IRLBench: A Multi-modal, Culturally Grounded, Parallel Irish-English Benchmark for Open-Ended LLM Reasoning Evaluation

Khanh-Tung Tran
University College Cork
Cork, Ireland
123128577@umail.ucc.ie

Duc-Hai Nguyen
University College Cork
Cork, Ireland
125109073@umail.ucc.ie

Barry O'Sullivan
University College Cork
Cork, Ireland
b.osullivan@cs.ucc.ie

Hoang D. Nguyen
University College Cork
Cork, Ireland
hn@cs.ucc.ie

Abstract

Recent advances in Large Language Models (LLMs) have demonstrated promising capabilities, yet their performance in multilingual and low-resource settings remains modest. Existing benchmarks often exhibit cultural bias, restrict evaluation to text-only, rely on multiple-choice formats, and, more importantly, are ineffectual for extremely low-resource languages. To address these gaps, we introduce IRLBench, presented in parallel English and Irish, which is considered definitely endangered by UNESCO. Our benchmark consists of 12 representative subjects developed from the 2024 Irish Leaving Certificate exam, enabling fine-grained analysis of model capabilities across domains. By framing the task as long-form generation and leveraging the official marking scheme, it supports not only a comprehensive evaluation of correctness but also language fidelity. Our extensive experiments of leading closed-source and open-source LLMs reveal a persistent performance gap between English and Irish, in which models produce valid Irish responses less than 80% of the time, and answer correctly 55.8% of the time compared to 76.2% in English for the best-performing model. With Irish as the case study, our work exposes systemic weaknesses in today's multilingual LLMs and provides a rigorous benchmark for evaluating true multilingual capabilities. We release IRLBench[1] and an accompanying evaluation codebase[2] to enable future research on robust, culturally aware multilingual AI development.

CCS Concepts

• **Computing methodologies** → **Natural language processing**; **Natural language generation**; **Language resources**.

[1]https://huggingface.co/datasets/ReliableAI/IRLBench
[2]https://github.com/ReML-AI/IRLBench

KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2258-5/2026/08
https://doi.org/10.1145/3770854.3785693

Keywords

Large Language Model, Large Vision-Language Model, Extremely Low-resource, Benchmarking

ACM Reference Format:
Khanh-Tung Tran, Duc-Hai Nguyen, Barry O'Sullivan, and Hoang D. Nguyen. 2026. IRLBench: A Multi-modal, Culturally Grounded, Parallel Irish-English Benchmark for Open-Ended LLM Reasoning Evaluation. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3770854.3785693

1 Introduction

Large Language Models (LLMs) and their variants, including Vision–Language Models (VLMs) and Large Reasoning Models (LRMs), have recently achieved remarkable performance [5, 14], reshaping the field of natural language processing and rapidly overtaking traditional benchmarks [44]. As capabilities have advanced, existing reasoning benchmarks, such as mathematical and STEM reasoning, once considered challenging, have become saturated, prompting the community to develop more rigorous and advanced evaluations to probe the limits of current models [30, 36]. However, these benchmarking efforts predominantly focus on resource-rich and dominant languages such as English. Consequently, performance in other languages remains poorly understood, and the unique challenges of low-resource settings remain under-explored [9, 39].

Multilingual benchmarks that do exist often suffer from several shortcomings. They may embed cultural bias (e.g., U.S.-centric names, currency, and scenarios), limit themselves to text-only questions, rely on multiple-choice formats, or omit extremely low-resourced languages. This leads to known-limitations such as memorization problem, unable to test language generation capability in multi-choice benchmark [26, 32, 37], and unable to expose weaknesses of recent frontier models, as shown in Figure 1 [7, 12, 41]. Moreover, the translation of English-centric benchmarks into other languages often proves insufficient, as the resulting tasks may not correlate effectively with local human preferences and knowledge contexts [2, 44]. A key example of such a gap is Irish, classified as definitely endangered by UNESCO [40]. Irish currently lacks detailed, open-ended evaluation datasets. Existing Irish datasets

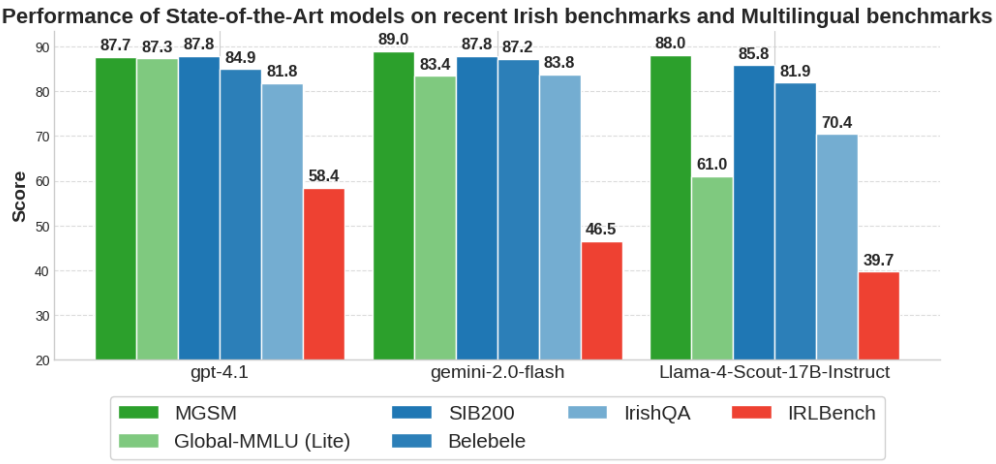

Figure 1: Performance of State-of-the-Art models on recent Irish benchmarks and Multilingual benchmarks compared to IRLBench. While existing benchmarks are saturated, unable to show meaningful differences between models, IRLBench illustrates a significant drop in performance, filling the gap for true multilingual benchmarks. We benchmark more models on IRLBench in Section 4. Results on the MGSM dataset are taken from [41], and results on the Global-MMLU (Lite) dataset are taken from [7, 12], except for Llama-4-Scout. The rests are reproduced by us, following the default evaluation hyperparameters.

(e.g., SIB200 [1], Belebele [4], IrishQA [38]) focus on more traditional tasks (e.g., topic classification), and are quickly saturated by state-of-the-art models, with accuracy rates approaching 90%, leaving little room for meaningful comparison or deeper analysis of multilingual transfer, as shown in Figure 1.

To address these issues, we introduce IRLBench, a novel multi-modal, culturally grounded, parallel English–Irish benchmark for open-ended evaluation. Inspired by recent studies leveraging educational materials as benchmarks [29], IRLBench is derived directly from the 2024 Irish Leaving Certificate exams, which stand as a significant and consequential gateway to individual higher education and national development. It comprises twelve representative subjects grouped into four key domains: Science, Applied Science, Business Studies, and Social Studies. We balance diversity and comparability across domains, yielding a varied pool for the subsequent automated extraction step. To construct the dataset, we utilized official exam papers and corresponding marking schemes in PDF format, employing an automated vision-language pipeline to extract and format content efficiently. More importantly, a final human verification step ensures high data quality and accuracy. This choice aligns with findings from visually-rich document understanding: messy layouts (tables, multi-column text, nested fields) remain challenging for VLMs [42]. Because each exam is available in parallel English and Irish versions, with identical content, IRLBench naturally supports direct multilingual comparisons in high-quality. By formulating each question as a long-form, open-ended generation task aligned with official marking criteria, the benchmark enables comprehensive evaluation of factual correctness, domain-specific academic knowledge, linguistic fidelity, and depth of reasoning simultaneously in both languages (Figure 2). Furthermore, IRLBench includes multi-modal questions, requiring models to interpret figures, tables, and diagrams alongside textual descriptions, as shown in Figure 2, where a Chemistry question necessitates understanding of an associated diagram depicting a chemical process. Our deliberate focus on a single low-resource language ensures cultural authenticity. Research shows that localized benchmarks align better with human preferences than translated benchmarks [44]. While our pipeline centers on Irish as a case study, its language-agnostic construction nature offers a replicable blueprint for other endangered or low-resource languages communities to develop their own culturally grounded benchmarks [46], thereby extending its value to the wider multilingual-LLM community.

