RoBiologyDataChoiceQA: A Romanian Dataset for improving Biology understanding of Large Language Models

Anonymous ACL submission

Abstract

In recent years, large language models (LLMs) have demonstrated significant potential across various natural language processing (NLP) tasks. However, their performance in domain-specific applications and non-English languages remains less explored. This study introduces a novel Romanian-language dataset for multiple-choice biology questions, carefully curated to assess LLM comprehension and reasoning capabilities in scientific contexts. Containing approximately 14,000 questions, the dataset provides a comprehensive resource for evaluating and improving LLM performance in biology.

We benchmark several popular LLMs, analyzing their accuracy, reasoning patterns, and ability to understand domain-specific terminology and linguistic nuances. Additionally, we perform comprehensive experiments to evaluate the impact of prompt engineering, fine-tuning, and other optimization techniques on model performance. Our findings highlight both the strengths and limitations of current LLMs in handling specialized knowledge tasks in low-resource languages, offering valuable insights for future research and development.

1 Introduction

While LLMs excel in many general NLP tasks, challenges persist in specialized domains and non-English languages, making Romania's rich tradition in biology an ideal context for evaluating LLMs scientific reasoning in a relatively low-resource language.

To address this, we introduce a novel Romanian-language dataset consisting of multiple-choice biology questions sourced from two prestigious national platforms: the Romanian Biology Olympiad and medical school admission examinations. The Romanian Biology Olympiad is the country's largest and most popular biology competition, catering to students from middle school through high school, while medical school entrance exams rigorously test pre-university candidates on their foundational biology knowledge. Together, these sources offer a comprehensive and challenging set of questions covering a wide range of biological concepts, levels of difficulty, and linguistic complexity.

This study goes beyond mere benchmarking of LLMs. We conduct extensive experiments to explore model performance variations under different experimental conditions, such as prompt engineering, model source, and domain-specific fine-tuning. Statistical analyses provide insights into how well models grasp biological concepts in Romanian, identify common failure patterns, and highlight differences in models' performances.

Our work contributes to advancing the understanding of LLM performance in several key ways:

Dataset Creation: We introduce a carefully curated Romanian-language biology dataset suitable for both benchmarking and research in specialized domains.

Benchmarking: We assess the capabilities of leading LLMs, identifying their strengths and limitations in scientific reasoning (which is something LLMs generally struggle with, as shown by Huang and Chang, 2023) within a low-resource language setting.

Experimental Analysis: We explore the impact of various factors on model performance, offering insights that can inform future improvements in LLM development and deployment for specialized tasks.

By presenting these findings, we aim to foster further research on LLM applications in non-English languages and specialized domains, as well as to promote NLP advancements tailored to educational and scientific contexts. Our dataset plays a crucial role in enhancing LLMs' performance in biology by enabling fine-tuning on domain-specific data. The benchmarking methodology established in this work supports continued exploration in this critical area.

2 Related work

Biomedical question-answering (QA) datasets have played a crucial role in advancing domain-specific language models. PubMedQA (Jin et al., 2019) introduced a large-scale English-language biomedical QA dataset with 1,000 expert-annotated, 61,200 unlabeled, and 211,300 artificially generated *yes/no/maybe* questions. While valuable for scientific text comprehension, it does not include multiple-choice questions, which require more complex reasoning over structured information.

A more relevant effort is MedQA (Jin et al., 2021), an open-domain multiple-choice QA dataset collected from professional medical board exams. MedQA covers three languages — English (12,723 questions), simplified Chinese (34,251 questions), and traditional Chinese (14,123 questions) — and requires models to select the correct answer from multiple options rather than extracting answers directly from text. Similarly, MedMCQA (Pal et al., 2022) is an English-language multiple-choice QA dataset designed for medical entrance exams, containing over 194,000 questions. Unlike MedQA, which focuses on board exam questions, MedMCQA emphasizes a wide range of medical knowledge, testing over ten different reasoning abilities.

Efforts to develop language models specialized for Romanian biology are quite limited. One notable contribution is RoQLlama, a lightweight Romanian-adapted language model designed to enhance NLP performance in Romanian-language applications (Dima et al., 2024). RoQLlama was evaluated using the RoMedQA dataset (Crăciun, 2023), a specialized collection of Romanian medical school examination questions.

Our work surpasses this effort by introducing a carefully curated and extended Romanian-language biology dataset extracted from multiple sources, going beyond single-choice questions. We also fine-tune promising models and perform multiple benchmarks. Fine-tuning on our dataset significantly improves LLM performance, making it a valuable resource for enhancing language models in biology. By focusing on this domain, our dataset diversifies the range of available domain-specific resources for Romanian, complementing previous contributions in the medical field and aiming for deeper reasoning.

Guidance on creating and documenting high-quality NLP datasets is essential for ensuring the utility of research outcomes. The dataset documentation framework proposed by Gebru et al., 2018 provided foundational insights for structuring the description and documentation of our dataset.

The use of LLMs in biology has shown significant potential for transforming research in the life sciences. Bhattacharya et al., 2023 explored the evolution of LLMs from textual comprehension tools to multimodal systems capable of analyzing complex biological data and contributing to advances in molecular biology and medicine. Their findings highlight the importance of LLMs in handling scientific reasoning and specialized terminology, which is central to our work.

