A Culturally Rich Romanian NLP Dataset from ’Who Wants to Be a Millionaire?’ Videos

Alexandru-Gabriel Ganea * **Antonia-Adelina Popovici** * **Adrian Marius Dumitran**
University of Bucharest University of Bucharest University of Bucharest
antoniaadelina94@gmail.com
ganeaalex222@gmail.com marius.dumitran@unibuc.ro

Abstract

Large Language Models (LLMs) demonstrate varying performance across languages and cultural contexts. This study introduces a novel, culturally-rich, multilingual dataset derived from video recordings of the Romanian game show "Who Wants to Be a Millionaire?" (Vrei să fii Milionar?). We employed an innovative process combining optical character recognition (OCR), automated text extraction, and manual verification to collect question-answer pairs, enriching them with metadata including question domain (e.g., biology, history), cultural relevance (Romanian-specific vs. international), and difficulty. Benchmarking state-of-the-art LLMs, including Romanian-adapted models, on this dataset revealed significant performance disparities: models consistently achieve higher accuracy (80-95%) on international questions compared to Romanian-specific cultural questions (50-75%). We further investigate these differences through experiments involving machine translation of Romanian questions into English and cross-lingual tests using a comparable dataset in French. Our findings underscore the impact of cultural context and data source on LLM performance and offer practical insights for building robust, culturally-aware multilingual NLP systems, especially in educational domains. The dataset is publicly available at Hugging Face.

1 Introduction

The rapid advancement of large language models (LLMs) has transformed many NLP tasks, including question answering, summarization, and translation. However, most evaluations focus on high-resource languages like English, leaving a gap in understanding LLM performance in lower-resource and culturally diverse contexts.

To address this, we introduce a Romanian-language dataset derived from *Vrei să fii milionar?*, the local version of "Who Wants to Be a Millionaire?" [1]. Publicly available on Hugging Face [2], the dataset was compiled directly from video recordings, requiring structured data extraction from a dynamic visual format rather than standard text corpora. It reflects authentic language in a culturally specific, conversational quiz show setting. Each question is annotated for cultural relevance—Romanian-specific vs. international—and labeled by difficulty (easy, medium, hard).

Evaluating LLMs in Romanian presents challenges due to complex grammar, rich vocabulary, and limited NLP resources. We examine how current models handle Romanian multiple-choice question answering (MCQA), focusing on cultural knowledge and the impact of Romanian-specific fine-tuning. Benchmarking open-source LLMs, including Romanian-adapted ones, reveals how language and cultural grounding affect performance. Our results highlight both the strengths and limitations of multilingual LLMs in culturally rich settings and support the development of more robust NLP systems for underrepresented languages like Romanian.

2 Related Work

2.1 Linguistic and Cultural Diversity in LLM Evaluation

LLMs such as GPT-4 (OpenAI, 2023), PaLM (Chowdhery et al., 2022), and LLaMA (Touvron et al., 2023) have demonstrated strong performance on many benchmarks. However, most evaluation datasets remain heavily skewed toward English and

*Equal contribution.

[1] https://en.wikipedia.org/wiki/Who_Wants_to_Be_a_Millionaire

[2] https://huggingface.co/datasets/WWTBM/wwtbm

Proceedings of Recent Advances in Natural Language Processing,pages 378–387
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-046

other high-resource languages (Wang et al., 2022; Nangia et al., 2021), limiting our understanding of model generalization to diverse linguistic and cultural contexts. While multilingual benchmarks like XTREME (Hu et al., 2020) and XGLUE (Liang et al., 2020) address this to some extent, they often lack deep cultural grounding and omit many underrepresented languages.

More recently, efforts like the BOLD benchmark (Dhamala et al., 2021) have begun to incorporate cultural dimensions, particularly in evaluating bias.

2.2 Romanian NLP Datasets and Resources

Romanian, as a mid-resource language, has seen increasing support through dedicated datasets and benchmarks. The RoBERT model (Ştefan Daniel Dumitrescu et al., 2020), based on the original BERT architecture and trained specifically for Romanian, is not to be confused with RoBERTa (Liu et al., 2019), a robustly optimized BERT approach pretrained on larger, English-focused corpora. Foundational resources like CoRoLa (Tufiş et al., 2017) have further enabled robust pretraining and evaluation. More recently, the "Vorbesti Românește?" initiative (Masala et al., 2024) introduced a large-scale instruction-tuned Romanian benchmark suite, along with open-source models and datasets that significantly advance the capabilities of Romanian LLMs across multiple evaluation categories.

