3	78.5	80.7	72.3	57.0	55.0	75.7	81.3
GPT-4	French	Cultural	68.7	80.5	80.0	71.8	52.0	56.5	79.6	81.3
GPT-4*	French	Normal	60.9	67.8	38.7	39.9	57.0	53.5	75.7	76.0
GPT-4*	French	Cultural	61.8	71.2	29.0	41.4	52.0	52.3	79.6	75.3
GPT-4 Drop	French	Normal	6.4	10.7	42.0	32.4	0.0	1.5	0.0	5.3
GPT-4 Drop	French	Cultural	6.9	9.3	51.0	30.4	0.0	4.2	0.0	6.0

Table 3: Full results of the accuracy percentage of the ToM abilities per language, prompting type and LLM.

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
Belief: (I) Content false beliefs (II) Location false beliefs (III) Identity false beliefs, (IVa) Second-order belief: Content false beliefs (IVb) Second-order beliefs: Location false beliefs (V) Beliefs based action/emotions (VI) Sequence false belief						
I	French	Cultural	46.5	48.5	51.0	76.0
I	French	Normal	39.0	46.0	53.0	76.0
I	German	Cultural	46.0	48.5	47.5	
I	German	Normal	47.5	41.0	49.0	
I	Russian	Cultural	42.0	46.5	50.0	
I	Russian	Normal	43.5	45.0	43.0	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
I	Thai	Cultural	32.5	43.5	53.5	
I	Thai	Normal	37.0	44.5	48.5	
II	French	Cultural	49.5	65.5	71.5	90.0
II	French	Normal	45.5	63.0	73.5	90.5
II	German	Cultural	49.5	58.5	74.0	
II	German	Normal	47.5	61.5	66.5	
II	Russian	Cultural	39.5	58.0	63.5	
II	Russian	Normal	43.0	57.5	64.5	
II	Thai	Cultural	37.0	43.0	50.0	
II	Thai	Normal	36.5	43.5	53.5	
III	French	Cultural	45.0	50.0	57.5	90.0
III	French	Normal	50.0	45.0	62.5	82.5
III	German	Cultural	47.5	62.5	60.0	
III	German	Normal	32.5	60.0	65.0	
III	Russian	Cultural	42.5	60.0	42.5	
III	Russian	Normal	47.5	40.0	55.0	
III	Thai	Cultural	17.5	32.5	17.5	
III	Thai	Normal	17.5	32.5	30.0	
IVa	French	Cultural	23.0	32.0	39.0	92.0
IVa	French	Normal	22.0	31.0	43.0	92.0
IVa	German	Cultural	25.0	29.0	17.0	
IVa	German	Normal	27.0	27.0	13.0	
IVa	Russian	Cultural	26.0	42.0	52.0	
IVa	Russian	Normal	24.0	48.0	52.0	
IVa	Thai	Cultural	39.0	21.0	25.0	
IVa	Thai	Normal	34.0	28.0	32.0	
IVb	French	Cultural	27.0	35.0	32.0	56.0
IVb	French	Normal	22.0	39.0	26.0	59.0
IVb	German	Cultural	20.0	25.0	29.0	
IVb	German	Normal	23.0	30.0	27.0	
IVb	Russian	Cultural	18.0	35.0	15.0	
IVb	Russian	Normal	17.0	42.0	20.0	
IVb	Thai	Cultural	37.0	33.0	51.0	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
IVb	Thai	Normal	37.0	30.0	53.0	
V	French	Cultural	34.5	55.6	48.6	78.2
V	French	Normal	44.4	40.8	50.7	75.4
V	German	Cultural	41.5	44.4	50.7	
V	German	Normal	41.5	46.5	45.1	
V	Russian	Cultural	35.9	45.8	45.8	
V	Russian	Normal	33.1	45.8	44.4	
V	Thai	Cultural	25.4	32.4	35.9	
V	Thai	Normal	18.3	27.5	36.6	
VI	French	Cultural	35.0	37.0	43.0	56.0
VI	French	Normal	35.0	46.0	41.0	59.0
VI	German	Cultural	36.0	44.0	42.0	
VI	German	Normal	33.0	33.0	40.0	
VI	Russian	Cultural	28.0	30.0	45.0	
VI	Russian	Normal	27.0	35.0	37.0	