3 Dataset Composition

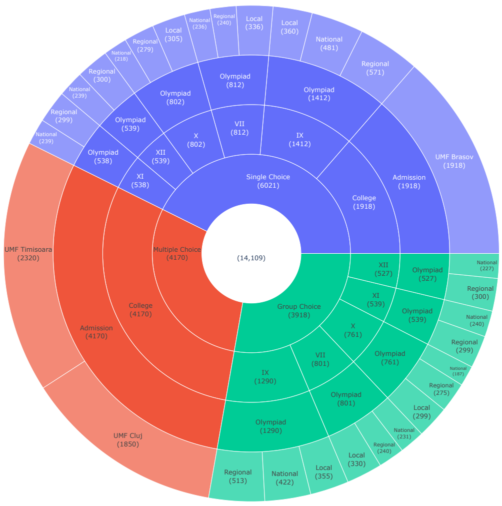

Figure 1: The data distribution based on question type and collection sources details.

3.1 Olympiads

The *Romanian National Biology Olympiad* is a multiple-choice-based competition structured in multiple stages, covering all high school grades and occasionally including middle school. A typical Olympiad exam consists of three primary question categories:

- **Single-choice questions** – Typically, 30 choice questions with a single correct answer.

- **Group-choice questions** – Another 30 questions, where each answer can be one of five predefined lettered combinations (further details in A).

- **Complex single-choice questions** – A set of 10 advanced problems requiring analytical problem-solving to determine the correct answer.

There are exceptions to this standard format, particularly in older exams or localized stages, where the structure may differ, featuring only single-choice questions or a varying number of items.

Olympiad data is collected exclusively from **PDF documents** available online, typically hosted on news websites, archived school portals, or dedicated Olympiad platforms such as olimpiade.ro.

As shown in Figure 3, we extract only **single-choice** and **group-choice** questions from multiple grades, covering various competition stages and years (Figure 2). Given that the source documents are predominantly text-based PDFs (with occasional Word files, which we manually convert into PDFs), **PyMuPDF4LLM** (Artifex, 2024) is used to extract content in Markdown format. The extracted text is subsequently parsed into question instances using **regular expressions**.

A major challenge in this process is **word fragmentation** due to inconsistencies in document formatting. To address this, we employ **Gemini 1.5 Flash** and **Gemma2 9B Instruct** for grammar correction, followed by manual validation. This suggests that LLMs exhibit a tendency to favor logically correct statements, indicating that they have either encountered similar data during training or have developed an implicit understanding of correctness through their learned representations.

3.2 College Admission

Several Romanian universities use **multiple-choice-based admission exams**, with each university providing a dedicated question book (Matusz et al., 2020; Costache et al., 2020; Opincariu et al., 2018). These books, authored by university professors, serve as the **primary study resource** for candidates, as the actual exam questions are guaranteed to be similar to them. Our dataset includes approximately **6,000 questions** collected from the admission preparation books of three universities (Figure 3).

Unlike the Olympiad materials, these documents are **scanned books in image-based PDFs**, necessitating Optical Character Recognition (OCR). The lack of Romanian-specialized OCR tools presents a challenge. While **docTR** (Liao et al., 2023) , a library known for strong English OCR performance, was tested, it proved inadequate for Romanian text. The most viable alternative was **Tesseract OCR**, optimized with **OpenCV-based noise removal preprocessing** (Kotwal et al., 2021). However, this approach introduced challenges:

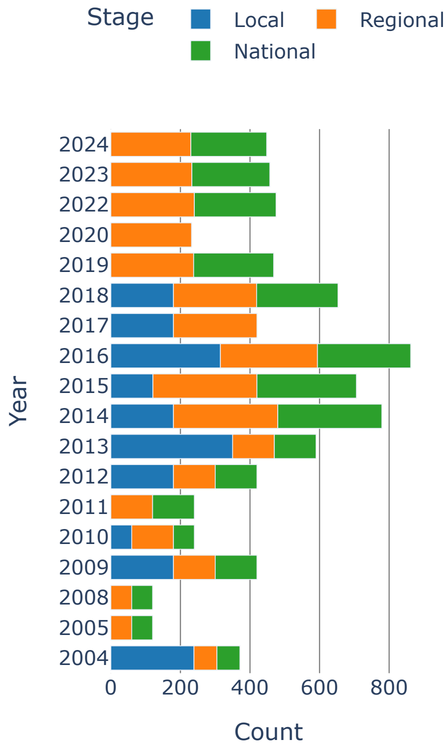

Figure 2: How many questions were collected from each year and of which type.

- **Inconsistent noise removal** – Some techniques improved OCR accuracy for one page while degrading performance on others.

- **Language constraints** – The texts, although in Romanian, contain **Greek letters** used for specialized terminology (e.g., α, β, γ). While Tesseract supports multiple languages, enabling both Romanian and Greek led to **higher misinterpretation rates** rather than improved detection of Greek symbols.

To mitigate these issues, we explored **AI-based OCR solutions**, relying on context-aware processing for improved accuracy. The **Gemini Flash 1.5** model provided better results in recognizing text within scanned images. However, occasional hallucinations—such as **unintended duplication of questions**—necessitated **manual verification** to ensure proper extraction.

