MultiNRC: A Challenging and Native Multilingual Reasoning Evaluation Benchmark for LLMs

Anonymous ACL submission

Abstract

Although recent Large Language Models (LLMs) have shown rapid improvement on reasoning benchmarks in English, the evaluation of such LLMs' multilingual reasoning capability across diverse languages and cultural contexts remains limited. Existing multilingual reasoning benchmarks are typically constructed by translating existing English reasoning benchmarks, biasing these benchmarks towards reasoning problems with context in English language/cultures. In this work, we introduce the Multilingual Native Reasoning Challenge (**MultiNRC**), a benchmark designed to assess LLMs on more than 1,000 native, linguistic and culturally grounded reasoning questions written by native speakers in French, Spanish, and Chinese. MultiNRC covers four core reasoning categories: language-specific linguistic reasoning, wordplay & riddles, cultural/tradition reasoning, and math reasoning with cultural relevance. For cultural/tradition reasoning and math reasoning with cultural relevance, we also provide English equivalent translations of the multilingual questions by manual translation from native speakers fluent in English. This set of English equivalents can provide a direct comparison of LLM reasoning capacity in other languages vs. English on the same reasoning questions. We systematically evaluate current 14 leading LLMs covering most LLM families on MultiNRC and its English equivalent set. The results show that (1) current LLMs are still not good at native multilingual reasoning, with none scoring above 50% on MultiNRC; (2) LLMs exhibit distinct strengths and weaknesses in handling linguistic, cultural, and logical reasoning tasks; (3) Most models perform substantially better in math reasoning in English compared to in original languages (+10%), indicating persistent challenges with culturally grounded knowledge.

1 Introduction

Enhancing the reasoning capabilities of Large Language Models (LLMs) is a critical area of ongoing research and development (OpenAI, 2024; Guo et al., 2025). Consequently, thorough and diverse reasoning evaluation benchmarks in English have been developed to assess LLMs' improved reasoning capabilities. These evaluation datasets cover a variety of reasoning tasks, including commonsense reasoning (Zellers et al., 2019; Sakaguchi et al., 2021; Clark et al., 2018), language-understanding reasoning in English (Hendrycks et al., 2020; Wang et al., 2025), STEM reasoning (Cobbe et al., 2021; Zhong et al., 2024; Hendrycks et al.), and logical reasoning (Liu et al., 2020; Yu et al.; Geva et al., 2021). These benchmarks play a critical role in guiding the evolution of the reasoning capacity of LLMs.

While a large number of these benchmarks have been developed in English, the landscape of multilingual reasoning benchmarks is less developed, comprehensive, and equitable. Existing multilingual reasoning benchmarks are largely created by translating existing English reasoning benchmarks (automatically, manually, or hybrid) into target languages (Shi et al.; Patel et al., 2021). Benchmarks created by directly translating English-centric content may reflect the cultural framing and linguistic characteristic of English, which are not globally representative. As a result, reasoning tasks that require native linguistic or cultural understanding are often missing, leaving gaps in the evaluation of non-English language reasoning abilities. Consequently, we are still largely unaware of how state-of-the-art (SOTA) LLMs perform on genuinely native multilingual reasoning tasks that demand both linguistic diversity and cultural or contextual depth.

Motivated by this gap, we develop the Multilingual Native Reasoning Challenge (MultiNRC), a

new multilingual evaluation benchmark that contains native and challenging reasoning questions of the targeted language, to assess LLMs' reasoning capability under a natural and native context. Specifically, MultiNRC contains four categories of reasoning questions: language-specific linguistic reasoning, wordplay riddles, cultural/tradition reasoning, and finally mathematical reasoning with cultural relevance. We first release MultiNRC in French, Spanish, and Chinese; Table 1 shows examples of the four categories in different languages. To build MultiNRC, we recruit native speakers of each language to create challenging reasoning questions and ground-truth final answers according to the definitions of the four categories above. We only accept reasoning questions that cause at least 3 out of 5 SOTA LLMs to fail, and we facilitate MultiNRC with automatic evaluation with LLM-as-a-judge, for fast and accurate model assessments. Automatic evaluation is made possible by only including reasoning questions with an objective and short ground-truth final answer; we find that our automatic evaluation has more than 95% alignment with human judgment on MultiNRC.

We thoroughly evaluate current SOTA LLMs on MultiNRC. We find that the best performing LLM, o3-pro (OpenAI, 2025) released in April 2025, achieves only 49% accuracy according to automatic evaluation with respect to our ground-truth final answers, indicating the high difficulty level of MultiNRC for SOTA models. We find that o3 variants along with Gemini-2.5-Pro (June 2025) (Mallick and Kilpatrick, 2025) greatly outperform other models, and that LLMs also rank differently across MultiNRC reasoning categories, suggesting that the taxonomy of MultiNRC provides a more informative view to guide LLM performance improvement. For example, Gemini-2.5-Pro ranks third on average accuracy among models but ranks first in math reasoning with cultural relevance.

We also augment MultiNRC with English equivalent translations of the non-English prompts by having our annotators translate the prompts into English while retaining the structure and solvability of the original question. In this English equivalent set, we only translate cultural reasoning and math categories. This is because linguistic and wordplay categories cannot be translated into reasonable equivalents; for example, a play on words in French translated to English would no longer make sense. This English-translated evaluation set can provide a direct comparison of LLM reasoning capability in other languages vs. English on same reasoning questions. We find that LLMs perform much better on the resulting set of English math reasoning, achieving 10% higher accuracy on average, compared to their performance on the original multilingual math reasoning with cultural relevance. Our case study shows that the reason for this phenomenon is likely that LLMs are able to better retrieve cultural context/knowledge in English for math problems, compared to retrieving such cultural context/knowledge in its original language. However, in the category of cultural reasoning, there is no large performance difference in English equivalent prompts compared to the original multilingual prompts. Our case study shows that the cultural/tradition context in this category is much more specific and nuanced compared to the math category, and thus often absent in both the LLM knowledge base of English or original language, leading to no performance improvement on English-equivalent prompts.

Our contributions are as follows. First, we introduce MultiNRC, a native, challenging multilingual reasoning benchmark authored by native speakers to cover reasoning questions with linguistic and cultural nuances. Second, we systematically evaluate 14 strong LLMs in French, Spanish, and Chinese, on both native questions and English translations, to understand the true multilingual reasoning capabilities of models. Third, we perform extensive ablation studies across languages and reasoning categories, as well as the effect of model translation and reasoning in English to disentangle the impact of language, translation, and cultural context on model reasoning. Our analyses highlight the underlying cultural and contextual knowledge required to answer MultiNRC questions. Finally, we will release the dataset to facilitate further research on multilingual and culturally informed evaluation.