We conduct extensive experiments on leading closed-source (e.g., gpt-4.1, gemini-2.0-flash) and open-source (e.g., Llama-4-Scout-Instruct) models. Using LLM-as-a-judge for automatic evaluation [45, 48], our results reveal substantial performance gaps between English and Irish ($p=0.004$, effect sizes up to $d>1.0$). While human students perform equivalently in both languages, the best LLM achieves only 55.8% accuracy in Irish versus 76.2% in English. We uncover fundamental weaknesses: poor confidence calibration ($r<0.25$), multimodal degradation particularly in Irish, and persistent gaps despite translation-based agentic workflows. These findings underscore critical challenges in achieving true multilingual transfer for open-ended reasoning in extremely low-resource languages.

Our key contributions are as follows:

- **IRLBench**. We introduce IRLBench, a parallel Irish-English, multi-modal, open-ended academic reasoning dataset drawn from the 2024 Irish Leaving Certificate.
- **Comprehensive Evaluation**. We systematically evaluate a range of closed and open LLMs, analyzing their performance across language (English vs. Irish) and domain (Science, Applied Science, Business Studies, and Social Studies).
- **Insights on Multilingual Transfer**. Our experiments reveal statistically robust English-Irish performance gaps (up to 20% for SoTA models, bootstrap-confirmed with effect

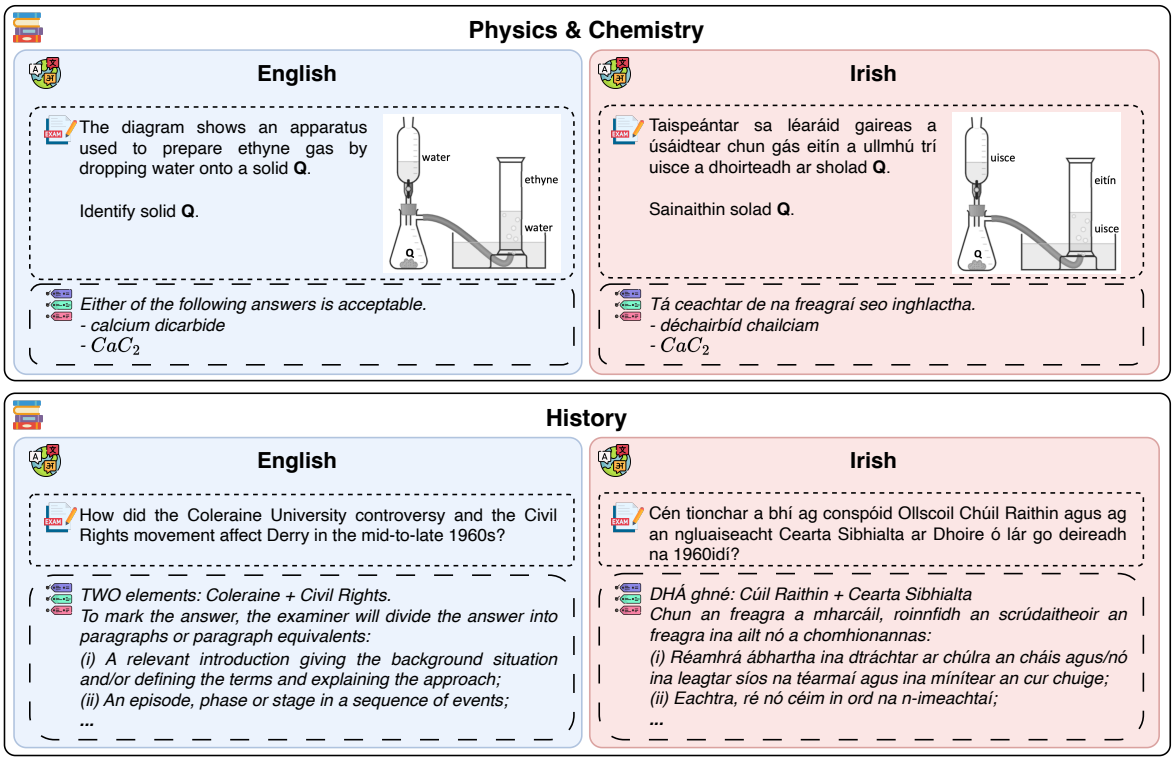

Figure 2: Illustrative examples showcasing the multi-modal, Irish-English parallel nature and the diverse tasks of our proposed IRLBench dataset.

sizes d>1.0). We quantify language fidelity, confidence miscalibration (r<0.25), and demonstrate multimodal degradation in Irish. We evaluate translation-based agentic approaches and contrast LLM limitations with equivalent human performance, providing comprehensive directions for low-resource reasoning research.

- **Public Benchmark Release**. Last but not least, we publish the complete IRLBench dataset alongside open-source evaluation code, enabling researchers to test new models or replicate our findings with minimal setup.

2 Related Works

Multilingual reasoning benchmark. Recent multilingual reasoning datasets have attempted to evaluate LLMs across multiple languages. For instance, MGSM [33], a multilingual variant of GSM8K [6], provides human-translated arithmetic word problems in 10 languages, including two extremely low-resource languages. However, these problems predominantly reflect U.S.-centric contexts e.g., American names, dollar-denominated currency, and everyday scenarios, introducing potential cultural biases [37]. Another notable benchmark, MMLU (Massive Multitask Language Understanding) [13], has inspired multilingual derivatives: MMMLU[3], a professionally translated set of MMLU questions across 14 languages, and Global_MMLU [34], which highlights cultural biases by identifying that approximately 28% of original MMLU questions heavily rely on Western-centric concepts. To mitigate this, Global_MMLU categorizes its dataset into culturally sensitive and culturally agnostic subsets across 42 languages. Further, there have been efforts to create language-specific variants of MMLU using local exam questions and educational materials. Examples include CMMLU (Chinese)[18], KMMLU (Korean)[35], and ArabicMMLU (Arabic)[17]. Recently, M3Exam addressed cultural bias by collecting real-world exam questions from various countries, but its non-parallel structure across languages limits fair cross-lingual comparisons. Moreover, all these aforementioned benchmarks primarily use multiple-choice or short-answer questions, testing factual recall rather than in-depth generative reasoning. Such tasks risk memorization biases and fail to rigorously assess the language generation capabilities of models[31].

Recent English-centric datasets, including LLM4DyG [47], ST-Bench [20], and Humanity's Last Exam [30], have transitioned towards open-ended, free-form generative evaluation. However, there is still a need for such datasets for extremely low-resource languages. Our work extends this direction by introducing IRLBench, the first multilingual, culturally grounded, parallel benchmark specifically designed for open-ended generative evaluation in an extremely low-resource language scenario, Irish. Table 1 provides a comparative summary highlighting IRLBench's unique features relative to existing multilingual benchmarks.

[3]https://huggingface.co/datasets/openai/MMMLU

Table 1: Comparison of IRLBench and other related low-resource scenario benchmarks.