3.3 Deduplication

When identical questions with the same answer options appear across different tests or problem sets, we assign them a shared `dupe_id`, a unique UUID identifying a group of duplicates. Each group contains at least two instances. A question is considered a duplicate if both its text and answer options match, regardless of option order, which, as a matter of fact, could impact performance (Pezeshkpour and Hruschka, 2024). To detect slight rephrasings, we compare text embeddings generated with **jina-embeddings-v3** (Sturua et al., 2024).

Rather than removing duplicates, we mark them, as it is unclear which instance should be deleted. Duplication data may also reveal relationships between different subjects. While duplicates remain in the dataset, users can filter them using the `dupe_id` if needed. We ensure that no duplicates exist between the training, validation, and test splits to maintain dataset integrity.

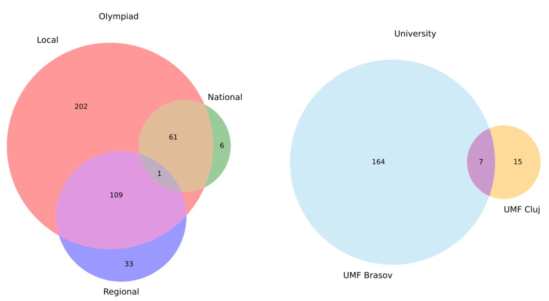

Figure 3: Duplication groups by stage. Overlaps indicate that the same question appears across all the participating stages. There is no duplicate question to be present in both olympiad and university subjects at the same time.

4 Experiments

We conducted comparisons and benchmarks based on multiple criteria, including zero-shot vs. few-shot settings, heuristics for group choice, and combined vs. individual predictions. Notably, all experiments were performed with the temperature set to zero to enhance reproducibility. All experiments were conducted using a Google Colab Pro subscription and various API subscriptions, with a total cost of under $50. While we do not have an exact estimate for continuous runtime, the experiments were carried out over 2–3 months of intermittent activity.

4.1 Benchmarking on RoBiologyDataChoiceQA

Acknowledging good benchmarking practices explored by Liang et al., 2023, we evaluate multiple LLMs on the test split of the RoBiologyDataChoiceQA dataset and report their accuracies in Table 1. The selected models include those offering accessible API usage as well as competitive open-source Romanian models. Details regarding the prompts used can be found in the Appendix (B).

Despite the dataset being in Romanian, the Romanian-trained models (*Rogemma2, Rollama3-8B-Instruct-Imat, and Romistral-7B-Instruct*) did not show a significant advantage over multilingual or primarily English-trained models. Given their explicit training on Romanian (Masala et al., 2024), we expected them to perform better due to their stronger grasp of Romanian syntax and semantics. However, the observed improvements were marginal, suggesting that language understanding alone is not enough to solve this task. Instead, performance appears to be primarily constrained by the models' ability to reason about biological concepts and apply domain knowledge rather than by linguistic factors.

Studies (Nguyen et al., 2025; Gao et al., 2024) have shown that running the same models from different providers could yield slightly different accuracies in some contexts. This was not our case, since doing this resulted in nearly identical accuracies, with variations of at most 0.04. Therefore, we do not specify the source for each model. We conduct evaluations both locally and via external providers.

Model	Single Acc.	Group Acc.	Multi Acc.
gemini-2.0-flash	**0.733**	0.524	**0.585**
gemini-2.0-flash-exp	0.719	**0.537**	0.539
qwen-max-2025-01-25	0.699	0.472	0.573
llama-3.1-405B-Instruct-Turbo	0.685	0.426	0.464
gemini-1.5-flash	0.668	0.419	0.406
DeepSeek-V3	0.665	0.453	0.474
llama-3.3-70B-Instruct-Turbo	0.629	0.413	0.378
rogemma2-9b-instruct (Q8)	0.543	0.298	0.198
gemma2-9b-it	0.529	0.346	0.226
llama3-8b-instruct	0.405	0.250	0.093
phi-3.5-mini-instruct (F32)	0.379	0.208	0.080
eurollm-9b-instruct (F16)	0.384	0.220	0.102
rollama3-8b-instruct-imat (FP16)	0.371	0.235	0.102
romistral-7b-instruct (Q8)	0.371	0.252	0.077
mistral-7b-instruct-v0.1 (Q8)	0.221	0.199	0.046
Baseline	0.245	0.200	0.032

Table 1: Accuracies of models benchmarked on zero shot.

Running the models with a few-shot approach did not yield substantial improvements (phenomenon also found in Hendrycks et al., 2021 and Kojima et al., 2023); in fact, some models performed worse, as shown in Figure 4. Notably, certain LLMs exhibited a tendency to overfixate on specific letters after being presented with examples—interestingly, not necessarily the ones included in the prompt. The few-shot examples were provided to the LLMs within the system prompt, as described in Appendix B.