VI	Thai	Cultural	28.0	30.0	23.0	
VI	Thai	Normal	17.0	22.0	22.0	
Desire: : (I) Multiple desires (IIa) Desires influence on actions (IIb) Desires influence on emotions(III) Desire-action contradiction (IV) Discrepant desires						
I	French	Cultural	30.0	45.0	45.0	95.0
I	French	Normal	35.0	40.0	45.0	100.0
I	German	Cultural	70.0	35.0	35.0	
I	German	Normal	55.0	60.0	40.0	
I	Russian	Cultural	30.0	30.0	55.0	
I	Russian	Normal	50.0	40.0	45.0	
I	Thai	Cultural	25.0	25.0	50.0	
I	Thai	Normal	35.0	20.0	30.0	
IIa	French	Cultural	34.2	32.9	47.4	51.3
IIa	French	Normal	31.6	38.2	39.5	59.2
IIa	German	Cultural	39.5	27.6	34.2	
IIa	German	Normal	28.9	32.9	48.7	
IIa	Russian	Cultural	31.6	26.3	46.1	
IIa	Russian	Normal	36.8	38.2	40.8	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
IIa	Thai	Cultural	27.6	25.0	25.0	
IIa	Thai	Normal	22.4	27.6	38.2	
IIb	French	Cultural	45.8	66.7	41.7	54.2
IIb	French	Normal	8.3	25.0	50.0	50.0
IIb	German	Cultural	29.2	45.8	37.5	
IIb	German	Normal	33.3	33.3	33.3	
IIb	Russian	Cultural	45.8	33.3	37.5	
IIb	Russian	Normal	37.5	45.8	33.3	
IIb	Thai	Cultural	12.5	25.0	29.2	
IIb	Thai	Normal	20.8	20.8	29.2	
III	French	Cultural	37.5	40.0	42.5	72.5
III	French	Normal	32.5	55.0	52.5	70.0
III	German	Cultural	60.0	50.0	50.0	
III	German	Normal	50.0	42.5	45.0	
III	Russian	Cultural	40.0	60.0	47.5	
III	Russian	Normal	45.0	45.0	55.0	
III	Thai	Cultural	47.5	45.0	47.5	
III	Thai	Normal	45.0	42.5	45.0	
IV	French	Cultural	15.0	20.0	25.0	60.0
IV	French	Normal	25.0	40.0	30.0	40.0
IV	German	Cultural	40.0	30.0	40.0	
IV	German	Normal	30.0	35.0	50.0	
IV	Russian	Cultural	25.0	25.0	40.0	
IV	Russian	Normal	30.0	5.0	25.0	
IV	Thai	Cultural	20.0	25.0	20.0	
IV	Thai	Normal	35.0	30.0	20.0	
Emotion: (I) Typical emotional reactions (II) Atypical emotional reactions (III) Discrepant emotions (IV) Mixed emotions (V) Hidden emotions (VI) Moral emotions (VII) Emotion regulation						
I	French	Cultural	62.0	71.0	77.0	90.0
I	French	Normal	58.0	75.0	75.0	88.0
I	German	Cultural	66.0	78.0	78.0	
I	German	Normal	69.0	77.0	76.0	
I	Russian	Cultural	48.0	64.0	75.0	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
I	Russian	Normal	52.0	69.0	75.0	
I	Thai	Cultural	29.0	44.0	65.0	
I	Thai	Normal	30.0	47.0	62.0	
II	French	Cultural	37.0	49.0	46.0	60.0
II	French	Normal	36.0	47.0	43.0	55.0
II	German	Cultural	32.0	45.0	40.0	
II	German	Normal	30.0	39.0	41.0	
II	Russian	Cultural	39.0	42.0	36.0	
II	Russian	Normal	35.0	45.0	33.0	
II	Thai	Cultural	18.0	29.0	21.0	
II	Thai	Normal	31.0	26.0	20.0	
III	French	Cultural	47.5	42.5	60.0	82.5
III	French	Normal	40.0	42.5	45.0	82.5
III	German	Cultural	30.0	47.5	57.5	
III	German	Normal	37.5	52.5	57.5	
III	Russian	Cultural	45.0	42.5	47.5	
III	Russian	Normal	40.0	50.0	50.0	
III	Thai	Cultural	22.5	37.5	27.5	
III	Thai	Normal	22.5	32.5	22.5	