2 Related Work

Reasoning benchmarks in English. A wide range of benchmarks have been developed to evaluate the reasoning capabilities of LLMs across commonsense, logical, and STEM reasoning domains. Commonsense reasoning benchmarks aim to evaluate an LLM's ability to apply everyday knowledge and logical inference to under-

Language	Category	Prompt	Answer
French	Linguistic	Si je suis un mot masculin seul, féminin au pluriel, je ne suis jamais bouclé. Qui suis-je?	Délice(s)
Chinese	Wordplay	你可以用谐音梗，用一个成语回答我：为什么井越浅越好？因为 _____。	好景不长
Spanish	Cultural/Trad	Me voy de viaje a Cancún el próximo año en 2026. Quiero estar allá para el Día de la Candelaria, pero quiero tener dos fines de semana completos para disfrutar la playa. ¿Cuáles tendrían que ser las fechas de mis vuelos?	Ida: viernes, 30 de enero, 2026 Regreso: lunes, 9 de febrero, 2026
French	Math	J'ai trouvé une maison à acheter en viager en 2025 : le bouquet est de 90 000€ et la mensualité de 500€. L'occupante est née l'année de la sixième édition du Festival de Cannes, la maison est estimée à 150 000€, et va prendre 1% de valeur par an. A partir de quel âge de l'occupante cela devient une moins bonne affaire pour moi?	86 ans.

Table 1: Representative native-language example prompts and short answers from MultiNRC.

stand and interact with the world in a human-like way. HellaSwag (Zellers et al., 2019) and WinoGrande (Sakaguchi et al., 2021) are two widely-adopted examples. Logical and deductive reasoning benchmarks, such as ReClor (Yu et al.) and StrategyQA (Geva et al., 2021), and ARC-AGI (Chollet et al., 2024), are designed to evaluate an LLM's capacity to apply formal rules of logic and to perform deductive, inductive, or abductive inference given a set of premises. To comprehensively assess both general knowledge and reasoning, multitask benchmarks like MMLU-Pro (Wang et al., 2025) have emerged, expanding on prior efforts in MMLU (Hendrycks et al., 2020) by including more rigorous reasoning components and a broader coverage of academic subjects, rather than just factual recall. STEM and math reasoning benchmarks, such as AIME (Mathematical Association of America, 2024) and MATH (Hendrycks et al.), are specifically designed to assess an LLM's proficiency in solving complex, multi-step mathematical problems. Distinct from prior benchmarks, we focus on evaluating reasoning over linguistic, wordplay, cultural/tradition, and math questions, each written by native speakers to capture local language and cultural nuance.

Multilingual reasoning benchmarks. Compared to LLM reasoning evaluation in English, multilingual reasoning benchmarks are less developed. A majority of them are translated from existing English reasoning benchmarks (Li et al., 2024; Shi et al.; Huang et al., 2025). Among them, MMLU-ProX (Xuan et al., 2025), is built by translating the existing English benchmark, MMLU-Pro (Wang et al., 2025), to evaluate advanced reasoning across multiple languages. The initial translation of the English questions into the other 28 languages was performed using multiple LLMs, followed by an expert review phase. Similarly, MGSM (Shi et al.) and BenchMAX (Huang et al., 2025) are translated from GSM8K (Cobbe et al., 2021) and GPQA (Rein et al.). BenchMAX is built with an initial machine translation and a meticulous post-editing process in which each sample was independently reviewed by three distinct native-language annotators. MGSM was created by manually translating 250 problems selected from the English GSM8K dataset. An approach to avoid translating existing English benchmarks is by using exam questions in non-English languages (Hardalov et al., 2020; Zhang et al., 2023; Wang et al., 2024). However, such benchmarks, with M3Exam as an example, have mostly multichoice questions that assess knowledge instead of reasoning and are too easy for SOTA models, as suggested by their reported GPT-4 results in 2023 (Zhang et al., 2023). Another example of data collected from native languages is MultiLoko (Hupkes and Bogoychev, 2025), which collects SQuAD-style (Rajpurkar et al., 2016) questions from the most popular Wikipedia pages for each language, rated for their local relevance to the language. While this work also analyzes model performance on English equivalents, the focus of our work is on reasoning questions over linguistic and cultural knowledge and nuance, and not just factual knowledge.

3 Evaluation Benchmark Overview

Below we describe the Multilingual Native Reasoning Challenge (MultiNRC) benchmark, detailing our choice of reasoning categories, English equivalent set of multilingual prompts, as well as dataset construction methodology and statistics.

3.1 Reasoning Categories

MultiNRC contains 4 categories of reasoning questions. They are language-specific linguistic rea-

soning, wordplay riddles, cultural/tradition reasoning, and finally, math reasoning with cultural relevance. We extract and summarize the 4 categories from preliminary crowd-sourced native and challenging reasoning questions that we collect for multiple languages and cultures. The 4 categories are uniquely defined from each other to assess the different capabilities of LLM required in multilingual reasoning tasks. Table 1 shows example tasks in all categories, which are described in the following paragraphs. Additional examples can be found in the Appendix in Table 7.

Language-specific Linguistic Reasoning questions are based on grammatical rules, honorifics, or language-specific conventions that only exist in the target language but not English. Such questions require reasoning about language structure or usage. Examples under this category include word formation problems, relation inference based on language conventions, etc. We enforce that the questions should require multi-step reasoning, instead of mere linguistic feature identification without reasoning. In the French example in Table 1, the question requires recognizing that the noun délice is masculine in the singular and feminine in the plural, an uncommon grammatical characteristic not present in English.

Wordplay & Riddles contain puzzles that rely on the clever and ambiguous use of the target language, often involving multiple meanings, homophones, or puns. Since such reasoning questions usually use very language-specific homophones or puns, there are no English translations that can provide the exact meaning of such questions. Our evaluations later show that this category is one of the most difficult for LLMs. In the Chinese example shown in Table 1, the question requires the respondent to use a Chengyu (Ancient Chinese idiom) involving a pun or homophone as an answer to a playful riddle, thus demanding clever reasoning.

Cultural/Tradition Reasoning questions require reasoning through timelines, behaviors, or customs derived from local traditions, holidays, or ceremonies. Similar to other categories, we enforce the assessment of LLM reasoning capabilities and exclude purely factual questions. In the Spanish example in the table, the question requires identifying the date of a local holiday (Día de la Candelaria) and reasoning about travel dates to ensure two full weekends are spent at the destination, demonstrating an understanding of local traditions and applying multi-step temporal reasoning.

Math Reasoning with Cultural Relevance questions involve calculations based on cultural-specific counting systems, calendars, currencies, or numerical phrasing. As above, it must require at least 1 reasoning step based on a culturally specific element, and we do not allow pure arithmetic or logic without cultural grounding. In the French example in the table, the question requires the application of culturally specific knowledge about the 'viager' real estate system and historical events (such as the date of the Cannes Film Festival) to determine the financial break-even point for a property purchase, thereby testing mathematical reasoning within a local context.