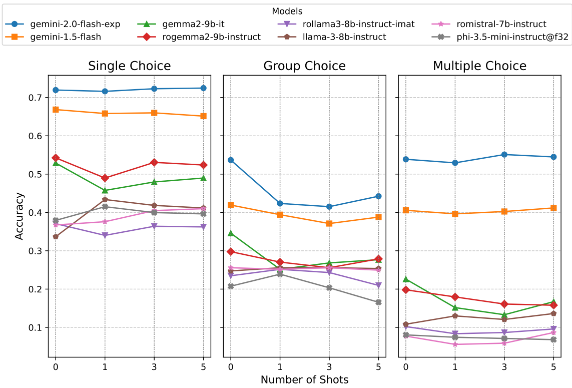

Figure 4: Accuracies of some models over few shot prompting.

4.2 Benchmarking by source type

Multiple	Single Acc.		Multiple Acc.	
	Olympiad	UMF Brasov	UMF Timisoara	UMF Cluj
gemini-2.0-flash-exp	0.704	0.824	0.615	0.415
qwen-max-2025-01-25	0.679	0.838	0.655	0.439
llama-3.1-405B-Instruct-Turbo	0.665	0.824	0.565	0.301
gemini-1.5-flash	0.658	0.743	0.485	0.276
DeepSeek-V3	0.650	0.770	0.540	0.366
llama-3.3-70B-Instruct-Turbo	0.611	0.757	0.445	0.268
rogemma2-9b-instruct (Q8)	0.531	0.622	0.230	0.146
gemma2-9b-it	0.502	0.716	0.255	0.179
llama3-8b-instruct	0.409	0.378	0.130	0.033
eurollm-9b-instruct (F16)	0.393	0.270	0.110	0.073
phi-3.5-mini-instruct (F32)	0.387	0.324	0.085	0.073
romistral-7b-instruct (Q8)	0.374	0.324	0.085	0.065
rollama3-8b-instruct-imat (FP16)	0.372	0.365	0.120	0.073
mistral-7b-instruct-v0.1 (Q8)	0.210	0.297	0.055	0.033
Baseline	0.250	0.200	0.032	0.032

Table 2: Accuracies of models, separated by source.

We compare model performance on Olympiad data versus university admission data. As shown in Figure 2, models tend to perform better on university-level questions with a single correct answer, suggesting they are more accustomed to medical admission data than to biology Olympiad questions. Alternatively, this may indicate that olympiad questions are potentially more challenging, requiring deeper knowledge and reasoning skills.

In Figure 2, we highlight instances where Olympiad scores surpass university admission scores. Even in these cases, the difference is generally small. However, when university admission scores are higher, the margin tends to be larger.

Comparing the difficulty levels of the three universities, we observe that the UMF Brașov exam appears to be the easiest, as it consists solely of single-answer questions. In contrast, the UMF Timișoara and UMF Cluj exams contain multiple-answer questions, making them more challenging and not directly comparable to UMF Brașov. Additionally, UMF Cluj's exam seems to be the most difficult, as all models achieve higher scores on UMF Timișoara's admission questions. This aligns with the common perception that among the three universities analyzed, UMF Cluj has the most difficult admission exam, followed by UMF Timișoara, while UMF Brașov is considered the easiest.

4.3 Finetuning Gemini 1.5 Flash

Google AI Studio allows fine-tuning of the **Gemini 1.5 Flash** model with custom data by providing a CSV file where one column serves as the input and another as the model's output. Using the training split of the RoBiologyDataChoiceQA dataset, we set the input as the benchmarking prompt, replacing %question-text% with the formatted question entry. The output corresponds to the correct answer field without additional formatting.

Once training is complete, we evaluate the fine-tuned model on the test split. We train multiple versions with different parameter settings (e.g., number of epochs, batch size) as detailed in Figure 5. Our fine-tuned models achieve new state-of-the-art accuracies, as shown in Table 3.

Model	Single Accuracy	Group Accuracy	Multiple Accuracy
gemini-2.0-flash	0.733	0.524	**0.585**
tuned_batch16_epochs5	**0.752**	0.627	0.486
tuned_batch16_epochs3	0.738	**0.642**	0.505
tuned_batch16_epochs1	0.733	0.614	0.486
tuned_batch32_epochs5	0.728	0.608	0.471
tuned_batch32_epochs3	0.748	0.629	0.533
tuned_batch32_epochs2	0.750	0.633	0.505
tuned_batch32_epochs1	0.745	0.637	0.464
tuned_batch16_epochs2	0.748	0.639	0.505
tuned_batch64_epochs3	0.733	0.612	0.517
gemini-1.5-flash	0.668	0.419	0.406

Table 3: Accuracies of fine-tuned Gemini 1.5 Flash models

4.4 Finetuning Gemma 2 9B Instruct

After successfully improving Gemini's performance through fine-tuning, we extend this approach to a smaller model, Gemma 2 9B Instruct, and observe similar accuracy gains, as shown in Figure 6.