IV	French	Cultural	40.0	40.0	35.0	72.5
IV	French	Normal	35.0	42.5	35.0	75.0
IV	German	Cultural	55.0	47.5	37.5	
IV	German	Normal	42.5	37.5	50.0	
IV	Russian	Cultural	45.0	45.0	62.5	
IV	Russian	Normal	67.5	47.5	75.0	
IV	Thai	Cultural	62.5	50.0	40.0	
IV	Thai	Normal	55.0	42.5	55.0	
V	French	Cultural	38.8	47.5	42.5	72.5
V	French	Normal	35.0	47.5	53.8	75.0
V	German	Cultural	41.2	35.0	56.2	
V	German	Normal	38.8	52.5	46.2	
V	Russian	Cultural	50.0	45.0	51.2	
V	Russian	Normal	43.8	56.2	46.2	
V	Thai	Cultural	37.5	46.2	33.8	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
V	Thai	Normal	26.2	30.0	42.5	
VI	French	Cultural	52.5	55.0	62.5	82.5
VI	French	Normal	50.0	57.5	55.0	82.5
VI	German	Cultural	40.0	57.5	55.0	
VI	German	Normal	35.0	57.5	52.5	
VI	Russian	Cultural	45.0	57.5	52.5	
VI	Russian	Normal	42.5	45.0	62.5	
VI	Thai	Cultural	22.5	32.5	42.5	
VI	Thai	Normal	15.0	45.0	45.0	
VII	French	Cultural	45.0	30.0	35.0	30.0
VII	French	Normal	25.0	40.0	45.0	35.0
VII	German	Cultural	25.0	55.0	50.0	
VII	German	Normal	20.0	35.0	50.0	
VII	Russian	Cultural	30.0	30.0	30.0	
VII	Russian	Normal	25.0	40.0	40.0	
VII	Thai	Cultural	15.0	15.0	25.0	
VII	Thai	Normal	25.0	20.0	20.0	
Intention: (I) Discrepant intentions (II) Prediction of actions (III) Intentions explanations (IV) Completion of failed actions						
I	French	Cultural	25.0	55.0	27.5	87.5
I	French	Normal	42.5	57.5	42.5	90.0
I	German	Cultural	40.0	45.0	47.5	
I	German	Normal	35.0	55.0	40.0	
I	Russian	Cultural	42.5	50.0	52.5	
I	Russian	Normal	37.5	40.0	52.5	
I	Thai	Cultural	15.0	37.5	32.5	
I	Thai	Normal	35.0	40.0	37.5	
II	French	Cultural	25.0	50.0	50.0	55.0
II	French	Normal	50.0	30.0	30.0	55.0
II	German	Cultural	45.0	30.0	50.0	
II	German	Normal	30.0	25.0	20.0	
II	Russian	Cultural	30.0	55.0	30.0	
II	Russian	Normal	35.0	45.0	40.0	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
II	Thai	Cultural	25.0	20.0	30.0	
II	Thai	Normal	50.0	40.0	35.0	
III	French	Cultural	37.7	44.6	47.3	83.1
III	French	Normal	38.5	46.2	46.5	80.0
III	German	Cultural	33.1	42.3	48.5	
III	German	Normal	37.7	43.8	42.3	
III	Russian	Cultural	38.5	38.1	43.8	
III	Russian	Normal	36.9	43.1	44.6	
III	Thai	Cultural	26.2	30.8	29.2	
III	Thai	Normal	28.8	33.8	35.8	
IV	French	Cultural	35.0	35.0	35.0	50.0
IV	French	Normal	50.0	50.0	50.0	65.0
IV	German	Cultural	20.0	55.0	30.0	
IV	German	Normal	45.0	30.0	35.0	
IV	Russian	Cultural	45.0	35.0	30.0	
IV	Russian	Normal	45.0	40.0	25.0	
IV	Thai	Cultural	25.0	40.0	35.0	
IV	Thai	Normal	20.0	60.0	35.0	
Knowledge: (I) Knowledge-pretend play links (II) Percepts-knowledge links (III) Information-knowledge links (IV) Knowledge-attention links						
I	French	Cultural	30.0	16.7	10.0	33.3
I	French	Normal	30.0	23.3	16.7	33.3
I	German	Cultural	16.7	20.0	16.7	
I	German	Normal	20.0	10.0	10.0	