2.3 Language Models for Romanian and Other Underrepresented Languages

Several transformer-based models have been trained or adapted specifically for Romanian, such as RoBERT and XLM-R (Ştefan Daniel Dumitrescu et al., 2020; Conneau et al., 2020), which include Romanian in their training data. However, performance varies considerably across domains and task types.

In the broader context of low-resource or typologically diverse languages, initiatives like Masakhane (Nekoto et al., 2020) and AmericasNLP (Mager et al., 2021) promote community-driven, culturally aware NLP development. These efforts highlight the value of local expertise and participatory approaches in building fair and effective tools.

Related efforts include a Turkish “Who Wants to Be a Millionaire?” dataset built from quiz show questions and crowdsourced gameplay data (Aydin et al., 2017), highlighting hybrid human-AI reasoning. In contrast, our dataset is derived from broadcast recordings and focuses on culturally grounded LLM evaluation in Romanian.

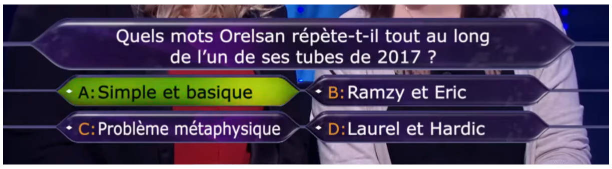
In line with efforts to support underrepresented languages, our work leverages non-traditional data sources for NLP. Unlike typical Romanian datasets based on written corpora, we extract question-answer pairs from quiz show recordings. This yields a more dynamic linguistic sample—capturing spoken patterns and cultural references—and shows how alternative modalities can enhance Romanian NLP resources.

3 Dataset Creation

We developed a multilingual dataset centered on the Romanian edition of the quiz show *Who Wants to Be a Millionaire?*. The core of this resource is the highly curated and annotated Romanian dataset. To facilitate comparative and cross-lingual analysis, we also translated this primary Romanian data into English and curated parallel datasets in English and French, following a similar structure where possible. The creation process for the primary Romanian dataset involved multiple steps, outlined below.

3.1 Data Collection and Frame Extraction

The final dataset comprises 1,000 multiple-choice questions collected from publicly available video recordings. Approximately 400 questions originated from episodes aired between 2011–2012 [3], obtained via Google Drive, and around 600 questions were extracted from 2018–2019 episodes downloaded from YouTube. [4]



Quels mots Orelsan répète-t-il tout au long de l'un de ses tubes de 2017 ?

A: Simple et basique

B: Ramzy et Eric

C: Problème métaphysique

D: Laurel et Hardic

Figure 1: Example of question frame

Each video was analyzed frame-by-frame to capture screenshots precisely when the correct answer turned green, signifying correctness. To avoid capturing redundant frames, we skipped approximately 500 frames after each correctly identified question-answer screenshot.

[3] https://drive.google.com/drive/folders/1zjO0P_awwSm52uWQKc9gMpEh25LrVfnC

[4] https://www.youtube.com/playlist?list=PLvC_Gs1fsycShkx65zNpIqUhlgTuVI8UB

3.2 Text Extraction and Diacritic Correction

Text was extracted using Google's `gemini-1.5-flash-002` (Georgiev et al., 2024) and structured into Q&A pairs. Romanian diacritics were automatically restored using a Romanian fine-tuned model `mt5-base-romanian-diacritics` [5] version of MT5 (Xue et al., 2021), followed by minimal manual correction.

3.3 Duplicate Removal

Duplicates were identified by calculating cosine similarity using embeddings generated with `jina-embeddings-v3` (Sturua et al., 2024). Questions exceeding a similarity threshold of 0.9 were manually reviewed, resulting in the removal of approximately 10 duplicates.

3.4 Metadata Annotation

Questions were enriched with relevant metadata, including:

- **Episode Air Date**.
- **Monetary Value (RON)**: Used as a proxy for question difficulty.
- **Difficulty Level**: Easy, medium, and hard (based on monetary value).
- **Category**: Art and Culture, Cinematography, Gastronomy...
- **Cultural Context**: Romanian or International. For example if a question asks about "Titanic" ->International whereas if the question is about "Filantropica" -> then it is Romanian.