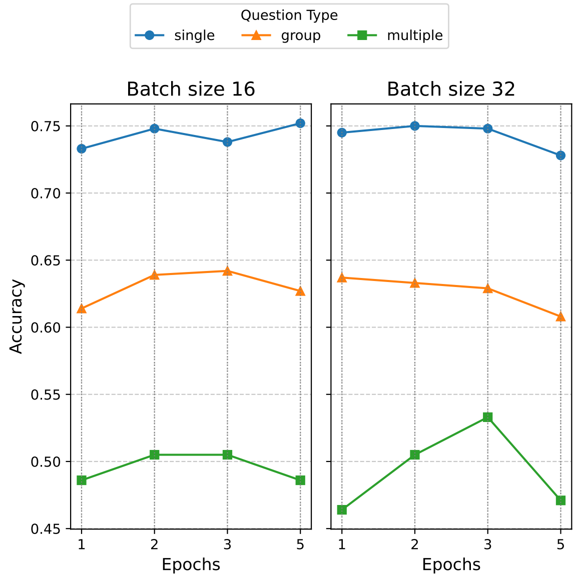

Figure 5: Accuracies of fine-tuned versions of Gemini 1.5 Flash.

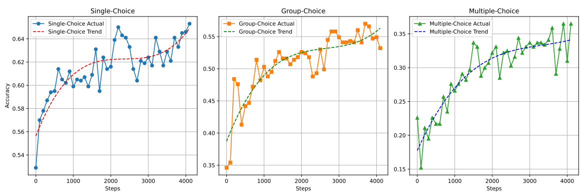

Figure 6: Performance of Gemma 2 9B Instruct on the test split over fine-tuning training steps.

For fine-tuning, we employ the LoRA technique via the Unsloth framework, training the model for approximately four epochs, with 1,000 steps per epoch. Accuracy is evaluated at intervals of 100 steps. While we halted training at four epochs, the observed trend suggests that further improvements may still be possible, particularly for single-choice and group-choice questions.

	Single Acc.	Group Acc.	Multiple Acc.
gemma2-9b-it	0.529	0.346	0.226
finetune step 3700	0.641	**0.570**	0.291
finetune step 3900	0.645	0.547	**0.365**
finetune step 4100	**0.653**	0.532	**0.365**
max increase	0.124	0.186	0.139

Table 4: Best accuracies of the model during fine-tuning.

Table 4 reports the highest accuracies obtained during fine-tuning. Compared to the initial model, Gemma 2 9B Instruct achieves improvements of over 12 percentage points. The fine-tuned model attains performance comparable to larger models, significantly narrowing the gap with Gemini 1.5 Flash on single-choice and multiple-choice questions (falling behind by only 1.5 and 4.1 percentage points, respectively). For group-choice questions, it outperforms all models from the initial benchmark, surpassing the previous state-of-the-art by 3.3 percentage points.

4.5 Treating group choice questions as multiple choice

Inspired by Balepur et al., 2024, we hypothesized that LLMs might struggle to correctly apply the grouping rules, particularly in cases where the multiple-choice accuracy was higher. To test this, we reformulated the questions into a multiple-choice format, ran them as if they were multiple-choice questions, and then manually mapped the groupings to their respective answers.

For cases where the model produces invalid combinations that cannot be mapped to a valid answer, we select the first letter (essentially randomizing the answer). This results in a new accuracy, which sometimes exceeds the original.

To further improve this accuracy, we implemented heuristics instead of relying on the random approach for invalid groups. For example, the combination (1, 2) is mapped to (1, 2, 3); (1) or (3) is mapped to (1, 3); (2, 3, 4) is mapped to (1, 2, 3, 4), and so on. For most models, the use of heuristics yields better results than the random selection, as shown in Table 5.

Model	Group	Group As Multiple	With Heuristics
gemini-2.0-flash-exp	0.537	0.449	0.499
DeepSeek-V3	0.453	0.388	0.423
llama-3.1-405B-Instruct-Turbo	0.426	0.453	0.484
gemini-1.5-flash	0.419	0.447	0.480
gemma2-9b-it	0.346	0.300	0.314
rogemma2-9b-instruct (Q8)	0.298	0.258	0.275
llama3-8b-8192	0.252	0.235	0.245
rollama3-8b-instruct-imat (FP16)	0.235	0.241	0.256
phi-3.5-mini-instruct (F32)	0.208	0.231	0.247

Table 5: The accuracies obtained on group choice questions with all strategies. Highlighting signifies a better score with the group-as-multiple approach compared to the initial strategy.

4.6 Model Ensemble

Building upon the insights from the LLM-Synergy framework proposed by Yang et al., 2023, we implemented a simplified ensemble learning approach to enhance the performance of our models on our dataset. Yang et al. employed Majority Weighted Vote to combine outputs from multiple large language models for biomedical question answering tasks.

In our approach, we formed three distinct model groups, each consisting of three models with very similar individual performances. These groups were as follows: (1) top-performing models, (2) mid-range models, and (3) models that include Romanian language in their fine-tuning. Since the models within each group exhibited comparable accuracies and there are only three models in each group, we used straightforward Majority Voting without the need for assigning weights (the vote results would remain unchanged).

Throughout this experiment, only zero-shot learning has been used, and everything has been computed separately for single, group, and multiple choice questions.

Table 6, 7, and 8 present the results of these ensemble experiments.

Model	Single	Group	Multiple
gemini-2.0-flash	**0.733**	0.524	**0.585**
qwen-max-2025-01-25	0.699	0.472	0.573
llama-3.1-405B-Instruct-Turbo	0.685	0.426	0.464
All of the above combined	0.719	**0.534**	0.560

Table 6: The accuracy of Majority Voting compared to the individual accuracies.