Dataset	Multilingual	Parallel	Culture-aware	Multi-task	Multi-modal	Evaluation
MGSM [33]	✓	✓	✗	✗	✗	Multi-choice
Global_MMLU [34]	✓	✓	✓	✓	✗	Multi-choice
MMMLU	✓	✓	✗	✓	✗	Multi-choice
M3Exam [46]	✓	✗	✓	✓	✓	Multi-choice
HLE [30]	✗	✗	✗	✓	✓	Open-ended
IRLBench (ours)	✓	✓	✓	✓	✓	Open-ended

Benchmarks for the Irish Language. Irish, classified as definitely endangered by UNESCO, significantly lacks resources for AI and natural language processing development. Current datasets, such as IrishQA [38] or multilingual benchmarks including SIB200 [1] and Belebele [1], primarily target language understanding tasks such as topic classification. These datasets have been rapidly saturated by SoTA LLMs, achieving accuracies approaching 90% (as shown in Figure 1), which limits their effectiveness in differentiating model capabilities and analyzing deeper aspects of multilingual reasoning and transfer. IRLBench aims to fill this critical gap by providing challenging, multimodal, open-ended tasks that comprehensively evaluate the generative reasoning abilities of models in both Irish and English.

3 IRLBench

In this section, we provide a detailed description of IRLBench, covering our data collection methodology, dataset composition, and benchmarking task designed to evaluate SoTA LLMs. First, we outline our automated extraction pipeline and subsequent human verification process. Next, we present the dataset's composition, organized by subjects and subject groups. Finally, we discuss our proposed evaluation procedure, highlighting how IRLBench effectively assesses models' open-ended generative abilities in extremely low-resource multilingual scenarios.

3.1 Data Collection

The Irish Leaving Certificate is Ireland's final secondary-school system examination, which also serves as the standard university matriculation examination, typically requiring students at least two years of preparation[4]. Exams are available in parallel Irish and English versions, accompanied by standardized marking schemes, and cover five broad subject groups: Languages, Natural Sciences, Applied Sciences, Social Sciences, and Business Studies. Each subject can be examined at one of three difficulty levels: Higher (informally called Honours), Ordinary (Pass), or Foundation, with Higher Level exams posing the most difficult challenge. To rigorously evaluate SoTA LLMs, IRLBench exclusively includes Higher Level exams. Additionally, to facilitate direct performance comparisons between English and Irish and to avoid confounding effects from language-specific content, we exclude the Languages subject group, which is only available in one language. From each remaining group, we select three representative subjects, forming a diverse set of 12 subjects. This cross-subject design ensures broad domain coverage across Natural Sciences, Applied Sciences, Social Sciences, and Business Studies. Each subject's official marking scheme enables precise and structured evaluation. The structure and characteristics of the selected exams are illustrated in the upper portion of Figure

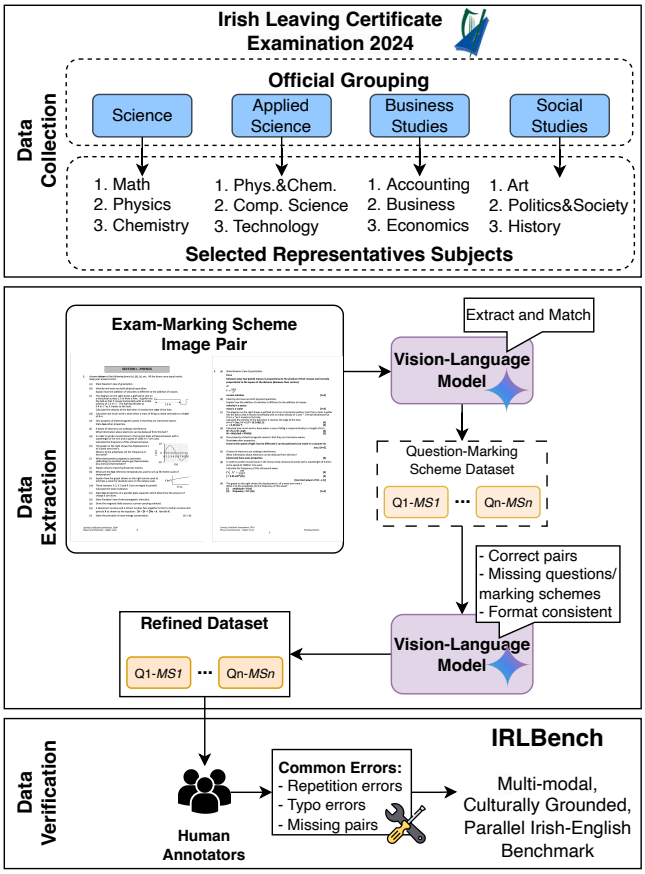

Figure 3: An overview of the IRLBench data collection pipeline, solving key challenges such as visual and parallel document understanding. Notably, we employ a human verification step at the end to ensure the high quality of IRLBench.

[4]https://en.wikipedia.org/wiki/Leaving_Certificate_%28Ireland%29

3, and example question-marking scheme pairs are provided in Figure 2, demonstrating the multi-modal, parallel, and diverse nature of IRLBench.

The original exam papers and marking schemes are publicly available as PDFs containing complex layouts with embedded tables, figures, and equations[5]. Given the challenges in accurately converting these documents into editable formats (e.g., Markdown or Word), we process them directly as images. To achieve efficient and accurate extraction, we leverage the frontier VLM gemini-2.0-flash, chosen for its strong performance in agentic tasks and cost-effectiveness [10, 11]. Specifically, the model extracts pairs of exam questions and corresponding marking schemes from the PDF images, aligning textual and graphical elements, as illustrated in the lower part of Figure 3. For reproducibility, the complete prompts used to guide the vision-language model in this complex data extraction process are provided in our supplementary materials.

The initial extraction result reveal several systematic issues, including incorrect question-marking scheme pairings, incomplete graphic extraction, and inconsistent formatting (in table and equations). To address these challenges before human annotation, we implement an automated verification step using the same VLM. This step significantly reduces systemic errors by detecting misalignments and extraction inaccuracies. Then, human annotators are only involved in a final verification phase, manually resolving the three most common remaining issues: duplicated outputs (primarily from lengthy Irish text segments), typographical mistakes, and missing question-scheme pairs. The resulting benchmark, IRLBench, is a multi-modal, culturally grounded dataset with parallel Irish-English samples, enabling detailed evaluation of multilingual academic reasoning capabilities.

3.2 Data Statistics

IRLBench consists of 1,700 samples, each containing a pair of question and marking scheme developed from the 2024 Irish Leaving Certificate exams. Samples are equally distributed between English and Irish, with parallel content for comparative evaluation. The 1,700 samples span twelve subjects (three per domain), ensuring cross-domain coverage; and 18.11% of samples are multi-modal, further diversifies task types.

Table 2 summarizes the dataset distribution across four subject groups: Science, Applied Science, Business Studies, and Social Studies. The Applied Science group has the highest number of samples (590), primarily from Physics and Chemistry (370). Science follows closely with 582 samples, with Chemistry (256) and Physics (216) comprising the majority. Business Studies contains 326 samples, evenly split between Business and Economics, but with fewer samples from Accounting. Social Studies is the smallest group, but still consists of 202 samples, distributed among History (106), Politics and Society (56), and Art (40).

Figure 4 visualizes the proportional distribution of these samples, showing the dominance of Applied Science and Science, highlighting their substantial representation within IRLBench. This structured diversity allows fine-grained analyses across subjects and language modalities, providing a robust foundation for evaluating multilingual and multi-modal generative capabilities.

[5]https://www.studyclix.ie

Table 2: Amount of samples collected for IRLBench. Each sample is a pair of question and marking scheme, including both text and illustrations in the original Leaving Certificate exam. 18.11 % of the samples include graphical illustrations, and the amount of samples are equally split between English and Irish, with the same task contents.