3.2 English Equivalent Prompts

In order to test the reasoning gap in English vs. non-English languages, we further ask annotators to translate the native prompts and ground-truth final answers into English. Annotators judge the validity of machine-translated prompts from 5 SOTA models and choose the most relevant machine-translated prompt from the above pipeline as a base for translation, or translate from scratch as needed. Our annotators focus on translating the logic, not literal words, ensuring that the English translation is also culturally grounded and requires similar reasoning. We only produce English translations for the Cultural/Tradition Reasoning and Math Reasoning categories. The translated English version retains the structure and solvability of the original, with the translator trying to translate cultural or linguistic characteristics into English. Examples of English translated prompts and answers are shown in the Appendix in Table 8. We found that the categories of Language-specific Linguistic Reasoning and Wordplay & Riddles could not be reliably translated into English, as they often rely on linguistic phenomena that do not translate or exist in English. Thus, we exclude these categories from the English equivalent set.

3.3 Data Collection/Review and Statistics

To collect data for MultiNRC, native speakers start by writing reasoning prompts according to the definition of one of the four categories mentioned above. For each prompt, the annotator also writes a ground-truth final answer (GTFA), consisting of either a single word or a short phrase/sentence. We then pull model responses from five top models at the time of creation

Lang	# Tot.	Linguistic	Wordplay	Cultural/Tradition	Math
French	338	75	79	107	77
Spanish	392	71	93	122	106
Chinese	325	71	86	90	78
Total	1055	217	258	319	261

Table 2: MultiNRC dataset statistics by language and category.

across model providers: Google's Gemini-2.5-pro (DeepMind, 2025), o4-mini (OpenAI, 2025), Claude 3.7-Sonnet (Anthropic, 2025), Deepseek R1 (DeepSeek-AI, 2025), and Llama-4 Maverick (AI, 2024). The 5 model responses are presented to annotators in a random order. We only keep the reasoning questions that 3 or more of the 5 models fail to correctly answer.

After the raw data is collected, we have two native-speaker reviewer layers to assess the quality of the data from different aspects. The first native-speaker reviewer layer is to assess the quality of the prompt and GTFA pairs. At this layer we would assess both the GTFA's accuracy and the prompt's reasonableness and alignment with category definition. The second native-speaker reviewer layer assesses whether there are truly 3 out of 5 SOTA model failures, given the verified prompt and GTFA pairs. This is to verify the difficulty level of the collected data. We show the statistics of the final dataset after these collection/review steps in Table 2. We collect roughly the same amount of examples across categories and languages, more than the 250 per-language count of the popular MGSM dataset (Shi et al.). We also annotate subcategories of the cultural and linguistic nuances present in the prompts. The results of this analysis are found in the Appendix in Table 9.

4 Experimental Settings

Reasoning Settings We perform experiments in the following settings: 1) **Original** in which the model is given the original multilingual question, and prompted to provide an explanation and a short final answer in the native language of the question. 2) **En** in which the model answers human-translated English equivalent questions (illustrated in Section 3.2) in English. By comparing **En** vs **Original**, we can see the difference in model's abilities to understand and reason about local linguistic and cultural nuance in the local language versus English. (Hupkes and Bogoychev, 2025) refer to this as the Mother Tongue Effect. 3) **EnT/R** in which the model is prompted to produce an English translation of the native prompt by itself, reason/explain in English and return the native final answer. 4) **EnR** in which the model is asked to reason in English on the native prompt and return the native language response. The results of these last two settings are found in Appendix A.2.

Automatic Evaluation As described above, we collect human judgment of the correctness of model responses for 5 SOTA models on our benchmark. In order to scale annotations, we leverage an LLM judge that takes the prompt, GTFA, and model response and produces a correct/not correct judgment as well as reasoning behind the judgment. We found that GPT-4.1(OpenAI, 2024) as the judge results in an over 95% agreement with our human judgments and a Scott's Pi (Scott, 1955) of 0.88, indicating strong reliability of our metric, and thus we utilize this automatic judge for all reported results. Note that all experimental results use pass@1.

Model Selection and Inference Parameters We select 14 model variations across model providers, to obtain as generalizable results as possible given current SOTA models. We test a range of recent LLMs including Google's Gemini-2.5-pro (DeepMind, 2025), OpenAI's o3 and o4-mini (OpenAI, 2025), gpt-4.1 (OpenAI, 2024), Anthropic's Claude (Opus, Sonnet, including thinking and non-thinking models), and Llama-4 (AI, 2024). Note that -high or -medium refers to the reasoning effort setting. We generate with max completion tokens of 32k, thinking budget of 16k where applicable, and default to the model provider's default temperature.

5 Results and Analysis

The main results on MultiNRC are found in Table 3, showing model performance in the Original setting, along with 95% confidence intervals. We note that o3 variants along with Gemini-2.5-

Model	Original
o3-pro-2025-06-10-high	**49.00 (±3.02)**
o3-2025-04-16-high	**45.50 (±3.00)**
Gemini-2.5-Pro-Preview-06-05	**45.12 (±3.00)**
o3-2025-04-16-medium	44.45 (±3.00)
Claude-4-Opus-20250514-thinking	33.93 (±2.86)
Claude-4-Opus-20250514	29.00 (±2.74)
Claude-3.7-Sonnet-thinking	27.77 (±2.70)
Deepseek-R1-0528	27.58 (±2.70)
Deepseek-R1	24.27 (±2.59)
o4-mini-high	22.18 (±2.51)
GPT-4.1	21.23 (±2.47)
Claude-4-Sonnet-20250514	18.39 (±2.34)
GPT-4o	12.42 (±1.99)
Llama-4-Maverick	8.44 (±1.68)

Table 3: Accuracy (%) of models in the **Original** setting (reasoning and answering in the original language) on the full MultiNRC benchmark, with 95% confidence intervals in parentheses.

pro-preview-06-05 score clearly better than any other model. However, no model surpasses 50% accuracy on the dataset, underscoring the difficulty of MultiNRC and the substantial performance gaps even among top models. We also find trends of progress in recent models; for example, Deepseek-R1-0528 improves considerably over the initial Deepseek-R1, and GPT-4.1 outperforms GPT-4o. Furthermore, thinking model variants (e.g., Claude-4-Opus-20250514-thinking) consistently show better performance within the same family, indicating the benefit of enhanced reasoning effort.

In Table 4, we break down model performance on MultiNRC in the Original setting by both category and language. Math reasoning questions are the most challenging overall, with an average accuracy of just 23.3% across all questions, and especially low scores in Spanish. In contrast, linguistic and cultural reasoning tasks see higher scores, with average accuracies of 34.8% and 31.1%. Wordplay remains a consistently difficult category for all languages, although the models score better in French for this category. Spanish stands out as the most challenging language across categories, particularly for math and wordplay. Some models exhibit distinct strengths in particular languages and categories, as seen in Deepseek-R1-05-28's higher rank on Chinese Math and Linguistic categories. These results illustrate that model performance is affected not only by the reasoning category itself but also by the interaction between languages and reasoning categories.