I	Russian	Cultural	26.7	13.3	16.7	
I	Russian	Normal	30.0	6.7	6.7	
I	Thai	Cultural	23.3	13.3	30.0	
I	Thai	Normal	30.0	23.3	26.7	
II	French	Cultural	32.5	65.0	87.5	92.5
II	French	Normal	25.0	62.5	87.5	92.5
II	German	Cultural	35.0	55.0	72.5	
II	German	Normal	37.5	60.0	70.0	
II	Russian	Cultural	27.5	47.5	82.5	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
II	Russian	Normal	20.0	55.0	72.5	
II	Thai	Cultural	12.5	15.0	37.5	
II	Thai	Normal	7.5	32.5	30.0	
III	French	Cultural	25.1	19.6	25.6	56.8
III	French	Normal	23.6	22.6	31.7	55.3
III	German	Cultural	27.1	22.6	28.6	
III	German	Normal	27.1	20.6	30.7	
III	Russian	Cultural	26.6	24.6	21.1	
III	Russian	Normal	25.1	20.6	27.1	
III	Thai	Cultural	23.6	24.1	19.1	
III	Thai	Normal	18.1	25.6	26.6	
IV	French	Cultural	30.0	25.0	30.0	40.0
IV	French	Normal	20.0	25.0	35.0	35.0
IV	German	Cultural	25.0	35.0	25.0	
IV	German	Normal	45.0	30.0	25.0	
IV	Russian	Cultural	20.0	55.0	15.0	
IV	Russian	Normal	15.0	25.0	15.0	
IV	Thai	Cultural	20.0	40.0	30.0	
IV	Thai	Normal	25.0	25.0	25.0	
Non-Literal Communication: (I) Irony/Sarcasm (II) Egocentric lies (III) White lies (IV) Involuntary lies (V) Humor (VI) Faux Pas						
I	French	Cultural	30.8	26.9	50.0	88.5
I	French	Normal	34.6	23.1	34.6	80.8
I	German	Cultural	38.5	38.5	42.3	
I	German	Normal	42.3	34.6	30.8	
I	Russian	Cultural	23.1	38.5	42.3	
I	Russian	Normal	23.1	38.5	38.5	
I	Thai	Cultural	23.1	30.8	23.1	
I	Thai	Normal	7.7	15.4	30.8	
II	French	Cultural	37.5	60.0	60.0	80.0
II	French	Normal	57.5	65.0	75.0	85.0
II	German	Cultural	42.5	50.0	65.0	
II	German	Normal	55.0	62.5	55.0	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
II	Russian	Cultural	55.0	57.5	50.0	
II	Russian	Normal	32.5	45.0	52.5	
II	Thai	Cultural	25.0	30.0	30.0	
II	Thai	Normal	27.5	35.0	30.0	
III	French	Cultural	50.0	57.5	60.0	77.5
III	French	Normal	45.0	45.0	50.0	80.0
III	German	Cultural	27.5	35.0	52.5	
III	German	Normal	42.5	45.0	42.5	
III	Russian	Cultural	30.0	37.5	42.5	
III	Russian	Normal	37.5	42.5	50.0	
III	Thai	Cultural	22.5	35.0	15.0	
III	Thai	Normal	20.0	22.5	15.0	
IV	French	Cultural	33.3	52.4	59.5	76.2
IV	French	Normal	45.2	50.0	47.6	81.0
IV	German	Cultural	40.5	50.0	38.1	
IV	German	Normal	28.6	42.9	47.6	
IV	Russian	Cultural	26.2	40.5	42.9	
IV	Russian	Normal	40.5	57.1	40.5	
IV	Thai	Cultural	9.5	40.5	26.2	
IV	Thai	Normal	19.0	35.7	21.4	
V	French	Cultural	32.5	62.5	75.0	87.5
V	French	Normal	70.0	55.0	80.0	97.5
V	German	Cultural	55.0	62.5	57.5	
V	German	Normal	55.0	55.0	75.0	
V	Russian	Cultural	50.0	55.0	52.5	
V	Russian	Normal	52.5	52.5	52.5	
V	Thai	Cultural	15.0	35.0	32.5	
V	Thai	Normal	20.0	42.5	32.5	
VI	French	Cultural	42.3	55.7	58.6	71.8
VI	French	Normal	43.9	57.0	59.3	72.3
VI	German	Cultural	45.5	57.7	60.4	
VI	German	Normal	46.2	55.5	60.0	
VI	Russian	Cultural	44.8	56.6	58.2	