Group	Subject	Samples
Science	Math	110
	Physics	216
	Chemistry	256
	Group Total	*582*
Business Studies	Business	132
	Accounting	62
	Economics	132
	Group Total	*326*
Applied Science	Physics and Chemistry	370
	Computer Science	100
	Technology	120
	Group Total	*590*
Social Studies	Art	40
	Politics and Society	56
	History	106
	Group Total	*202*
Total Samples		**1700**

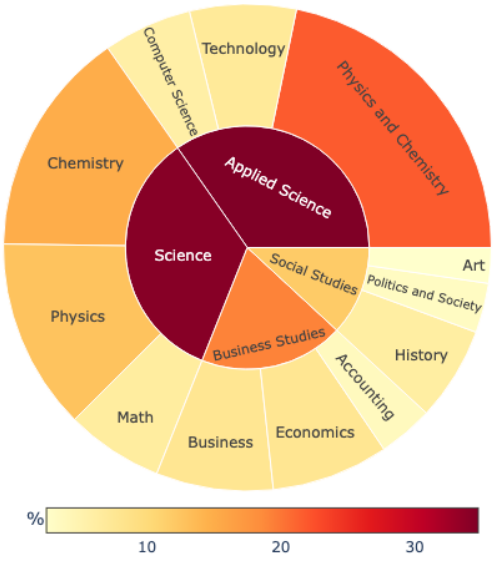

Figure 4: Distributions of subject groups and subjects as percentages of total amount of samples in IRLBench.

3.3 Benchmarking Task

Evaluating open-ended generation tasks poses significant challenges, as answers that are correct often vary widely in wording and

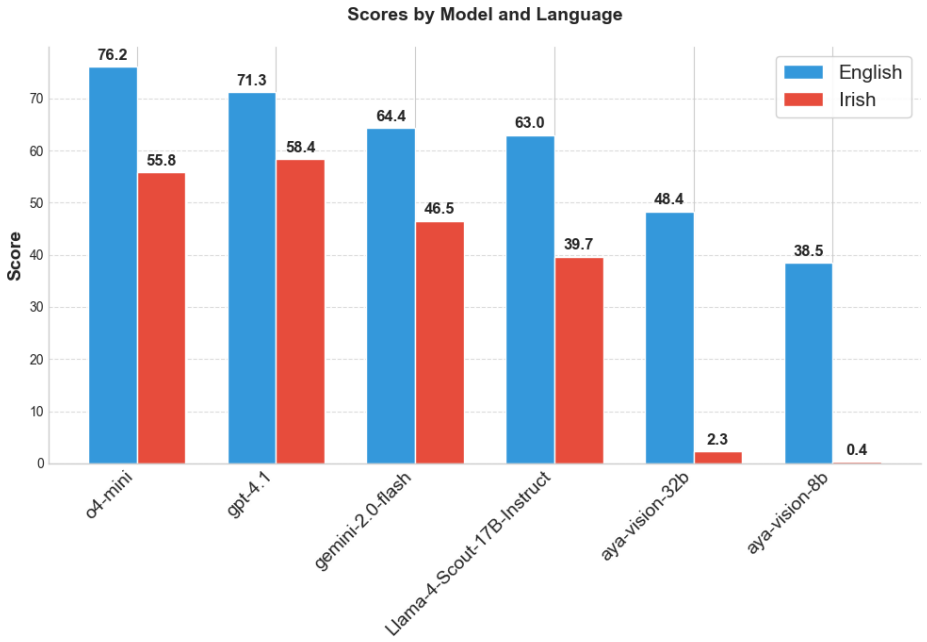

Figure 5: Accuracy scores on IRLBench per model and language. The results highlight a clear gap between English and Irish, and that open-ended reasoning remains substantially more challenging in extremely low-resource languages such as Irish.

detail, making traditional metrics like BLEU [27] and ROUGE [21] inadequate. Recent works [45, 48, 49] address these limitations by leveraging instruction-tuned LLMs as automated evaluators, utilizing their inherent ability to assess natural language outputs in context. Following this approach of LLM-as-a-judge, we employ an evaluation strategy combined with official marking schemes to reliably assess performance on IRLBench. We select gemini-2.5-flash [10] as our evaluation model due to its strong performance in instruction-following and reasoning tasks. For each evaluation, the judge model receives a prompt containing the original exam question, its official marking scheme, and the candidate model's generated response. The judge then assigns a binary label: correct or incorrect, simplifying the evaluation task and reducing the inherent variability associated with LLM-based evaluation. The complete prompts used to elicit model answers and guide the judge, along with detailed examples of correct and incorrect evaluation decisions, are provided in our supplementary materials. Although this binary judgment deviates somewhat from the granular scoring outlined in official marking schemes, it ensures consistent and comparable performance metrics across benchmarked models.

In addition to evaluating correctness, we assess the capability of models to generate text in Irish, an extremely low-resource language. To this end, we use a FastText-based language identification model [15, 16] to determine the language of responses generated for Irish-language questions. As the language detection model operates at the sentence level, we split generated answers into individual sentences. We empirically tuned the decision threshold and found 50% to perform most accurately. If more than 50% of sentences are detected as English rather than Irish, we classify the entire response as non-Irish. This criterion provides an estimate of language fidelity, emphasizing models' practical ability to handle extremely low-resource multilingual generation tasks.

Finally, we benchmark and compare prominent SoTA VLMs. This includes leading closed-source commercial models such as o4-mini [25], gpt-4.1 [24], and gemini-2.0-flash, known for their superior performance across various reasoning tasks, as well as open-source models, notably Llama-4-Scout-Instruct [3] and the aya-vision series (32B and 8B parameters) [8]. *We note that to the best of our knowledge, none of the frontier multilingual models claim to support the Irish language.*

4 Results and Analysis

4.1 Performance Gap Between Irish and English

Figure 5 compares the accuracy scores of various models evaluated on IRLBench, segmented by language (English vs. Irish). Our results highlight that open-ended reasoning remains substantially more challenging in extremely low-resource languages such as Irish. For instance, on the English split, the top-performing model, o4-mini, achieves a strong average accuracy of 76.2%, closely followed by gpt-4.1 at 71.3%. In contrast, all evaluated models demonstrate considerably lower accuracy on the Irish version, scoring below 60% despite being presented with the same questions but written in Irish. This indicates a statistically significant performance gap exceeding 10% between the two languages across all models (p-*value* = 0.004).

Generally, models with superior performance on English also perform relatively better on Irish. However, reasoning-adapted

Table 3: Statistical robustness of English-Irish performance gaps across 6 models. All 95% confidence intervals exclude zero, confirming significant differences. Effect sizes (Cohen's d) indicate practical significance.

Model	Gap	95% CI	p	d
gpt-4.1	12.9%	[4.9%, 20.4%]	0.006	0.135
gemini-2.0-flash	17.9%	[9.7%, 25.8%]	0.010	0.125
Llama-4-Scout	23.3%	[14.5%, 31.6%]	<0.001	0.260
o4-mini	20.4%	[12.3%, 28.2%]	<0.001	0.209
aya-vision-32b	46.1%	[38.3%, 53.8%]	<0.001	1.072
aya-vision-8b	38.2%	[30.6%, 45.7%]	<0.001	1.051

models (e.g., through reinforcement learning with test-time scaling), such as o4-mini, exhibit a notably larger gap (20.4%) compared to more general models like gpt-4.1 (12.9%) and gemini-2.0-flash (17.9%). This suggests that current SoTA models, particularly those optimized explicitly for reasoning, have not adequately addressed multilingual transfer, resulting in pronounced performance disparities in low-resource language contexts. IRLBench, therefore, provides an essential resource for evaluating and driving progress in this multilingual frontier.