(Hupkes and Bogoychev, 2025) shows the Mother Tongue Effect, calculated as the average performance in the original language minus the average performance in English over all questions. However, in contrast to (Hupkes and Bogoychev, 2025), here in Table 10 in the Appendix we note a negative effect on linguistically and culturally nuanced reasoning questions. We also include model consistency, as a measure of the percentage of questions that a model answers correctly in either original or En divided by the number of questions the model answers correctly in total. For most models, the consistency is around or below 50%, but we note that the top-performing models also have higher consistency, and that consistency has a strong rank correlation with original. The GAP column quantifies this difference between the best-performing language and the worst for each model, with an average gap of 13.90 points across all models. We observed that some models show especially large differences across languages; for example, Deepseek-R1-0528 scores 22.81 points higher in Chinese than in Spanish, and the original Deepseek-R1 a 29.10 difference. This underscores that even strong multilingual models perform very differently across languages.

Within categories, we see large performance gaps between models. For example, although top-performing models generally struggle on Math overall, Gemini performs much better than others due to its strong results in Spanish Math questions. Similarly, o3-pro demonstrates a significant improvement in Wordplay, especially French Wordplay (75.9%). As a result, models often rank differently across MultiNRC reasoning categories and languages. These results suggest that the taxonomy of MultiNRC provides a more informative view to guide LLM performance improvements.

The results described above show a full picture of LLM performance in Original setting of MultiNRC. Table 5 presents the performance delta, calculated as the difference between accuracy on English-translated questions and the original native questions, for both Cultural and Math tasks across French, Spanish, and Chinese. We observe that scores on English-translated versions of Chinese and Spanish questions are often higher than the original scores, although this is model-dependent, and several top-performing models, such as o3 variants, perform better in French compared to Original. These results suggest that for certain languages and categories, language-specific strengths and weaknesses remain a signif-

Model	Cultural			Math			Linguistic			Wordplay		
	Fr	Es	Zh	Fr	Es	Zh	Fr	Es	Zh	Fr	Es	Zh
o3-pro	46.7	**45.9**	50.0	**57.1**	**24.5**	24.4	**62.7**	**47.9**	**53.5**	**75.9**	**53.8**	**55.8**
o3-2025-04-16-high	**49.5**	**46.7**	48.9	**50.6**	23.6	24.4	**65.3**	**52.1**	45.1	**67.1**	**43.0**	**37.2**
Gemini-2.5-Pro-Preview-06-05	**52.3**	40.2	**61.1**	37.7	**41.5**	**43.6**	52.0	39.4	**49.3**	53.2	36.6	36.0
o3-2025-04-16-medium	**48.6**	**46.7**	**52.2**	**41.6**	20.8	25.6	**66.7**	**46.5**	42.3	**67.1**	**37.6**	**44.2**
Claude-4-Opus-20250514-thinking	36.4	20.5	**54.4**	32.5	21.7	**32.1**	34.7	32.4	**46.5**	48.1	32.3	25.6
Claude-4-Opus-20250514	29.9	23.8	45.6	27.3	19.8	28.2	28.0	23.9	39.4	36.7	29.0	20.9
Claude-3.7-Sonnet-thinking	26.2	23.8	21.1	32.5	**26.4**	20.5	40.0	40.8	39.4	34.2	22.6	15.1
Deepseek-R1-0528	27.1	23.8	51.1	22.1	13.2	**30.8**	25.3	42.3	**46.5**	12.7	17.2	27.9
Deepseek-R1	23.4	13.9	53.3	24.7	12.3	29.5	18.7	19.7	38.0	10.1	19.4	34.9
o4-mini-high	26.2	20.5	16.7	23.4	7.5	15.4	46.7	31.0	15.5	26.6	29.0	14.0
GPT-4.1	29.0	27.0	13.3	26.0	18.9	14.1	25.3	21.1	28.2	24.1	16.1	10.5
Claude-4-Sonnet-20250514	16.8	11.5	27.8	20.8	16.0	20.5	17.3	23.9	23.9	22.8	11.8	14.0
GPT-4o	18.7	13.9	15.6	9.1	9.4	6.4	13.3	21.1	21.1	11.4	4.3	5.8
Llama-4-Maverick	13.1	8.2	5.6	11.7	6.6	10.3	9.3	12.7	11.3	8.9	2.2	3.5
Average	**31.7**	**26.2**	**36.9**	**29.8**	**18.7**	**23.3**	**36.1**	**32.5**	**35.7**	**35.6**	**25.3**	**24.7**

Table 4: Model performance in the Original setting broken down by category and language.

Model	Cultural			Math		
	ΔFr	ΔEs	ΔZh	ΔFr	ΔEs	ΔZh
o3-pro	**-9.3**	4.1	7.8	-5.2	**17.9**	12.8
o3-2025-04-16-high	-5.6	-6.6	**6.7**	-7.8	**16.0**	10.3
Gemini-2.5-Pro-Preview-06-05	**-10.3**	1.6	0.0	2.6	0.9	**10.3**
o3-2025-04-16-medium	-4.7	-1.6	**-8.9**	0.0	**22.6**	10.3
Claude-4-Opus-20250514-thinking	**-12.1**	5.7	-7.8	3.9	**18.9**	15.4
Claude-4-Opus-20250514	-4.7	0.8	**-3.3**	3.9	**14.2**	10.3
Claude-3.7-Sonnet-thinking	-0.9	0.0	**5.6**	-3.9	**13.2**	11.5
Deepseek-R1-0528	-1.9	**17.2**	-2.2	9.1	33.0	**33.3**
Deepseek-R1	**-6.5**	4.1	-7.8	0.0	**22.6**	10.3
o4-mini-high	-3.7	7.4	**8.9**	1.3	**28.3**	14.1
GPT-4.1	-0.9	1.6	**7.8**	2.6	**17.9**	9.0
Claude-4-Sonnet-20250514	**-4.7**	4.1	-1.1	2.6	**13.2**	6.4
GPT-4o	-1.9	4.9	**6.7**	5.2	**23.6**	9.0
Llama-4-Maverick	-2.8	5.7	**7.8**	0.0	**25.5**	12.8
Average	**-5.0**	**3.5**	**1.4**	**1.0**	**19.1**	**12.5**

Table 5: Delta (English − Original) in accuracy (%) for Cultural and Math tasks, according to the Original language. **Bold** indicates the value with largest magnitude in each category per model. Positive values indicate a performance improvement on English translated queries vs Original.

icant factor in LLM performance.