Model	Single	Group	Multiple
DeepSeek-V3	0.665	0.453	**0.474**
gemini-1.5-flash	0.668	0.419	0.406
llama-3.1-405B-Instruct-Turbo	0.685	0.426	0.464
All of the above combined	**0.707**	**0.457**	0.439

Table 7: The accuracy of Majority Voting compared to the individual accuracies.

Model	Single	Group	Multiple
eurollm-9b-instruct (F16)	**0.384**	0.220	**0.102**
rollama3-8b-instruct-imat (FP16)	0.371	0.235	**0.102**
romistral-7b-instruct (Q8)	0.371	0.252	0.077
All of the above combined	0.372	**0.266**	**0.102**

Table 8: The accuracy of Majority Voting compared to the individual accuracies.

Although not by a significant difference, the Majority Voting surpassed the individual performances on group-choice questions in all of the chosen model subsets.

4.7 Accuracy by Stage

We compare the accuracies obtained on questions from the test split, grouped by the competition stage in which they were presented (local, regional, or national), and report the results in Figure 7.

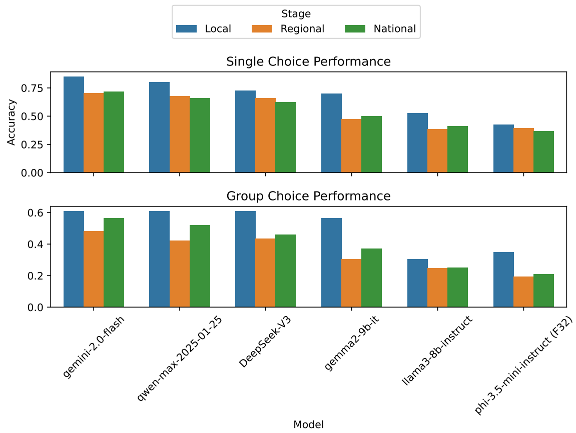

Figure 7: Accuracies of models on different competition stages.

For both single-answer and group-choice questions, models achieve the highest scores on the local stage, confirming that it is indeed the easiest of the three. For single-choice questions, the accuracy remains similar between the regional and national stages, suggesting comparable difficulty levels. However, for group-choice questions, models unexpectedly perform better on the national stage than on the regional stage, despite the expectation that the national stage should be more challenging.

4.8 Accuracy by Grade

We also compare the accuracies obtained on questions, grouped by the corresponding grade level.

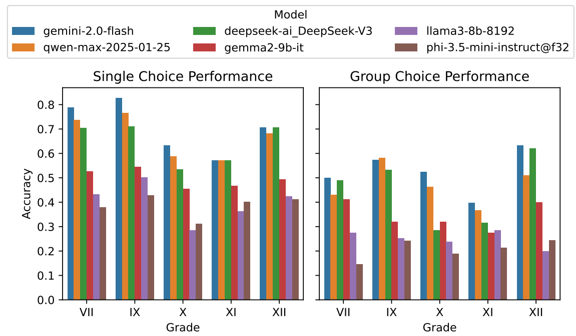

Figure 8: Accuracies of models, grouped by competition grade

As shown in Figure 8, models achieve the lowest scores on grades X and XI, while performing better on grades IX and XII. Performance on grade VII falls between these extremes.

Examining the curricula for these grade levels, we observe a correlation between subject focus and model accuracy. Grades IX and XII emphasize molecular biology and the interactions between biological systems, whereas grades X and XI focus on the physiology and functions of biological systems. Grade VII provides a broad introduction, covering aspects of all these topics while also including basic principles of hygiene and health.

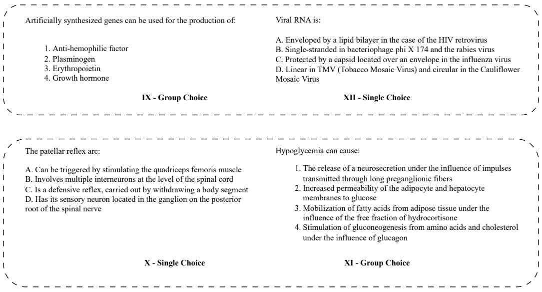

Figure 9: Examples of questions extracted and translated from the dataset

These results suggest that models perform better on topics related to molecular biology and genetics compared to those centered on the physiology of biological systems.

5 Conclusion

This study introduced RoBiologyDataChoiceQA, a novel Romanian-language dataset designed to evaluate the biology comprehension of large language models (LLMs). Sourced from both the Romanian Biology Olympiad and medical school entrance exams, this dataset provides a diverse and challenging benchmark for assessing domain-specific reasoning in a low-resource language.

Our benchmarking experiments revealed significant variations in model performance, highlighting both the strengths and limitations of LLMs in specialized knowledge tasks. While some models performed well on structured, single-answer university admission questions, their ability to handle grouped-choice and reasoning tasks remained inconsistent. Fine-tuning Gemini 1.5 Flash and Gemma 2 9B Instruct improved accuracy in certain cases, demonstrating that targeted adaptation can enhance performance.

Beyond model evaluation, our study offers insights into the impact of prompt engineering, fine-tuning strategies, and dataset characteristics on LLM performance. These findings contribute to the broader effort of advancing NLP applications in non-English languages and specialized scientific domains.