Human Performance Context. The Irish Leaving Certificate provides parallel exam versions in English and Irish with identical content. This bilingual system operates successfully at national scale for university admissions, demonstrating that human students handle both languages equivalently. In stark contrast, state-of-the-art LLMs exhibit substantial performance gaps (76.2% English vs. 55.8% Irish, p=0.004) on this identical content. This disparity reveals a fundamental difference: while human students demonstrate equal reasoning capabilities in both languages, LLMs behave as "proficient English reasoners but limited Irish reasoners." This gap represents a critical weakness absent in human cognition, highlighting the challenges of achieving true multilingual transfer in current language models.

Statistical Robustness. To verify the reliability of our findings, we conduct robustness analysis following established methodology [23]. We randomly sample 75% of questions (1,000 samples) 100 times and recompute model rankings, which remain highly stable (Spearman's $\rho = 0.985$), exceeding standard reliability thresholds [28]. Table 3 presents 95% confidence intervals and effect sizes (Cohen's d) for the English-Irish performance gap, computed via bootstrap resampling with 10,000 iterations. All models show statistically significant gaps with confidence intervals excluding zero and effect sizes ranging from small (gpt-4.1: d=0.135) to very large (aya-vision models: d>1.0).

4.2 Performance Across Subject Groups

Table 4 summarizes model accuracy across IRLBench's subject groups, providing detailed comparisons between English and Irish performance. Overall, performance varies across subject groups, with models tend to perform best on Social Studies subjects; in some cases, the gap can be more than 20% compared to other groups, such as in the case of gemini-2.0-flash and o4-mini on the Irish split. On the English subset, o4-mini achieves the highest overall accuracy (76.17%), followed by gpt-4.1 (71.29%). In general, open-source

Table 4: Model performances per subject group and language of IRLBench. Overall, models perform best on Social Studies, with larger variance across the remaining disciplines. Bold scores indicate the best performance, and underline highlights the second-best. Subject-group abbreviations: Sci. = Science, Appl. = Applied Science, Biz. = Business, Soc. = Social Studies, Avg. = overall average.

Model	Sci.	Appl.	Biz.	Soc.	Avg.
English					
aya-vision-8b	32.31	36.77	37.43	63.37	38.55
aya-vision-32b	42.76	47.11	44.17	71.29	48.36
Llama-4-Scout-Instr.	61.26	57.91	60.73	<u>78.22</u>	62.99
gemini-2.0-flash	62.77	61.83	60.12	76.24	64.41
o4-mini	**77.06**	**73.35**	**71.16**	**81.19**	**76.17**
gpt-4.1	<u>72.21</u>	<u>68.44</u>	<u>68.10</u>	72.28	<u>71.29</u>
Irish					
aya-vision-8b	0.22	0.00	0.62	1.00	0.36
aya-vision-32b	1.77	1.00	0.62	7.92	2.27
Llama-4-Scout-Instr.	34.10	39.31	44.98	49.50	39.67
gemini-2.0-flash	45.35	38.08	47.85	67.32	46.54
<u>o4-mini</u>	<u>51.58</u>	<u>49.87</u>	<u>53.99</u>	<u>72.28</u>	<u>55.82</u>
gpt-4.1	**54.12**	**54.31**	**60.12**	**73.27**	**58.43**

models lag behind their closed-source counterparts, particularly in science-focused categories (Science and Applied Science). However, an interesting exception occurs in the Social Studies group, where the open-source model Llama-4-Scout-Instruct outperforms the closed-source gemini-2.0-flash by a small margin of 1.98%.

Notably, performance gaps are even more significant in Irish, with smaller open-source models (aya-vision-32b and aya-vision-8b) failing almost entirely. Additionally, a consistent trend emerges: models excelling in English tend to perform better in Irish as well, underscoring that underlying reasoning proficiency strongly influences multilingual transfer. This result emphasizes the importance of robust linguistic and reasoning abilities in achieving strong multilingual generalization.

4.3 Error Analysis: Correctness vs. Language Fidelity

Beyond correctness, we also evaluate the ability of models to generate text in Irish, an extremely low-resource language. We employ a FastText-based language identification model [15], trained specifically for language detection to classify each response at the sentence level. Responses containing more than 50% (set empirically) English sentences are considered non-Irish, representing a lower bound for language fidelity.

Figure 6 illustrates the proportion of responses produced in Irish, broken down by three categories: for all responses, in correct responses only, or in incorrect responses only. We exclude models with extremely low accuracy (shown in previous results) due to insufficient valid samples for analysis. The analysis reveals that most evaluated models (with the exception of gpt-4.1) produce valid Irish responses less than 80% of the time, underscoring significant gaps in generative proficiency. Moreover, there is a clear correlation

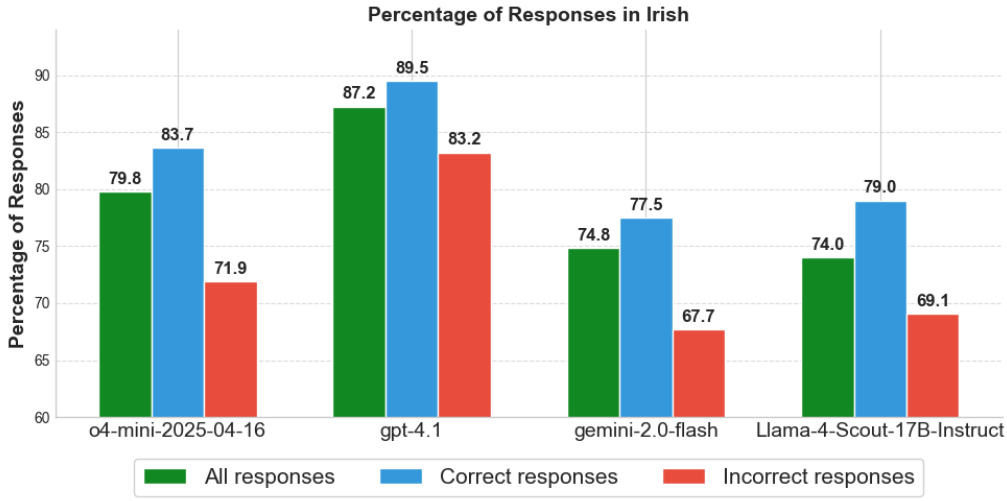

Figure 6: Percentage of responses generated by models that are in Irish (on Irish split of IRLBench), excluding low-performing models. All models except gpt-4.1 produce valid Irish responses less than 80% of the time, underscoring significant gaps in generative proficiency.

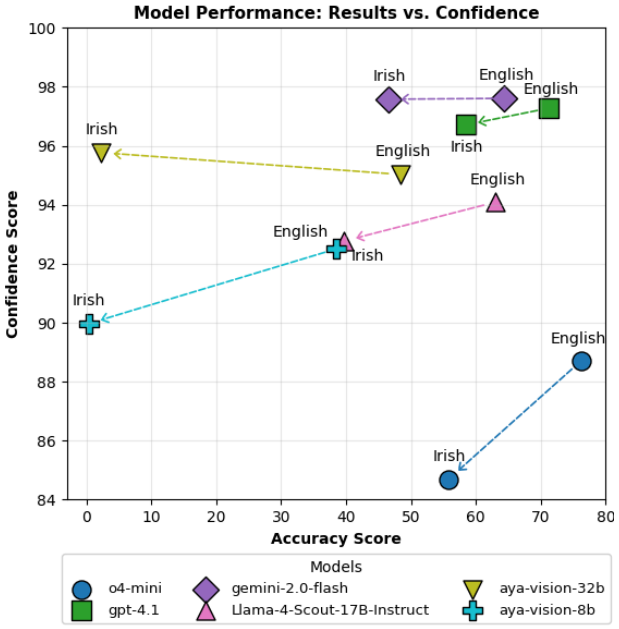

Figure 7: Model self-reported confidences compared to performances on IRLBench [43]. All models display substantial discrepancies, exhibiting overly high confidence (all above 80%) relative to their actual performance on Irish split.

between correctness and the capability to generate outputs in Irish (The percentage in Correct responses compared to the percentage in Incorrect responses), a gap of at least 6.3% in gpt-4.1, suggesting interdependence between reasoning and language-generation skills. Pooling the 3,400 individual responses from the four models, the point-biserial Pearson correlation between a response being in Irish and being rated correct is 0.1084 (p-$value < 0.000001$), a small but reliable effect. These results highlight a notable limitation of previous benchmarks, which often rely on multiple-choice formats and thus do not adequately measure generative language capabilities.