The distribution of questions across difficulty levels (Figure 2) shows the distribution of questions across difficulty levels

3.5 Cultural Context Categorization

Questions were automatically classified as Romanian-specific or international using `Qwen2.5-72B-Instruct` and manually validated, resulting in a 28.4% / 71.6% split.

3.6 Topic Categorization

Questions were automatically assigned to 12 topic domains using `Qwen2.5-72B-Instruct`, followed by manual verification and correction where necessary. The topic distribution is summarized in Figure 3, which shows the number of Romanian-specific and international questions across each domain.

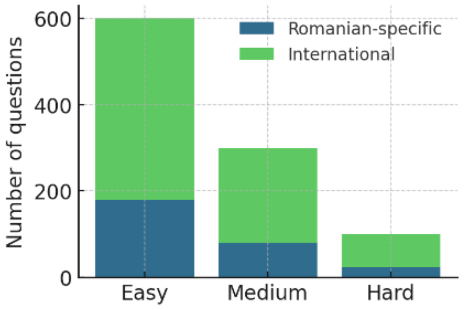


Figure 2: Distribution of difficulty levels across Romanian-specific and international cultural contexts. Note the limited number of "hard" questions reflecting contestant dropout at higher quiz levels.

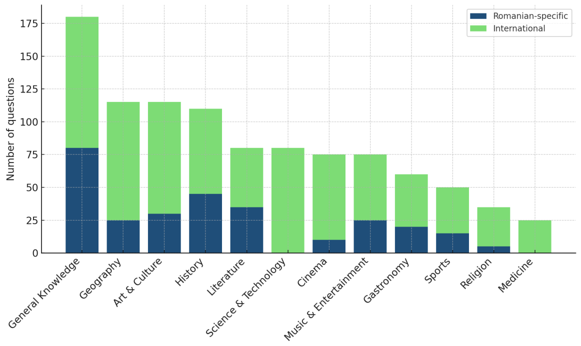


Figure 3: Distribution of questions across topic categories.

4 Methodology

We evaluated the performance of several state-of-the-art LLMs, specifically chosen to represent both general multilingual models and Romanian-specific adaptations. The evaluation framework, including model selection, configurations, prompting strategy, and testing conditions, is detailed below.

4.1 LLM Selection and Configuration

We benchmarked a diverse range of models (7B-72B parameters) covering various architectures, capabilities, and Romanian adaptations.

Inference for the largest models (`Qwen2.5-72B`, `Llama-3.3-70B`) utilized the Hyperbolic API[6]; others were evaluated locally (Kaggle, 2x NVIDIA T4 GPUs).

[5] https://huggingface.co/iliemihai/mt5-base-romanian-diacritics

[6] https://hyperbolic.xyz

Category	Model
General Multilingual	Llama-3.1-8B-Instruct (Grattafiori et al., 2024a) Llama-3.3-70B-Instruct (Grattafiori et al., 2024b) Gemma2-9B-Instruct (Team et al., 2024) Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) Aya-23-8B (Aryabumi et al., 2024) EuroLLM-9B-Instruct (Martins et al., 2024) Qwen2.5-72B-Instruct (Yang et al., 2024) Phi-4 (Abdin et al., 2024)
Romanian Fine-tuned	RoGemma2-9B-Instruct (Masala et al., 2024) RoLlama3.1-8B-Instruct (Masala et al., 2024) RoMistral-7B-Instruct (Masala et al., 2024) Pansophic-1-preview (Pansophic, 2024)

Table 1: Evaluated LLMs used in the benchmark.

4.2 Experiments

We evaluated model performance across multiple dimensions to assess capabilities in handling Romanian language and cultural nuances. The experiments were structured along three main axes:

1. **Category-Based Evaluation:** Performance was measured separately for each of the 12 annotated topic categories (e.g., Art, History, Science) to identify domain-specific strengths and weaknesses.

2. **Difficulty-Based Evaluation:** Models were assessed across easy, medium, and hard difficulty levels (derived from game show monetary value) to evaluate robustness to varying challenge levels.

3. **Cultural Context Evaluation:** We performed separate evaluations on Romanian-specific versus international questions to isolate the impact of cultural context and identify potential cultural knowledge gaps.

To further investigate cross-lingual and cultural generalization, we also conducted experiments using:

- **Comparable French and English Datasets:** To observe if performance patterns generalized across other languages, including another Romance language.