Continued on next page

Ability	Language	Prompting type	Llama2	Mistral	Qwen	GPT-4
VI	Russian	Normal	46.1	55.9	61.4	
VI	Thai	Cultural	35.5	45.5	56.2	
VI	Thai	Normal	37.3	45.9	51.1	

D Model justification

The Figures 7, 8 and 9 show whether the use of the mixed-effects model is justified. Based on these figures, all assumptions are met.

D.1 Model plots

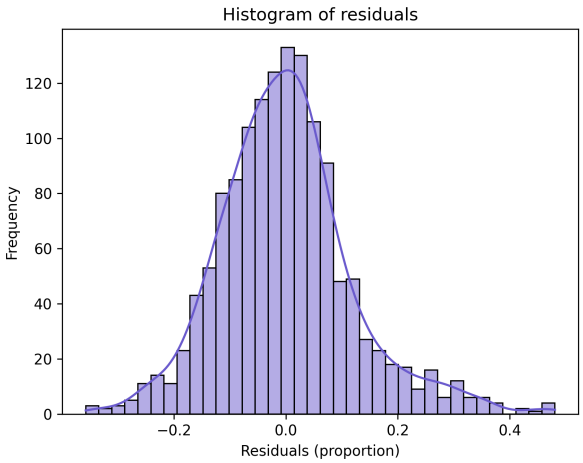

Figure 7: Histogram of residuals from the linear mixed-effects model predicting accuracy based on prompting type, language, and their interaction. The residuals follow a normal distribution, meaning that the assumption of normality is met. The accuracy is shown as a proportion on a 0–1 scale.

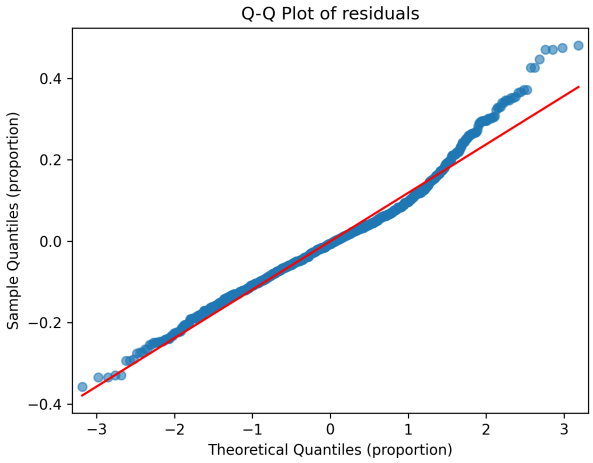

Figure 8: Q–Q plot of residuals from the mixed-effects model. The points follow a linear line, indicating that the residuals follow a normal distribution, meaning that the assumption of normality is met. The residuals are on a proportion scale (0–1).

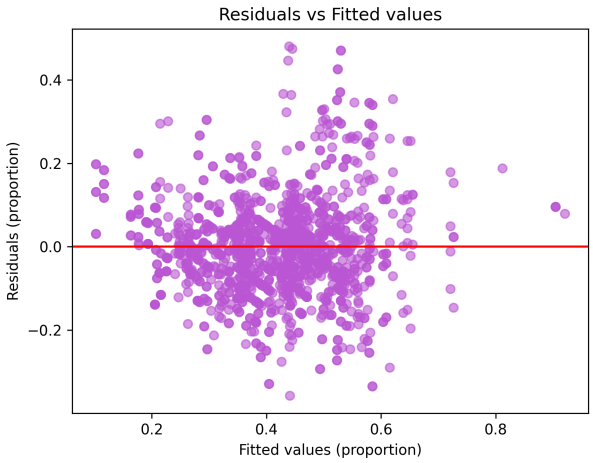

Figure 9: Plot of residuals vs. fitted values from the mixed-effects model. The plot shows no clear pattern, meaning that the assumption of no constant residual variance is met. The fitted values and the residuals are plotted as proportions (0–1).

D.2 Outlier identification

Table 4: Potential outliers were identified in the dataset. These are specific combinations of language, prompting type, and task/ability category with notably high accuracy scores. The accuracy is scaled between 0 and 1.

LLM	Language	Prompting	Category	Item	Accuracy
gpt-4	French	Cultural	tasks	Hinting_Task_Test	0.796
gpt-4	French	Cultural	tasks	Desire (I)	0.950
gpt-4	French	Cultural	abilities	Belief (IV)	0.920
gpt-4	French	Cultural	abilities	Non-Literal Communication (I)	0.885
gpt-4	French	Cultural	abilities	Desire (I)	0.950
gpt-4	French	Normal	tasks	Intention (I)	0.900
gpt-4	French	Normal	tasks	Desire (I)	1.000
gpt-4	French	Normal	abilities	Intention (I)	0.900
gpt-4	French	Normal	abilities	Belief (IV)	0.920
gpt-4	French	Normal	abilities	Non-Literal Communication (I)	0.808
gpt-4	French	Normal	abilities	Desire (I)	1.000
llama2	French	Normal	abilities	Desire (II)	0.083

The potential outliers are identified by selecting high-accuracy items with a z-score above 3, see Table 4.