For math reasoning, we observe a substantial improvement when models are given English equivalents, with especially pronounced gains for Spanish (+19.1) and Chinese (+12.5). Our case studies highlight that LLMs are able to surface the correct background knowledge and relevant values more reliably for the math problems presented in English, leading to higher accuracy on these tasks. In contrast, for cultural reasoning, the difference is negligible on average. This suggests that while translation to English can unlock knowledge for certain tasks such as math reasoning questions, performance improvement on culturally grounded tasks remains limited, likely due to the lack of exposure of the models to specific cultural knowledge, regardless of language. We present case studies of examples where o3 answers English equivalents correctly but provides incorrect answers to the original questions in Appendix A.2, along with more quantified analysis on model discrepancies and consistency.

6 Discussion and Limitations

We analyze the errors of o3-2025-04-high and Gemini models and include representative examples in Table 6. We observe that 66% of the problems are either solved correctly or incorrectly by both models. Among the problems that the two models perform differently on, our error analysis suggests that o3 performs better on French Wordplay examples, where Gemini is more likely to miss cultural nuance or wordplay (yielding literal or surface-level responses). However, Gemini per-

Prompt	Gold Answer	Gemini Response	o3 Response
En un documento antiguo se hablaba de que la cosecha de uva en el Ribeiro fue de 50 codos de ribera cúbicos y que en la bodega había 11 bocoys de 40 arrobas. Si el rendimiento de la uva es de un 75%. Según la correspondencia de medidas recogida en Wikipedia, tomando 3 decimales para la arroba y 5 para el codo de ribera, ¿cuántos porrones de litro y medio hacen falta para guardar el vino que sobra?	13	15 porrones	9 872 porrones
Tu dois trouver dans cette conversation la subtilité française: -J'en ai marre du chantier, j'ai assez donné aujourd'hui. -Pareil, j'en peux plus... -On en a marre tôt on vient juste de commencer. -Tu as raison, qu'est-ce qu'on pourrait faire pour se motiver entre nous ? -Juste siffloter et attendre la fin de la journée.	marre tôt -> marteau	La subtilité est la proposition d'une solution passive et résignée ("siffloter et attendre") à un problème qui appelle une action motivante.	le jeu de mots « marre tôt » / « marteau »
快到中秋节了，家里已经准备好月饼和茶，期待当天跟家人一起赏月和提灯笼。今年是 2020 年后的第一个无春年，早上看了划龙舟比赛，下午还吃了肉粽。对了，从明天算起距离中秋节还有几天?	99 天	99 天	" 今年是 2020 年后的第一个无春年" 指的是...（2021 年）。... 今天是端午节...2021 年端午节... 是 6 月 14 日，所以" 明天" 是 2021-06-15。2021 年中秋节... 是 2021-09-21。计算... 天数差：6 月剩余 15 天（15 30 日）...7 月 31 天...8 月 31 天...9 月前 21 天...15 + 31 + 31 + 21 = 98 天。

Table 6: Representative native-language prompts and model outputs for o3 and Gemini models, illustrating both models' strengths and weaknesses. For the top example, both models fail to correctly retrieve the codo de ribera unit. In the second example, Gemini struggles to understand the French wordplay, while o3 incorrectly performs date calculations in the third row.

forms better on Chinese cultural prompts. The Spanish example in Table 6 is an example in which both models fail. Both models used incorrect values for the codo de ribera unit, leading to a difference in calculating the number of porrones required. In the French example shown, Gemini fails to recognize the homophonic pun between "marre tôt" and "marteau," while o3 correctly identifies and explains this wordplay. In the Chinese example, o3 undercounts the days between festivals by one.

These case studies highlight the sensitivity of large language models to linguistic and cultural nuances, especially when reasoning in languages other than English. In our work, we do not explore models designed or finetuned specifically to be multilingual, and we leave an evaluation of these models on the benchmark for future work. We selected French, Spanish, and Chinese as initial languages given the large number of global speakers and the authors' familiarity with these languages. To draw wider conclusions about multilingual performance, it will be be important to include additional languages in future work. We expect that extending this benchmark to low-resource languages would magnify the struggle of current models in dealing with local knowledge, linguistic phenomena, and complex reasoning. Furthermore, our analysis does not explicitly consider the effect of regional or dialectal variation, which is present in languages like Spanish and Chinese, and future work could examine how these differences impact model performance. This underscores the need for both more diverse training data and tailored model evaluation to ensure robust, equitable progress in multilingual LLMs.

7 Conclusion

We introduce MultiNRC, a challenging benchmark designed to evaluate the multilingual abilities of LLMs through linguistically and culturally-nuanced reasoning questions. By constructing diverse native reasoning questions in French, Spanish, and Chinese, we highlight ongoing challenges for LLMs when faced with linguistically and culturally nuanced reasoning tasks across multiple languages. Our evaluation shows that while English translation can boost performance on some tasks, substantial gaps remain for many culturally and linguistically nuanced questions. We hope that MultiNRC will serve as a robust testbed for future advances in multilingual LLM development, and we release our dataset to further research in this direction.

8 Limitations

While this work introduces a challenging and natively-sourced benchmark, several limitations should be acknowledged.

First, the scope of languages in MultiNRC is currently limited to French, Spanish, and Chinese. These languages were chosen based on the large global speaker population and the authors' familiarity. To formulate broader conclusions about multilingual reasoning, it will be crucial for future iterations to incorporate a wider variety of languages, especially low-resource ones, which are likely to amplify the observed challenges related to local knowledge and linguistic phenomena.

Second, the current analysis does not explicitly address the impact of regional or dialectal variations within the chosen languages, such as those present in Spanish and Chinese. Future work could explore how these differences affect model performance, offering a more granular understanding of multilingual capabilities.

Finally, the benchmark's construction methodology, which filters for questions that cause at least three out of five state-of-the-art models to fail, ensures a high degree of difficulty. However, this approach may not capture the full spectrum of multilingual reasoning abilities and could be expanded to include questions of varying difficulty to provide a more comprehensive assessment.

References

Meta AI. 2024. Llama 4: Advancing multimodal intelligence. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-06-21.

Anthropic. 2025. Claude 3.7 sonnet and claude code. *Anthropic News*. 5 min read.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. 2024. Arc prize 2024: Technical report. *arXiv preprint arXiv:2412.04604*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Google DeepMind. 2025. Gemini model and "thinking" updates: March 2025. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/. Accessed: 2025-06-21.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. Exams: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. *arXiv preprint arXiv:2011.03080*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Xu Huang, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. Benchmax: A comprehensive multilingual evaluation suite for large language models. *arXiv preprint arXiv:2502.07346*.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. Multiloko: a multilingual local knowledge benchmark for llms spanning 31 languages. *arXiv preprint arXiv:2504.10356*.