Moving forward, future research should focus on expanding the dataset with fine-grained subdomain annotations to enable deeper biological analysis, improving OCR processing to reduce errors in text extraction from scanned documents, and conducting further experiments with different fine-tuning strategies and model architectures. Additionally, addressing dataset biases by analyzing differences in model performance across Olympiad and university questions could provide valuable insights. Enhancing answer verification through expert validation will also be crucial in ensuring benchmark accuracy.

6 Limitations

While our study provides valuable insights into LLM performance on Romanian-language biology questions, several limitations should be considered when interpreting the results.

- **Lack of fine-grained tagging** – The dataset does not include detailed annotations distinguishing specific biological subdomains (e.g., genetics, physiology, ecology). This limits the ability to analyze model performance at a more granular level and identify knowledge gaps in specialized areas.

- **Potential inaccuracies in answer keys** – Although we rely on authoritative sources, occasional ambiguities or errors in the provided answer keys may affect benchmarking accuracy. While we performed additional verification, some uncertainties remain.

- **Challenges with OCR-extracted data** – The dataset includes content extracted from scanned PDFs, particularly for university admission exams. Despite preprocessing and manual validation, some errors introduced by OCR remain, potentially affecting model training and evaluation.

- **Limited scope of fine-tuning experiments** While we observed improvements when fine-tuning Gemini 1.5 Flash and Gemma 2 9B Instruct, additional experiments with different architectures and training strategies could yield further insights. Exploring other Romanian-adapted models could provide a broader perspective.

- **Domain-specific biases in LLMs** – Our results suggest that models perform better on university admission questions than on Olympiad questions, likely due to differences in training data exposure. Investigating whether this bias stems from pretraining corpora, difficulty of questions, or inherent reasoning limitations could further refine model evaluation.

7 Ethical Statement

To promote transparency and responsible use, we release the dataset under the *Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0)* license. This license allows for non-commercial use, sharing, and adaptation with proper attribution.

No personally identifiable or sensitive information is included in the dataset. We encourage ethical research practices and responsible AI development when using our dataset. However, a potential risk is that it could inadvertently encourage the use of LLMs in biology exams for cheating, rather than for legitimate educational or research purposes. We urge users to adopt responsible policies to prevent misuse in academic settings.

References

Artifex. 2024. Pymupdf4llm: A breakthrough in pdf to markdown conversion for python developers. Accessed: 2025-02-13.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.

Manojit Bhattacharya et al. 2023. Large language model to multimodal large language model: A journey to shape the biological macromolecules to biological sciences and medicine. *Molecular Therapy Nucleic Acids*, 35.

Cristea Costache, Daniela Diaconescu, Andreea Fleancu, Marius Alexandru Moga, Alina Mihaela Pascu, Sebastian Toma, Evelyn Cîrstea, and Alexandra Lazăr. 2020. *Teste de Biologie pentru Admiterea la Facultatea de Medicină [Biology Tests for Admission to the Faculty of Medicine]*, ediția a iii-a, revizuită și completată edition. Editura Universității Transilvania din Brașov, Brașov, Romania.

Cristian-George Crăciun. 2023. RoMedQA v1: A Dataset of Romanian Medical Examination Questions. Hugging Face.

George-Andrei Dima, Andrei-Marius Avram, Cristian-George Craciun, and Dumitru-Clementin Cercel. 2024. RoQLlama: A lightweight Romanian adapted language model. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4531–4541, Miami, Florida, USA. Association for Computational Linguistics.

Irena Gao, Percy Liang, and Carlos Guestrin. 2024. Model equality testing: Which model is this api serving? *arXiv preprint arXiv:2410.20247*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé, and Kate Crawford. 2018. Datasheets for datasets. *Communications of the ACM*, 64:86 – 92.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

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*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. *arXiv preprint*, arXiv:1909.06146.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Nikita Kotwal, Gauri Unnithan, Ashlesh Sheth, and Nehal Kadaganchi. 2021. Optical character recognition using tesseract engine. *International Journal of Engineering Research & Technology*, 10(9):1–5.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models.

Transactions on Machine Learning Research. Featured Certification, Expert Certification.

Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen Tu, Ravi Kumar Satzoda, R Manmatha, and Vijay Mahadevan. 2023. Doctr: Document transformer for structured information extraction in documents. *arXiv preprint arXiv:2307.07929*.

Mihai Masala, Denis C. Ilie-Ablachim, Alexandru Dima, Dragos Corlatescu, Miruna Zavelca, Ovio Olaru, Simina Terian-Dan, Andrei Terian-Dan, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "vorbeşti româneşte?" a recipe to train powerful romanian llms with english instructions.

Petru Matusz, Lavinia Noveanu, Horia Prundeanu, Pusa Gaje, and Carmen Tatu. 2020. *Teste de Biologie pentru Admiterea 2020 la Facultățile de Medicină și Medicină Dentară [Biology Tests for Admission 2020 to the Faculties of Medicine and Dental Medicine]*. Editura Victor Babeș, Timișoara, Romania.