- **Romanian-English Translation:** We translated the Romanian dataset into English and re-evaluated models to help disentangle linguistic understanding challenges from cultural knowledge factors.

4.3 Prompt Design

To ensure consistent and comparable results across all models, we standardized the prompt format. Each prompt included the question text followed by four multiple-choice answer options (`a, b, c, d`). Models were explicitly instructed to respond only with the letter corresponding to their chosen answer. The prompt was structured as follows:

> *Please respond with only the letter corresponding to the correct answer. Do not include any additional text, explanations, or punctuation.*

Additionally, a system-level instruction was included to further guide the model behavior:

> *You are a master of answering multiple-choice questions who responds only with the letter corresponding to the correct answer.*

Both prompts and system instructions were translated into Romanian to match the evaluation language of the dataset. All evaluations were conducted using a zero-shot prompting strategy.

4.4 Model Inference Configuration

To maintain consistent inference behavior and minimize randomness, the following parameters were uniformly applied across all model runs:

- **Max Tokens**: 1 (restricting output to one character).

- **Temperature**: 0 (fully deterministic responses).

This setup was designed to elicit concise responses limited to "a", "b", "c", or "d". Manual inspection showed that nearly all models followed this format. The main exception was `RoMistral-7B`, which produced longer outputs (e.g., *Răspunsul corect este b*, EN: The correct answer is b) in about 10–12% of cases.

4.5 Evaluation Pipelines

Evaluation procedures differed based on model access. **API-based models** (`Qwen2.5-72B`, `Llama-3.3-70B`) were queried via the Hyperbolic API using JSON requests, with built-in retries for rate limits. **Locally executed models** processed structured prompts directly using Hugging Face Transformers and local tokenizers. For both pipelines, failed or incomplete predictions were consistently marked with a placeholder (`'x'`) for standardized error handling.

4.6 Performance Metrics

Model predictions were systematically compared against ground-truth answers. Accuracy scores were calculated separately for each testing condition (cultural context, difficulty, and category), providing insights into models' relative strengths and limitations, particularly concerning their understanding of Romanian language and cultural-specific knowledge.

5 Results and Analysis

We evaluate model performance comprehensively, focusing separately on cultural context, topic categories, and question difficulty.

5.1 Performance by Cultural Context

Table 2 reveals model accuracy based on cultural context (Romanian-specific vs. international).

Model	Romanian	International	Overall
RoGemma2-9B	60.3	91.5	82.8
Gemma2-9B	62.8	89.2	81.8
Llama-3.1-8B	52.3	79.1	71.6
RoLlama3.1-8B	48.7	84.2	74.3
Pansophic-1	50.5	79.8	71.6
Aya-23-8B	46.9	78.1	69.3
Mistral-7B	32.1	59.8	52.0
RoMistral-7B	49.1	84.1	74.3
EuroLLM-9B	70.7	88.2	83.3
Qwen2.5-72B	63.9	94.4	85.9
Llama-3.3-70B	**75.8**	**96.5**	**90.7**
Phi-4	56.3	85.2	77.1

Table 2: Model accuracy (%) by cultural context.

Analysis and Observations:

- **Significant Cultural Gap:** We investigated whether there is a statistically significant difference in model performance between Romanian-specific and international questions.

  The null hypothesis states that there is no difference in mean accuracy between Romanian-specific and international questions across models. For each of the 12 models, we computed the mean accuracy separately on the Romanian-specific and international subsets, and then performed a paired t-test on these per-model differences.

  The test rejected the null hypothesis, showing a statistically significant difference in performance ($t(11) = 18.82$, $p < 0.001$). This indicates that models perform significantly better on international questions than on Romanian-specific ones, highlighting the challenge posed by cultural context.

- **Top Model Performance:** While `Llama-3.3-70B-Instruct` achieves the highest overall (90.7%) and international (96.5%) scores, its accuracy still drops significantly on Romanian-specific questions (75.8%), indicating that large scale does not fully overcome the need for specific cultural knowledge.

- **Difficulty Distribution Paradox:** This performance gap persists *despite* the Romanian-specific subset having proportionally fewer 'Hard' questions (Figure 2). This suggests that even 'Easy' or 'Medium' Romanian-specific items pose greater intrinsic difficulty, likely due to cultural facts, historical figures, or linguistic nuances underrepresented in training data.