Bryan Li, Tamer Alkhouli, Daniele Bonadiman, Nikolaos Pappas, and Saab Mansour. 2024. Eliciting better multilingual structured reasoning from llms through code. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5154–5169.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Shrestha Basu Mallick and Logan Kilpatrick. 2025. Gemini 2.5: Updates to our family of thinking models. https://developers.googleblog.com/en/gemini-2-5-thinking-model-updates/. Accessed: 2025-06-23.

Mathematical Association of America. 2024. Aime 2024. https://www.maa.org/math-competitions/aime. Problem set and results.

OpenAI. 2024. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2024-06-21.

OpenAI. 2024. New models and developer products announced at spring update. Accessed: 2024-06-13.

OpenAI. 2025. Introducing o3 and o4 mini. https://openai.com/index/introducing-o3-and-o4-mini/. Accessed: 2025-06-21.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

William A. Scott. 1955. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, and 1 others. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

Hongyu Wang, Jiayu Xu, Senwei Xie, Ruiping Wang, Jialin Li, Zhaojie Xie, Bin Zhang, Chuyan Xiong, and Xilin Chen. 2024. M4u: Evaluating multilingual understanding and reasoning for large multimodal models. *arXiv preprint arXiv:2405.15638*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2025. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, Nan Liu, Qingyu Chen, Douglas Teodoro, Edison Marrese-Taylor, Shijian Lu, Yusuke Iwasawa, Yutaka Matsuo, and Irene Li. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *Preprint*, arXiv:2503.10497.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36:5484–5505.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. Agieval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314.

A Appendix

A.1 Dataset Examples and Linguistic Phenomena

Additional Samples Table 7 includes a broader sample of MultiNRC questions. For example, one French cultural question requires knowledge of both local festivals and the French calendar of saints. The French wordplay question relies on recognizing that "mon chien Michel" is a homophone for "Mont Saint-Michel," combining everyday language with a cultural reference.

We also include English-equivalent prompts in Table 8. For example, the French mathematical question asks for the difference in height between Paris's iconic monument and the tallest residential tower in the country, which are translated from French while keeping the logic and required knowledge for the question intact.

Cultural and Linguistic Nuances In order to analyze the presence of cultural and linguistic nuances in our prompts, we create categories of linguistic and cultural nuance using an LLM. Then, we label each prompt in our dataset with its associated nuances. The results of the most common linguistic and cultural phenomena present in the dataset are shown in Table 9. The categories of Puns, Homophones, & Wordplay are very common as this includes all examples in the Wordplay category as well as many under Linguistic reasoning. Similarly, Calendar & Time Systems are often present in Math as well as Cultural/Traditions reasoning problems.

A.2 Additional Evaluation Results

Language Consistency and Gaps As in (Hupkes and Bogoychev, 2025), we show the Mother Tongue Effect, calculated as the average performance in the original language minus the average performance in English over all questions. However, in contrast to (Hupkes and Bogoychev, 2025), here we note a negative effect on linguistically and culturally nuanced reasoning questions. We also include model consistency, as a measure of the percentage of questions that a model answers correctly in either original or En divided by the number of questions the model answers correctly in total. For most models, the consistency is around or below 50%, but we note that the top-performing models also have higher consistency, and that consistency has a strong rank correlation with original. The GAP column quantifies this difference between the best-performing language and the worst for each model, with an average gap of 13.90 points across all models. We observed that some models show especially large differences across languages; for example, Deepseek-R1-0528 scores 22.81 points higher in Chinese than in Spanish, and the original Deepseek-R1 a 29.10 difference. This underscores that even strong multilingual models perform very differently across languages.

Original vs En Examples We provide two examples where o3-2025-04-16-high incorrectly answers the question in French while correctly responding to the English translation in Table 11. In the first example, o3 misinterprets the value of a coin in French but correctly identifies it and computes and resulting answer in English. In the second example, a riddle about the color of a sash, o3 incorrectly determines the color sequence and predicts the wrong gender, while in English the model correctly understands the intended colors and predicts the right answer. These cases illustrate how the model may better surface and reason over certain cultural information in English, leading to improved performance.

Machine Translation and English Reasoning When comparing the reasoning settings in Table 12, we find a slight improvement with EnR on average, but this effect varies depending on the model. Furthermore, EnT/R does not offer an improvement on average. The combination of translation error due to native prompt misunderstanding as well as the presence of prompts in the wordplay and linguistic reasoning categories that cannot be directly translated likely contribute to error propagation. We show the delta between Original and En/R according to categories and languages in Table 13.

We note some improvement from reasoning in English for Chinese prompts, with Math, Linguistic, and Wordplay categories showing average deltas above +1.7. However, we see no improvements in the Cultural category. Even within a particular model, the benefit of reasoning in English often varies depending on both category and language. These results highlight the inherent differences in model families and the heterogenous impact of translation and English reasoning, reinforcing the need for language-specific and model-specific analysis in multilingual benchmarks. .

Language	Category	Prompt	Answer
Chinese	Math	我正在写小说，这是小说里的一个场景：我参加外婆的寿宴时，外婆很自豪地说：“算起来，这年是我第九轮的第一年。我可是民国二十年生人。”。我妈说：”妈，时间过得真快，我也年过花甲了。“这时小姨插嘴道：”二姐，我比你小一岁。我是属马的。“我知道这里提到的都是虚岁，这时候我弟弟就问我，外婆，妈妈和小姨的到底实岁是几岁？外婆和妈妈相差几岁？	1
Spanish	Math	Mi jefa quedó engatusada con la oferta de un vendedor de refris en Tijuana, y quiero hacer que agarre el rollo. Se me hace que se la quieren fregar, pero nomás no le cae el veinte. Échame la mano con los números. El refri cuesta 20 milagritos ya con IVA. También que su pago mensual por 12 meses será igual al pago original del IVA. Este bato dice que le van a regresar el IVA del precio final que haya pagado. ¿Cuál es el resultado del total que pagó por el refri menos lo que le regresen?	$16,355.56
Chinese	Math	我最小的妹妹是 2010 年后出生的，出生在一个双春年的中秋节。我们家非常重视中秋节，所以虽然她出生证明上的日期是阳历生日，但她的生日派对每年都在中秋节这一天举行的。今年是闰六月年份，根据阳历计算，今年的中秋节比她出生那天的中秋节晚了一个数字 X，这个数字 X 刚好等于今年和她出生年份之间的相差年数。请问，我妹妹出生于几年几月几号，今年几岁了？	2020 年 10 月 1 日；5 岁
Chinese	Cultural	我姐夫的家乡在潮汕，当时他和姐姐的婚礼定在了当地俗称“天日”的这天，希望可以带来好兆头。我依稀记得他们结婚那年是长辈说的辛丑年，如果我和男朋友 2024 年也在相同的公历日期结婚，你知道确切的日子吗？	2024 年 2 月 20 日
French	Cultural	En 2025, comme tous les ans, les Fêtes de la Madeleine se dérouleront à Mont-de-Marsan. Quel prénom est célébré le dernier jour des Fêtes de la Madeleine selon le calendrier des saints en France ?	Marina
French	Wordplay	Explique cette devinette : Pourquoi un Normand et un Breton se disputent-ils ? Pour savoir à qui appartient mon chien Michel.	”mon chien Michel” sonne comme ”Mont Saint-Michel”.
Spanish	Wordplay	Mi compañera estaba riendo feliz mientras yo, tristemente, sacudía los muebles. ¿Qué hace mi compañera?	Barriendo.
Spanish	Linguistic	Estoy observando desde la puerta de una casa abandonada a la noche algunas cosas que están a mi lado, otras a unos pasos más alejadas y otros a varios metros de distancia: Aquello está un poco marchito. Esta cosa le falta un ojo. Esos tienen un mango Aquella contiene múltiples herramientas. Aquí tengo un juguete que tiene ruedas. Al día siguiente, con mayor iluminación volví a observar desde la puerta qué elementos hay. Esto es un tren Allá están las flores. Ahí está el coso para golpear. Eso es para cavar. Al lado mío está un muñeco de trapo. Allá está la caja. Al tercer día, un fantasma movio las cosas: Aquello lo trajo acá. Lo de ahí fue allá. Indicar cuantos objetos hay acá, ahí y allá.	4 acá, 0 ahí y 2 allá.