4.4 Confidence Analysis

Following [43], we analyze models' self-reported confidence by prompting them to output confidence scores alongside responses, by directly prompting them to output their confidence (prompt available in our supplementary materials.). Figure 7 compares actual accuracy against self-assessed confidence. Ideally, well-calibrated models should align confidence with accuracy. However, all models display substantial discrepancies, exhibiting overly high confidence (above 80%) relative to actual Irish performance (below 60%). For example, aya-vision-32b maintains above 90% confidence despite only 2.3% accuracy—a drop exceeding 40% from English. Such miscalibration could mislead users relying on confidence scores, likely arising from next-token prediction mechanisms prone to hallucinations.

To quantify the relationship between model confidence and correctness more rigorously, we compute point-biserial correlations between self-reported confidence scores and actual correctness. Table 5 presents results for all evaluated models across both languages. All frontier models exhibit weak positive correlations ($r < 0.25$), confirming the poor calibration observed in Figure 7. Notably, smaller models (aya-vision-32b and aya-vision-8b) show no significant correlation between confidence and correctness ($r < 0.04$, $p > 0.05$), indicating their confidence scores are essentially random and provide no meaningful signal of response quality. This further underscores fundamental calibration challenges in multilingual settings, where models cannot accurately assess their own performance, particularly in low-resource languages.

Table 5: Point-biserial correlation between model confidence and correctness. All correlations are weak (r<0.25), confirming poor calibration across models. ***p<0.001, **p<0.01, ns=not significant.

Model	English	Irish
o4-mini	0.247***	0.138***
gpt-4.1	0.221***	0.184***
gemini-2.0-flash	0.184***	0.194***
Llama-4-Scout-Instruct	0.105**	0.144***
aya-vision-32b	0.009 (ns)	0.033 (ns)
aya-vision-8b	0.031 (ns)	0.020 (ns)

4.5 Judge Reliability Analysis

We assess the reliability of our primary evaluator, gemini-2.5-flash, through three complementary experiments:

- Self-reliability: we sub-sample 1,000 responses across all models and subjects, then perform the same evaluation pipeline with gemini-2.5-flash on another random seed. The resulting judgments remain highly consistent, with a very strong Pearson correlation of 0.7750 (p-value < 0.000001), an agreement ratio of 88.69%, and a Cohen's kappa coefficient of $\kappa = 0.77$, indicating substantial agreement. This demonstrates the stability of gemini-2.5-flash and the reliability of our LLM-as-a-judge approach.
- Agreement with another strong evaluator: we evaluate the above subset of responses using both gemini-2.5-flash and gpt-4.1, which achieves the strongest Irish performance. The judgments of the 2 evaluators, again, exhibit a strong Pearson correlation of 0.7418 ($p\text{-}value < 0.000001$), an agreement ratio of 86.54%, and $\kappa = 0.73$.
- Agreement with a human judge: we conduct an evaluation with a native Irish speaker on 150 samples (a subset of the 1,000-sample set above), achieving an agreement ratio of 82.86%, indicating substantial agreement ($\kappa = 0.66$) between the human judge and gemini-2.5-flash. This demonstrates the robustness and reliability of our LLM-as-a-judge evaluation.

4.6 Impact of Visual Content

IRLBench includes 18.11% multimodal samples requiring interpretation of diagrams, tables, and figures alongside content. Figure 9 (detailed results in Appendix B) reveals that accuracy drops for all models when visual content is present. Gemini-2.0-flash exhibits the largest degradation (16.2 percentage points), despite being designed for multimodal tasks. This pattern holds across both English and Irish splits, indicating that visual content introduces substantial complexity beyond language barriers. The performance drop is particularly pronounced in Irish, where models must simultaneously handle low-resource language understanding and visual interpretation. These findings demonstrate that IRLBench's multimodal component provides a meaningful challenge for evaluating vision-language capabilities in low-resource contexts, an aspect underexplored in existing multilingual benchmarks.

4.7 Translation Agent Evaluation

A natural question arises: *can agentic workflows using translation circumvent Irish language limitations?* To investigate this, we evaluate translation agents that (1) translate Irish questions to English, (2) perform reasoning in English, and (3) translate answers back to Irish. We test two configurations using gpt-4.1 as the reasoning model:

NLLB-3.3B as translator. This specialized neural translation model [50] achieves high language fidelity (97.06% responses in valid Irish, compared to gpt-4.1's direct 87.2%). However, correctness drops dramatically to 53.78% - significantly lower than gpt-4.1's performance on both English (71.3%) and direct Irish (58.4%). This counterintuitive result suggests that translation introduces errors that compound reasoning failures.

GPT-4.1 as translator. Using the same model for all pipeline steps yields better results: 98.63% Irish fidelity and 66.0% correctness. This represents substantial improvement over the NLLB approach, demonstrating that translation quality critically affects downstream reasoning. However, performance remains 5.3 percentage points below the English baseline (71.3% vs. 66.0%).

These findings reveal inherent limitations in translation-based approaches. Neural translation models rely primarily on pattern recognition rather than deep linguistic understanding [19, 22], losing subtle contextual and cultural nuances essential for accurate reasoning. Even with state-of-the-art translation, the English-Irish performance gap persists, underscoring that IRLBench's challenges extend beyond language barriers to encompass fundamental issues of cultural grounding and multilingual reasoning transfer.

5 Conclusion

We introduce IRLBench, the first multilingual, multi-modal benchmark for open-ended generative evaluation in extremely low-resource scenarios. Our systematic evaluations reveal substantial performance gaps between English and Irish (over 10% across all models, p=0.004), underscoring persistent challenges in multilingual reasoning tasks, particularly in resource-limited contexts. The study highlights the need for LLMs to better transfer capabilities across languages despite data scarcity. By providing a rigorous, challenging benchmark, IRLBench lays a cornerstone for low-resource multilingual benchmarks and fosters future research toward truly multilingual language models and greater linguistic diversity in natural language processing.

Disclaimer. The data used throughout this research is based on the Regulations on Re-use of Public Sector Information Regulations 2005 (SI 279 of 2005), and should be used for non-commercial research and study only.

Acknowledgments

This publication has emanated from research supported in part by grants from Research Ireland under Grant [12-RC-2289-P2] and [18/CRT/6223] which is co-funded under the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and En-Shiun Annie Lee. 2024. SIB-200: A Simple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+ Languages and Dialects. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 226–245. https://aclanthology.org/2024.eacl-long.14/

[2] Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards Measuring and Modeling "Culture" in LLMs: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 15763–15784. doi:10.18653/v1/2024.emnlp-main.882

[3] Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. (2025). https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[4] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand and virtual meeting, 749–775. https://aclanthology.org/2024.acl-long.44

[5] Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Joanna Matthiesen, Kevin A. Smith, and Joshua B. Tenenbaum. 2024. Evaluating Large Vision-and-Language Models on Children's Mathematical Olympiads. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=YMAU2kJgzY

[6] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168* (2021).