- **Fine-tuning Effects:** Romanian fine-tuning yields mixed results. `RoGemma2-9B` and `RoMistral-7B` improved notably over their base models, mainly on international questions, suggesting better linguistic adaptation rather than direct cultural knowledge gains.

- **Smaller Model Struggles:** Smaller models like `Mistral-7B` and `Aya-23-8B` struggle significantly, particularly with Romanian-specific questions, highlighting the combined difficulty of handling a mid-resource language and specific cultural content.

These results underscore the importance of evaluating LLMs within specific cultural contexts. Further few-shot results (1–5 examples) are available online[7].

5.2 Performance by Topic Categories

To understand performance variations across different knowledge domains, we analyzed model accuracy for each of the 12 topic categories. Figure 4 presents these results as a heatmap, offering a visual comparison across models and topics. The topic categories on the x-axis are sorted by descending average accuracy across all models, indicating

[7] https://github.com/AntoniaPopovici/ranlp-2025-few-shot-results

generally "easier" topics on the left and "harder" topics on the right.

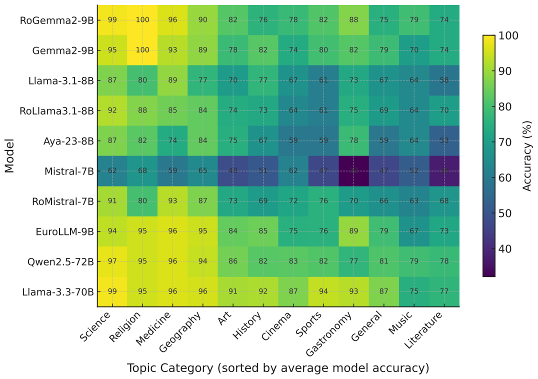


Figure 4: Model accuracies by topic (categories sorted by average score).

Analysis and Observations: The heatmap highlights several key trends:

- **Topic Difficulty Gradient:** A clear performance gradient exists, with topics like Science, Religion, and Medicine (left side) generally yielding higher accuracies than Literature, Music, and General Culture (right side).
- **Model Scale Matters:** Larger models (`Llama-3.3-70B`, `Qwen2.5-72B`) display consistently high performance (more yellow/bright green) across nearly all categories.
- **Specific Model Weaknesses:** Smaller models, particularly `Mistral-7B`, show significant weaknesses (darker cells) in multiple, often lower-performing, categories.
- **Fine-tuning Effects Visually:** Romanian fine-tuning shows noticeable benefits for `RoMistral-7B` compared to its base. Improvements for `RoGemma2-9B` appear more targeted (e.g., Medicine), while `RoLlama3.1-8B` shows less consistent visual improvement over its base.

Overall, the heatmap confirms that performance varies significantly by topic, influenced by model scale and targeted fine-tuning.

5.3 Performance by Difficulty Levels

Table 3 summarizes accuracy based on question difficulty.

Model	Easy	Medium	Hard
RoGemma2-9B	84.7	79.2	84.2
Gemma2-9B	84.3	77.9	73.6
Qwen2.5-72B	86.4	82.8	**94.7**
Llama-3.3-70B	**91.4**	**86.4**	89.4

Table 3: Accuracy (%) by question difficulty level.

Observations:

- `Llama-3.3-70B` and `Qwen2.5-72B` show robustness across all difficulty levels.
- Smaller models generally see significant accuracy drops on medium and hard questions.
- Romanian-fine-tuned models often outperform their base models at higher difficulties.

5.4 Cross-Language Testing on French and English Datasets

To investigate whether observed performance patterns generalize beyond Romanian, we evaluated models on comparable French and English datasets. For each language, we used a sample consisting of 35 hard, 165 medium, and 300 easy questions. Table 4 shows the comparative accuracy.

Model	Romanian	French	English
Gemma2-9B	81.4	76.4	88.8
Llama-3.1-8B	71.8	68.8	85.0
Aya-23-8B	73.4	66.0	77.6
Mistral-7B	52.6	47.2	75.6
EuroLLM-9B	74.2	70.6	79.6
Qwen2.5-72B	68.2	83.1	93.6
Llama-3.3-70B	77.0	82.2	94.0

Table 4: Accuracy (%) by language.

A consistent performance hierarchy emerges across models, with accuracy highest in English, followed by Romanian, and lowest in French (Table 4). This likely reflects the dominance of English in pre-training data. The lower French scores compared to Romanian suggest language-specific challenges or dataset differences beyond language family. While larger models like `Qwen2.5-72B` and `Llama-3.3-70B` show greater cross-lingual robustness, consistent performance across languages remains challenging—highlighting the need for continued language-specific evaluation in multilingual NLP.