Table 7: Additional native-language example and ground-truth answers from MultiNRC.

Language	Category	Prompt	Answer	English Prompt	English GTFA
French	Cultural Reasoning	Je suis agent chargé de la tranquillité publique à Coin-sur-Seille et il me reste 4 jours de congés à poser en 2025. J'aimerais profiter de toute la semaine qui inclut le Vendredi saint. Est-ce que c'est envisageable avec le solde qui me reste ?	Non.	I work as a public safety officer in Coin-sur-Seille and only have 4 days of leave left to take. Id like to take off the entire week that includes Good Friday. Is this possible with my remaining balance?	No, this is not an option.
French	Math	Quelle est la différence entre la hauteur du monument emblématique de Paris et la plus haute tour habitable du pays?	99 mètres	What is the difference between the height of Paris' iconic monument and the tallest residential tower in the country?	The difference between the two buildings is 99 meters
Spanish	Cultural Reasoning	Mi abuelo nació justamente 10 años antes de que Argentina retomara las Islas Malvinas, yo nací 30 años después el Día del Veterano y de los Caídos en la guerra de Malvinas ¿En qué día de la semana nací yo?	Martes	My grandfather was born exactly 10 years before Argentina recovered the Malvinas (Falkland) Islands, and I was born 30 years later on the Day of the Veteran and Fallen in the Malvinas War (Día de los Veteranos y Caicos in las Guerra de Malvinas). On what day of the week was I born?	Tuesday
Spanish	Math	Mis compas y yo somos futboleros de corazón. No hay partido de la Selección Nacional que dejemos de ver. Uno de mis amigos, Carlos, es 3 años menor que yo. El otro, Santi, es 2 años menor que Carlos y Fer, es 1 año mayor que yo. Si yo nací 18 años antes del segundo año de la temporada de Miguel Mejía Barón, como entrenador de la Selección, ¿cuántos años tenía Santi en julio del año pasado, 2024, si todos nacimos en el mes de agosto?	42 años.	My friends and I are diehard soccer fans. There isn' t a match of the National Team that we miss watching. One of my friends, Carlos, is three years younger than me. The other one, Santi, is two years younger than Carlos, and Fer is one year older than me. If I was born eighteen years before the second year of Miguel Mejía Barón' s tenure as coach of the National Team, how old was Santi in July of last year, 2024, if we were all born in August?,	42 years old.
Chinese	Cultural Reasoning	有一位清朝皇帝是首位出生和成长在北京的皇帝。请问他的最后一位皇后当了多久皇后就过世了？	1 天	There was a Qing dynasty emperor who was the first to be born and raised in Beijing. How long did his last empress serve as empress before she died?	One day.
Chinese	Math	据说关羽身高 9 尺的青龙偃月刀重 82 斤，请问是现在的多高，刀又有多少公斤？	217 厘米（据找到的资料一尺一般在 23.1 厘米到 23.3 厘米之间），18.04 公斤	It is said that Guan Yu was 9 chi tall and his Green Dragon Crescent Blade weighed 82 jin. Could you tell me how tall that is in modern measurements, and how many kilograms the blade weighed?	Guan Yu was approximately 2.08 meters tall, and his Green Dragon Crescent Blade weighed approximately 18.04 kilograms.

Table 8: Representative native-language prompts and answers from MultiNRC, with English prompt and reference answer.

Linguistic Nuances		Cultural Nuances	
Category	**Count**	**Category**	**Count**
Puns, Homophones & Wordplay	454	Calendar & Time Systems	324
Polysemy, Homonymy & Ambiguity	317	Specific Units & Currency	222
Discourse Inference & Pragmatics	167	Kinship & Social Hierarchies	128
Slang & Colloquial Language	165	Cross-System or Interlingual Reasoning	107
Idioms, Proverbs & Figurative Language	139	Institutional, Administrative, or Legal	80
Lexical Gaps or Untranslatables	126	Embedded or Hidden-Form Solutions	51
Gendered Language & Agreement	58	Game Rules & Contextual Calculation	42
Formal Register & Politeness	31		

Table 9: Clusters of top categories of linguistic and cultural nuances present in MultiNRC prompts.

Model	MTE	Consistency	GAP
o3-pro	-4.66	61.98	15.12
o3-2025-04-16-high	-1.90	52.04	14.04
Gemini-2.5-Pro-Preview-06-05	-0.34	54.47	12.19
o3-2025-04-16-medium	-2.93	47.68	11.00
Claude-4-Opus-20250514-thinking	-3.79	37.28	22.99
Claude-4-Opus-20250514	-3.28	33.46	15.57
Claude-3.7-Sonnet-thinking	-4.14	34.19	7.97
Deepseek-R1-0528	-14.66	30.42	22.81
Deepseek-R1	-3.97	31.51	29.10
o4-mini-high	-9.48	27.75	10.53
GPT-4.1	-6.21	31.82	14.03
Claude-4-Sonnet-20250514	-3.45	23.40	10.81
GPT-4o	-7.93	17.79	3.36
Llama-4-Maverick	-8.28	12.41	5.04
Average	-5.36	35.44	13.90

Table 10: MTE (Mother Tongue Effect: average difference Original − English), Consistency (proportion correct in both settings), and GAP (difference between highest and lowest original language score) for each model. Note that GAP is calculated over all datapoints, while MTE and Consistency over those with English translations.