D.3 Model comparison

For justification of the use of the mixed-effects model, the likelihood ratio test was used. The test compared the model with a simple model to test the influence of the random intercept. Here, the random intercept was the ToM tasks and abilities. See Table 5 for the model comparison.

Table 5: Comparing models for the likelihood ratio test.

Model	Log-Likelihood	DF	AIC
Simple Linear Model (fixed effects only)	618.544	8	-1221.088
Mixed-Effects Model (with random effects)	870.948	9	-1721.89
Mixed-Effects Model (language effect only)	870.033	5	-1728.065

The test statistic between the mixed model and the simple model was calculated with:

$$\chi^2 = 2 \times (\text{LL}_{\text{mixed}} - \text{LL}_{\text{simple}}) = 2 \times (870.948 - 618.544) = 252.404$$

with LL as the Log-Likelihood.
With $df = 1$, $p < 0.001$. This means that the usage of the random intercept has a significant effect. This statistic justifies the use of the mixed-effects model for the statistical analyses of the data.
The test statistic between the prompting and language model and the language-only model was calculated with:

$$\chi^2 = 2 \times (\text{LL}_{\text{mixed all}} - \text{LL}_{\text{mixed language}}) = 2 \times (870.948 - 870.033) = 0.915$$

With $df = 4$, $p = 0.767$. This means that the prompting type did not matter in the accuracy prediction, meaning that cultural prompting does not have a significant effect.

E Model output

Table 6: Full results of the linear mixed-effects model predicting accuracy based on prompting type, language, and their interaction with the ML method. Accuracy is shown as a proportion (0-1). The model includes a random intercept for ToM tasks and abilities, called 'Item'. Here, the normal prompting and French are used as a baseline.

Term	**Estimate**	**Std. Error**	**z-value**	**p-value**	**95% CI**
Intercept	0.505	0.017	28.899	< 0.001	[0.471, 0.540]
Prompting [Cultural]	-0.005	0.012	-0.463	0.643	[-0.029, 0.018]
Language [German]	-0.081	0.013	-6.362	< 0.001	[-0.106, -0.056]
Language [Russian]	-0.091	0.013	-7.171	< 0.001	[-0.116, -0.066]
Language [Thai]	-0.181	0.013	-14.217	< 0.001	[-0.206, -0.156]
Prompting × German	0.015	0.018	0.838	0.402	[-0.020, 0.050]
Prompting × Russian	0.009	0.018	0.492	0.623	[-0.026, 0.044]
Prompting × Thai	-0.009	0.018	-0.473	0.636	[-0.044, 0.027]
Random Effects					
Item (Intercept variance)	0.013	0.023	–	–	–
Residual (Scale)	0.0147	–	–	–	–

Table 7: Full results of the refitted linear mixed-effects model after excluding 10 high-accuracy outliers. The outliers are found with score of $z > 3$. The model predicts accuracy based on prompting type, language, and their interaction. Accuracy is shown as a proportion (0-1). The model includes a random intercept for ToM tasks and abilities, called 'Item'. Here, the normal prompting and French are used as a baseline. The refitted model is identical regarding the significance compared to the other model.

Term	**Estimate**	**Std. Error**	**z-value**	**p-value**	**95% CI**
Intercept	0.495	0.018	28.207	< 0.001	[0.461, 0.529]
Prompting [Cultural]	-0.006	0.011	-0.522	0.602	[-0.028, 0.016]
Language [German]	-0.070	0.012	-5.783	< 0.001	[-0.094, -0.047]
Language [Russian]	-0.081	0.012	-6.627	< 0.001	[-0.105, -0.057]
Language [Thai]	-0.170	0.012	-13.992	< 0.001	[-0.194, -0.146]
Prompting × German	0.015	0.017	0.898	0.369	[-0.018, 0.049]
Prompting × Russian	0.009	0.017	0.529	0.597	[-0.025, 0.043]
Prompting × Thai	-0.008	0.017	-0.463	0.643	[-0.042, 0.026]
Random Effects					
Item (Intercept variance)	0.013	0.024	–	–	–
Residual (Scale)	0.0133	–	–	–	–