[7] Compare AI by FoundTT. 2025. GPT-4.1: AI Technical Specifications and Review. https://compare-ai.foundtt.com/en/gpt-4-1/. Accessed: 2025-07-28.

[8] Saurabh Dash, Yiyang Nan, John Dang, Arash Ahmadian, Shivalika Singh, Madeline Smith, Bharat Venkitesh, Vlad Shmyhlo, Viraat Aryabumi, Walter Beller-Morales, Jeremy Pekmez, Jason Ozuzu, Pierre Richemond, Acyr Locatelli, Nick Frosst, Phil Blunsom, Aidan Gomez, Ivan Zhang, Marzieh Fadaee, Manoj Govindassamy, Sudip Roy, Matthias Gallé, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2025. Aya Vision: Advancing the Frontier of Multilingual Multimodality. arXiv:2505.08751 [cs.CL] https://arxiv.org/abs/2505.08751

[9] Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The Multilingual Mind : A Survey of Multilingual Reasoning in Language Models. arXiv:2502.09457 [cs.CL] https://arxiv.org/abs/2502.09457

[10] Google. 2024. Introducing Gemini 2.0: our new AI model for the agentic era. (2024). https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/#ceo-message

[11] Google. 2025. Gemini 2.0: Flash, Flash-Lite and Pro. (2025). https://developers.googleblog.com/en/gemini-2-family-expands/

[12] Google DeepMind. 2025. Gemini Flash. https://deepmind.google/models/gemini/flash/. Accessed: 2025-07-28.

[13] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)* (2021).

[14] Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards Reasoning in Large Language Models: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 1049–1065. doi:10.18653/v1/2023.findings-acl.67

[15] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[16] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification. *arXiv preprint arXiv:1607.01759* (2016).

[17] Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing Massive Multitask Language Understanding in Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 5622–5640. doi:10.18653/v1/2024.findings-acl.334

[18] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. CMMLU: Measuring massive multitask language understanding in Chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11260–11285. doi:10.18653/v1/2024.findings-acl.671

[19] Jiahuan Li, Hao Zhou, Shujian Huang, Shanbo Cheng, and Jiajun Chen. 2024. Eliciting the Translation Ability of Large Language Models via Multilingual Finetuning with Translation Instructions. *Transactions of the Association for Computational Linguistics* 12 (2024), 576–592.

[20] Wenbin Li, Di Yao, Ruibo Zhao, Wenjie Chen, Zijie Xu, Chengxue Luo, Chang Gong, Quanliang Jing, Haining Tan, and Jingping Bi. 2025. Stbench: Assessing the ability of large language models in spatio-temporal analysis. In *Companion Proceedings of the ACM on Web Conference 2025*. 749–752.

[21] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013/

[22] A. Muñoz-Ortiz, C. Gómez-Rodríguez, and D. Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review* 57 (2024), 265.

[23] Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2020. Are Neural Ranking Models Robust? *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–36.

[24] OpenAI. 2025. Introducing GPT-4.1 in the API. (2025). https://openai.com/index/gpt-4-1/

[25] OpenAI. 2025. Introducing OpenAI o3 and o4-mini. (2025). https://openai.com/index/introducing-o3-and-o4-mini/

[26] Yonatan Oren, Nicole Meister, Niladri S Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. 2023. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*.

[27] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Pierre Isabelle, Eugene Charniak, and Dekang Lin (Eds.). Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318. doi:10.3115/1073083.1073135

[28] Sam Parsons et al. 2022. Methods to Split Cognitive Task Data for Estimating Split-Half Reliability: A Comprehensive Review and Tutorial. *Behavior Research Methods* 54 (2022), 44–66.

[29] Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of Cultural Awareness in Language Models: Text and Beyond. arXiv:2411.00860 [cs.CL] https://arxiv.org/abs/2411.00860

[30] Long Phan et al. 2025. Humanity's Last Exam. arXiv:2501.14249 [cs.LG] https://arxiv.org/abs/2501.14249

[31] Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. 2025. None of the Others: a General Technique to Distinguish Reasoning from Memorization in Multiple-Choice LLM Evaluation Benchmarks. arXiv:2502.12896 [cs.CL] https://arxiv.org/abs/2502.12896

[32] Florian Schneider and Sunayana Sitaram. 2024. M5–A Diverse Benchmark to Assess the Performance of Large Multimodal Models Across Multilingual and Multicultural Vision-Language Tasks. *arXiv preprint arXiv:2407.03791* (2024).

[33] Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=fR3wGCk-IXp

[34] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. arXiv:2412.03304 [cs.CL] https://arxiv.org/abs/2412.03304

[35] Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. KMMLU: Measuring Massive Multitask Language Understanding in Korean. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 4076–4104. https://aclanthology.org/2025.naacl-long.206/

[36] Haoxiang Sun, Yingqian Min, Zhipeng Chen, Wayne Xin Zhao, Zheng Liu, Zhongyuan Wang, Lei Fang, and Ji-Rong Wen. 2025. Challenging the Boundaries of Reasoning: An Olympiad-Level Math Benchmark for Large Language Models. arXiv:2503.21380 [cs.CL] https://arxiv.org/abs/2503.21380

[37] Aditya Tomar, Nihar Ranjan Sahoo, Ashish Mittal, Rudra Murthy, and Pushpak Bhattacharyya. 2025. Mathematics Isn't Culture-Free: Probing Cultural Gaps via Entity and Scenario Perturbations. *arXiv preprint arXiv:2507.00883* (2025).

[38] Khanh-Tung Tran, Barry O'Sullivan, and Hoang D Nguyen. 2024. UCCIX: Irish-eXcellence Large Language Model. *ECAI 2024* (2024).

[39] Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Scaling Test-time Compute for Low-resource Languages: Multilingual Reasoning in LLMs. arXiv:2504.02890 [cs.CL] https://arxiv.org/abs/2504.02890
[40] UNESCO (Ed.). 2010. *Atlas of the world's languages in danger* (3 ed.). UNESCO.
[41] Vals AI, Inc. 2025. Models. https://www.vals.ai/models/. Accessed: 2025-07-28.
[42] Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. Vrdu: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 5184–5193.
[43] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. arXiv:2411.04368 [cs.CL] https://arxiv.org/abs/2411.04368
[44] Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The Bitter Lesson Learned from 2,000+ Multilingual Benchmarks. arXiv:2504.15521 [cs.CL] https://arxiv.org/abs/2504.15521
[45] Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=3GTtZFiajM
[46] Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3Exam: A Multilingual, Multimodal, Multilevel Benchmark for Examining Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=hJPATsBb3l
[47] Zeyang Zhang, Xin Wang, Ziwei Zhang, Haoyang Li, Yijian Qin, and Wenwu Zhu. 2024. LLM4DyG: can large language models solve spatial-temporal problems on dynamic graphs?. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4350–4361.
[48] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. https://openreview.net/forum?id=uccHPGDlao
[49] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. LIMA: Less Is More for Alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=KBMOKmX2he
[50] Wenhao Zhu et al. 2024. Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

Supplementary Materials

Due to conference page limitation, comprehensive supplementary materials - including all experimental prompts, additional statistical analyses, and extended results - are provided in a separate document available through our project repository at `https://github.com/ReML-AI/IRLBench`. Below we present key results referenced in the main text for reader convenience.