5.5 Translation-Based Comparison with English

To isolate the impact of linguistic versus cultural grounding, we translated all Romanian questions into English and re-evaluated model performance. By comparing responses to the original and translated versions, we analyzed whether discrepancies arose from limitations in Romanian language understanding or from gaps in cultural knowledge.

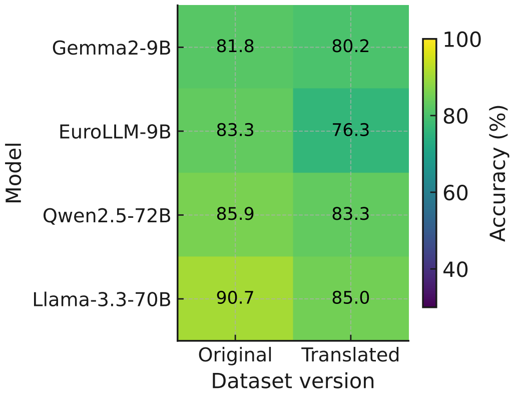


Figure 5: Comparison between original and translated Romanian dataset.

This experiment revealed that translation did not help and, in fact, led to slightly worse performance, highlighting the importance of native-language benchmarks for fair and accurate evaluation.

5.6 Illustrative Examples of Culturally-Specific Questions

To illustrate the nature of the challenges posed by the Romanian-specific subset (discussed in Section 5.1), here are selected examples requiring localized knowledge. English translations are provided.

- **Historical Context (Index 896):** *RO:* În afară de Marea Unire, domnia lui Ferdinand I a fost marcată și de: (*EN:* Besides the Great Union, the reign of Ferdinand I was marked by:) Answer: *Primul Război Mondial / World War I*. Model prediction: *loss of Bucovina*. *Commentary:* The English translation misses the Romanian historical context, where Ferdinand's reign is more closely associated with WWI than the loss of Bucovina.

- **Geographical Knowledge (Index 958):** *RO:* Ce munți trebuie să urci ca să vizitezi Babele? (*EN:* What mountains must you climb to visit Babele?) Answer: *Bucegi*. Model prediction: *Făgăraș*. *Commentary:* The English translation doesn't account for the specific location of Babele in the Bucegi Mountains, leading to the incorrect prediction of Făgăraș.

- **Idiomatic Expression (Index 918):** *RO:* O expresie veche românească spune că omul care speră lucruri irealizabile visează: (*EN:* An old Romanian expression says that the man who hopes for unachievable things dreams:) Answer: *cai verzi pe pereți / green horses on walls*. Model prediction: *green and dried*. *Commentary:* A literal translation misses the idiomatic meaning, which is culturally rooted and nonsensical without Romanian-specific knowledge.

These questions exemplify how the dataset probes knowledge beyond internationally common facts, demanding culturally embedded understanding.

6 Conclusions

We introduced **WWTBM**, a novel, culturally-rich multilingual dataset derived from "Who Wants to Be a Millionaire?" videos, designed for evaluating LLMs on Romanian cultural nuances. Our benchmarking revealed several key insights into model performance at the intersection of language and culture.

A **significant performance gap** consistently appears between international (high accuracy) and **Romanian-specific questions** (lower accuracy) across all tested models, including large-scale ones like `Llama-3.3-70B-Instruct`. This highlights that current pre-training does not fully capture deep cultural knowledge. Furthermore, Romanian-specific fine-tuning showed *mixed effects*, sometimes improving linguistic adaptation more than cultural knowledge retrieval.

Cross-lingual tests confirmed performance variations across languages (English > Romanian > French). Notably, translating Romanian questions to English **decreased accuracy** compared to the original, underscoring the value and potential subtleties of **native-language benchmarks**.

Overall, our findings stress the critical need for **culturally-grounded datasets** like WWTBM for robust LLM evaluation. Developing models adept across diverse cultural contexts remains challenging but crucial, particularly for building effective and equitable *educational NLP applications*. WWTBM provides a valuable resource for advancing this research.

7 Ethical Considerations

The dataset used in this research was obtained solely from publicly available video recordings of the Romanian edition of *Who Wants to Be a Millionaire?*, accessible on platforms such as YouTube. No private or restricted content was involved.