Prompt	GTFA	English Prompt	o3 original Answer	o3 English Answer
Je crois que j'ai perdu de l'argent de ma poche aujourd'hui. Je me rappelle qu'avant de sortir j'avais 4 billets verts et une pièce argentée avec contour doré, mais en rentrant j'ai trouvé dans ma poche uniquement la somme de 200. Quelle somme aurais-je perdu ?	201 euros	I think I lost some money from my pocket today. I remember that before I went out, I had 4 green bills and one silver coin with a golden contour, but when I came back, I only found 200 in my pocket. How much money would I have lost?	**Explication:** ... La « pièce argentée avec contour doré » correspond à la pièce bicolore de 2 €. ... **Exact Answer:** 202 €	**English Explanation:** In euro currency: • The only bill that is predominantly green is the €100 note.... • A coin that is silver-coloured ... with a golden ring around it matches the €1 coin. **Exact Answer:** €201
Ma femme est allée seule à l'échographie de notre second enfant permettant de savoir si c'est un garçon ou une fille...Je n'étais pas disponible. Elle a choisi de me le faire deviner sous forme d'une énigme basée sur les couleurs : 1. Je suis la couleur la plus proche du col de l'écharpe tricolore d'un parlementaire lorsqu'elle est portée en écharpe. 2. Je suis la couleur du ruban de l'insigne d'un chevalier de la Légion d'honneur. 3. Si les couleurs sont différentes, alors c'est un garçon. Si elles sont identiques, alors c'est une fille. Alors, verdict ?	Une fille (a girl)	My wife went alone to the ultrasound of our second child to find out if it's a boy or a girl... I wasn't available. She chose to have me guess in the form of a riddle based on colors: 1. I am the color closest to the collar of the tricolor sash of a parliamentarian when it is worn as a sash. 2. I am the color of the ribbon of the insignia of a Knight of the Legion of Honor. 3. If the colors are different, then it's a boy. If they are identical, then it's a girl. So, verdict?	... l'ordre des bandes est bleu-blanc-rouge ; la couleur la plus proche du col est donc le bleu. Le ruban de l'insigne de chevalier de la Légion d'honneur est entièrement rouge. Comme les deux couleurs obtenues ... ne sont pas identiques, la règle donnée ... le bébé est un garçon.	1. When a French parliamentarian ... wears the tricolour sash, the stripe that sits nearest the collar is red. 2. The ribbon from which the insignia of a Knight (Chevalier) of the Légion d'honneur is suspended is entirely red. Both requested colours are therefore red. Because your wife's rule ... the identical reds mean the baby is a girl.

Table 11: Examples where o3 incorrectly answers the French prompt but correctly answer the English-translated problem.

Model	Orig	EnT/R	EnR
o3-pro-2025-06-10-high	**49.00**	41.71	**45.97**
o3-2025-04-16-high	**45.50**	**45.02**	**46.54**
Gemini-2.5-Pro-Preview-06-05	45.12	44.08	45.50
o3-2025-04-16-medium	44.45	**45.21**	45.31
Claude-4-Opus-20250514-thinking	33.93	32.99	33.27
Claude-4-Opus-20250514	29.00	26.26	26.82
Claude-3.7-Sonnet-thinking	27.77	25.02	26.54
Deepseek-R1-0528	27.58	24.36	28.91
Deepseek-R1	24.27	21.99	24.17
o4-mini-high	22.18	24.64	25.21
GPT-4.1	21.23	22.65	22.09
Claude-4-Sonnet-20250514	18.39	16.68	17.91
GPT-4o	12.42	12.23	11.56
Llama-4-Maverick	8.44	9.48	10.81
Average	31.45	30.38	31.67

Table 12: Accuracy (%) of models under different multilingual reasoning settings: **Original** (reasoning in the original language), **EnT/R** (translate prompt to English, then reason in English), and **EnR** (reasoning in English).

Model	Cultural			Math			Linguistic			Wordplay		
	ΔFr	ΔEs	ΔZh	ΔFr	ΔEs	ΔZh	ΔFr	ΔEs	ΔZh	ΔFr	ΔEs	ΔZh
o3-pro	0.9	-1.6	**-4.4**	**-6.5**	-0.9	0.0	**-10.7**	-7.0	-4.2	-2.5	**-5.4**	2.3
o3-2025-04-16-high	-0.9	0.8	**5.6**	**-6.5**	3.8	5.1	4.0	**-5.6**	**5.6**	0.0	-7.5	**8.1**
Gemini-2.5-Pro-Preview-06-05	0.0	2.5	**-4.4**	5.2	**-7.5**	1.3	1.3	**2.8**	1.4	-1.3	2.2	**3.5**
o3-2025-04-16-medium	2.8	-1.6	**-8.9**	3.9	**4.7**	1.3	2.7	1.4	**11.3**	**-3.8**	0.0	-1.2
Claude-4-Opus-20250514-thinking	2.8	-1.6	**-3.3**	3.9	**5.7**	1.3	4.0	-5.6	**-7.0**	**-7.6**	-3.2	0.0
Claude-4-Opus-20250514	-3.7	**-9.8**	-1.1	**-5.2**	0.0	-3.8	**-6.7**	4.2	4.2	**2.5**	-2.2	0.0
Claude-3.7-Sonnet-thinking	-4.7	-2.5	**6.7**	-2.6	-3.8	**5.1**	**-13.3**	1.4	2.8	**2.5**	-4.3	0.0
Deepseek-R1-0528	0.9	**4.1**	-3.3	**7.8**	1.9	3.8	-2.7	**-5.6**	2.8	**8.9**	-1.1	-2.3
Deepseek-R1	-2.8	-2.5	**-5.6**	-6.5	0.9	**7.7**	5.3	**9.9**	-7.0	**5.1**	-3.2	1.2
o4-mini-high	3.7	2.5	**5.6**	**2.6**	0.0	1.3	0.0	-1.4	**5.6**	**12.7**	-5.4	10.5
GPT-4.1	1.9	-0.8	**3.3**	-1.3	**2.8**	-1.3	**-5.3**	**8.5**	4.2	**-6.3**	1.1	3.5
Claude-4-Sonnet-20250514	-0.9	**-1.6**	-1.1	0.0	**-8.5**	-2.6	0.0	4.2	**5.6**	-3.8	**5.4**	1.2
GPT-4o	0.0	2.5	**-5.6**	**3.9**	0.0	1.3	**-4.0**	0.0	-2.8	-3.8	1.1	**-4.7**
Llama-4-Maverick	0.0	-2.5	**8.9**	1.3	0.9	**5.1**	0.0	4.2	**5.6**	0.0	**5.4**	2.3
Average	**0.0**	**-0.9**	**-0.6**	**0.0**	**0.0**	**1.8**	**-1.8**	**0.8**	**2.0**	**0.2**	**-1.2**	**1.7**

Table 13: Delta (EnR − Original) in accuracy (%) for each reasoning category and language. **Bold** indicates the value with largest magnitude in each category per model. Positive values indicate a performance improvement on queries when reasoning in English vs Original.