A Extended Model Evaluation

Table 6 provides detailed performance breakdowns across all subject groups for additional models (claude-3.7-sonnet and Qwen2.5-VL-32B-Instruct) that further validate our findings. Figure 8 presents the relationship between model self-reported confidence and actual performance, demonstrating consistent miscalibration across most of evaluated models.

B Impact of Visual Content

Figure 9 presents performance comparisons between text-only and multimodal samples, as referenced in Section 4.6. This analysis demonstrates that visual content introduces substantial additional complexity across all models and both language splits.

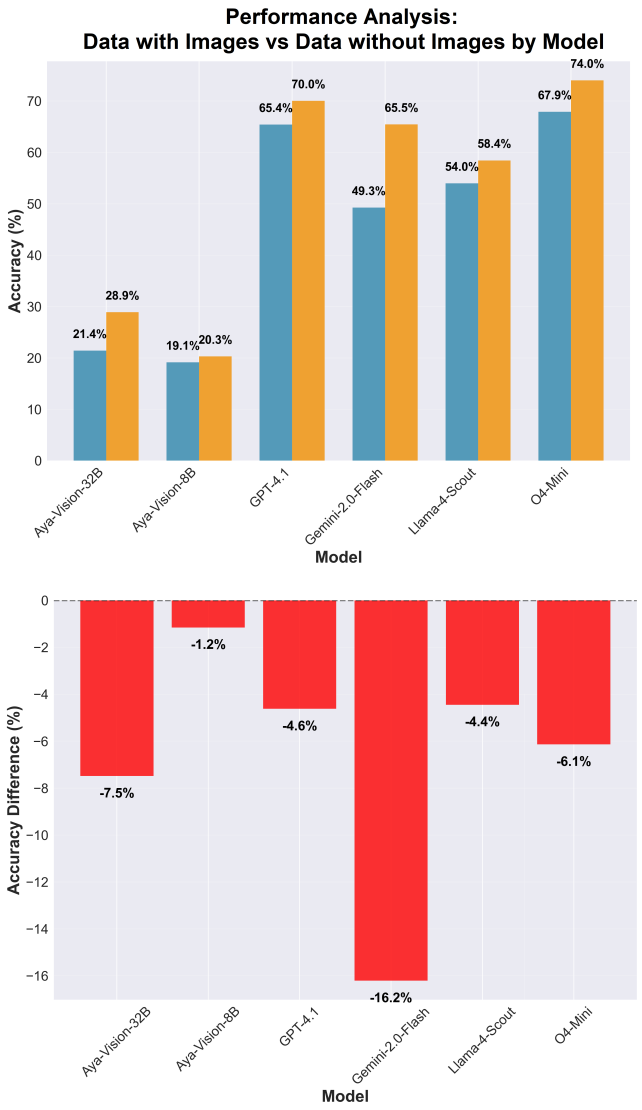

Figure 9: Model Performance on Data with Images and Data without Images

Model	Split	Science	Applied Science	Business Studies	Social Studies	Average
aya-vision-8b	English	32.31	36.77	37.43	63.37	38.55
aya-vision-32b		42.76	47.11	44.17	71.29	48.36
Llama-4-Scout-Instruct		61.26	57.91	60.73	78.22	62.99
gemini-2.0-flash		62.77	61.83	60.12	76.24	64.41
Qwen2.5-VL-32B-Instruct		66.57	63.01	60.73	78.22	66.22
claude-3.7-sonnet		70.86	64.33	63.19	77.23	69.37
o4-mini		**77.06**	**73.35**	**71.16**	**81.19**	**76.17**
gpt-4.1		72.21	68.44	68.10	72.28	71.29
aya-vision-8b	Irish	0.22	0.00	0.62	1.00	0.36
aya-vision-32b		1.77	1.00	0.62	7.92	2.27
Llama-4-Scout-Instruct		34.10	39.31	44.98	49.50	39.67
gemini-2.0-flash		45.35	38.08	47.85	67.32	46.54
Qwen2.5-VL-32B-Instruct		11.30	9.37	4.91	17.93	10.67
claude-3.7-sonnet		38.79	33.44	44.17	58.41	41.21
o4-mini		51.58	49.87	53.99	72.28	55.82
gpt-4.1		**54.12**	**54.31**	**60.12**	**73.27**	**58.43**

Table 6: Model performances per subject group and language of IRLBench.

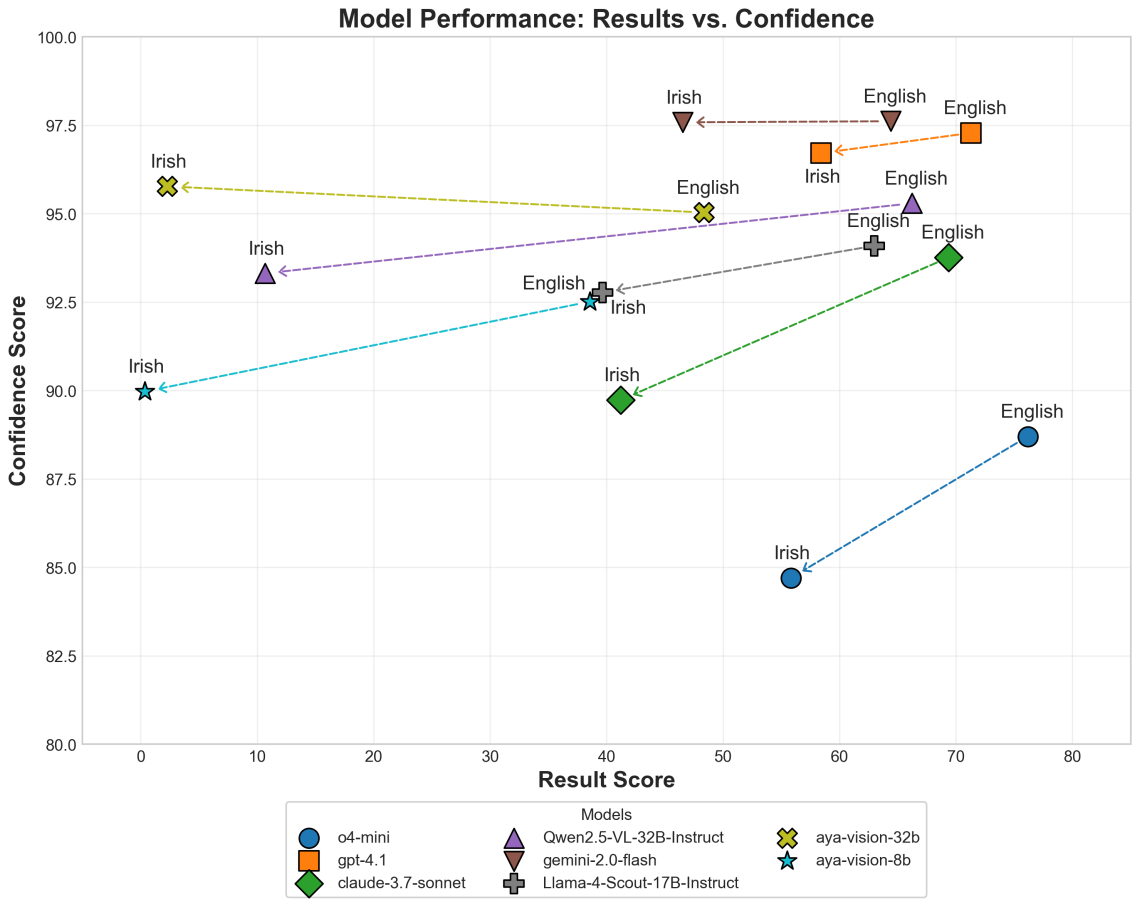

Figure 8: Model self-reported confidences compared to performances.