Screenshots were captured solely to extract text using OCR. The resulting question-answer pairs were manually verified and converted into structured, annotated text. No video, audio, or image content is redistributed—only the derived text dataset is shared. We believe this qualifies as academic fair use due to its educational purpose and transformative nature.

The dataset is released as an open-source resource to support multilingual and culturally-aware NLP research. No personally identifiable information of participants or audience members is included.

We acknowledge the importance of ethical use and will address any concerns raised by content owners. Any takedown or modification requests will be honored. Future users are encouraged to use the dataset responsibly and in accordance with applicable copyright.

8 Limitations

While this study offers insights into multilingual LLM performance on Romanian-specific and international content, several limitations remain.

First, the dataset is relatively small (1,000 questions) and sourced from a single quiz show. This may limit statistical power, particularly for rare topics and higher difficulty levels. Additionally, the data reflects the structure and cultural focus of the show, which may not fully represent broader Romanian knowledge domains.

Second, we relied on automated tools (OCR, LLM-based annotation) during dataset construction. Although manually validated, some annotation errors may persist.

Third, our evaluation focuses on multiple-choice questions with limited context, which may not capture the full range of language understanding. Performance could differ on open-ended or reasoning-heavy tasks.

Lastly, while most models evaluated are open-weight to support reproducibility, we included `Phi-4`, accessed via API, as a strong non-open baseline. Other proprietary models (e.g., GPT-4 (OpenAI, 2023), Claude (Anthropic, 2023), Gemini (DeepMind, 2023)) were excluded due to licensing restrictions, closed weights, and evolving APIs, which hinder consistent benchmarking.

Future work will expand dataset coverage, explore open-ended formats, and include both classical and proprietary baselines to better support Romanian-language NLP development.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *arXiv*.

Anthropic. 2023. Claude 2.1 overview. `https://www.anthropic.com/index/claude-2-1`. Accessed: 2025-07-22.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Bahadir Ismail Aydin, Yavuz Selim Yilmaz, and Murat Demirbas. 2017. A crowdsourced "who wants to be a millionaire?" player. In *Concurrency and Computation: Practice and Experience*, volume 29, page e4168.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2022. Palm: Scaling language modeling with pathways.

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

Google DeepMind. 2023. Gemini: Our largest and most capable ai model. `https://deepmind.google/technologies/gemini/`. Accessed: 2025-07-22.

Jwala Dhamala, Ronan Le Bras, Jihoon Kang, Chandra Bhagavatula, Yejin Choi, and 1 others. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *arXiv*.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, and the Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Aaron Grattafiori, Abhimanyu Dubey, and 1 others. 2024a. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. Llama-3.1-8B-Instruct model card: `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`.

Aaron Grattafiori, Abhimanyu Dubey, and 1 others. 2024b. The llama 3 herd of models. *Preprint*, arXiv:2407.21783. Llama-3.3-70B-Instruct model card: `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yichao Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Fang Kong, Shuguang Liu, Wei He, and 1 others. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6008–6018. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *arXiv*.

Manuel Mager, Pedro Ortiz Suárez, and 1 others. 2021. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. Eurollm: Multilingual language models for europe. *Preprint*, arXiv:2409.16235.

Mihai Masala, Denis C. Ilie-Ablachim, Dragoş Corlătescu, Miruna Zavelca, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. Openllm-ro: Technical report on open-source romanian llms. *Preprint*, arXiv:2405.07703. Models at `https://huggingface.co/OpenLLM-Ro`.

Nikita Nangia, Alex Wang, and Samuel R. Bowman. 2021. Winomt: Gender bias in machine translation revisited. *arXiv*.

Wilhelmina Nekoto, Vukosi Marivate, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *arXiv*.

Pansophic. 2024. Pansophic-1-preview. `https://huggingface.co/pansophic/pansophic-1-preview-LLaMA3.1-8b`.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv*.

Dan Tufiş, Elena Irimia, and 1 others. 2017. The corola corpus of contemporary romanian language. In *Proceedings of the Workshop on Corpora and Tools for Romanian Language*.

Yizhong Wang, Swaroop Mishra, Alisa Liu, and 1 others. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 24 others. 2024. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ştefan Daniel Dumitrescu, Andrei M. Avram, and Andrei M. Butnaru. 2020. The birth of romanian bert. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4324–4329. Association for Computational Linguistics.