Huy Cong Nguyen, Hai Phong Dang, Thuy Linh Nguyen, Viet Hoang, and Viet Anh Nguyen. 2025. Accuracy of latest large language models in answering multiple choice questions in dentistry: A comparative study. *PLOS ONE*, 20(1):e0317423.

Iulian Opincariu, Bianca Szabo, Carmen Crivii, Adriana Mureșan, Remus Orăsan, and Simona Clichici. 2018. *Biologie. Teste pentru Admitere [Biology. Admission Tests]*, ediția a 10-a revizuită edition. Editura Medicală Universitară „Iuliu Hațieganu”, Cluj-Napoca, Romania.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. In *Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

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.

Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, and Rui Zhang. 2023. One LLM is not enough: Harnessing the power of ensemble learning for medical question answering. *medRxiv*.

A Datasheet

A.1 Motivation for Dataset Creation

Why was the dataset created?

The dataset was developed to assess and enhance the performance of large language models (LLMs) on domain-specific tasks, specifically Romanian biology tests. It offers choice-based questions to evaluate LLM accuracy and can also be used for fine-tuning LLMs to understand specialized Romanian biology terminology.

What (other) tasks could the dataset be used for?

One potential application of this dataset is its use as training data for models designed to generate multiple-choice questions. Additionally, the dataset could be utilized for automatically assessing question difficulty.

A.2 Dataset Composition

What are the instances?

The instances consist of (single, group, or multiple) choice questions sourced from Romanian biology olympiads and college admission exam books. Each question is paired with its correct answer(s), extracted from the corresponding answer keys. Additional identifying information is also appended to each instance, as detailed in the following paragraphs.

Are relationships between instances made explicit in the data?

Yes, relationships between instances are explicitly marked. Using question identification metadata, instances can be grouped by attributes such as source, year, grade, and stage. When identical questions with identical answer options appear across different tests or problem sets, they are assigned a shared *dupe_id*.

Duplicates are retained rather than removed for several reasons:

- To analyze patterns of data repetition (e.g., identifying sources of inspiration between tests).
- To avoid arbitrarily deciding which instance to delete, leaving duplicate removal to the user's discretion.

All known duplicates are included exclusively in the training split.

How many instances of each type are there?
The dataset contains a total of 14,109 extracted questions:

- Single choice: 6,021
- Group choice: 3,918
- Multiple choice: 4,170

Of these, 8,021 questions are sourced from biology olympiads, while 6,088 come from college admission books. The tests span multiple years (2004-2024), although they are not uniformly distributed.

What data does each instance consist of?
We will explain each field:

- **question_number** = an integer stored as string; for olympiads it takes values from 1 to 80. Most tests tend to have at most 60, but the very old ones (2004) do not quite respect the format. As for college admissions, those take values from 1 to 800 (not uniformly, there are tests/chapters with random number of questions, no general rule).
- **question** = the question text
- **type** - can be one of the following:
  - *single-choice*: indicating the question has exactly one correct answer.
  - *group-choice*: indicating that the answer is a single letter, which corresponds to a combination of options being true together:

    A - if ONLY the options numbered by 1, 2 and 3 are correct
    B - if ONLY the options numbered by 1 and 3 are correct
    C - if ONLY the options numbered by 2 and 4 are correct
    D - if ONLY the option numbered by 4 is correct
    E - if ALL of the numbered options are correct

    The group choice is the only type that has options identified by numbers, while the others have them identified by letters.
  - *multiple-choice*: indicating that the answer is represented by any alphabetically ordered combination of the given options. Even though it is multiple, the answer CAN STILL be a single letter)
- **options** = a list of texts (usually statements or list of items) that in combination with the question text can be considered true or false. Olympiad tests have 4 options, while college admission tests have 5.
- **grade** = where the test/problem set was extracted from; it takes 6 values: *facultate* (college), *XII*, *XI*, *X*, *IX* (highschool), *VII* (middle school).
- **stage** = for college it is fixed on *admitere* (admission). For olympiad it represents the chain of theoretical importance and difficulty: *locala* -> *judeteana* -> *nationala* (local -> regional -> national).
- **year** = the year (as a string) in which the problem set/test was used in a competition
- **right_answer** = a letter for single-choice and group-choice (check the explanations above) and multiple (non-repeating) letters concatenated in a string with no other characters, in alphabetical order for multiple-choice.
- **source** = *olimpiada* (Olympiad of Biology in Romania) or, in the case of college, the university it was taken from (currently 3 possible values: *UMF Cluj*, *UMF Brașov*, *UMF Timișoara*)
- **id_in_source** = a string that has the purpose of further recognising the question within the problem set it was given, in case of ambiguity. Ensures uniqueness when combined with the other fields recommended for identifying the questions. Keep in mind that it contains spaces.
- **dupe_id** = a UUID that uniquely identifies a group of duplicated questions. The group may contain 2 or more instances. The instance is considered a duplicate if and only if both the question and options are the same (not necessarily in the same order for options). Two texts are considered the same if they are identical/use synonyms for common words/are obviously rephrased versions of each other. If

a text adds extra words but besides that it is identical with another text, it is *not* marked as a duplicate.

For uniquely identifying a question/instance we recommend the following combination of fields:

$$\left\{\begin{array}{l} \texttt{item['year'],} \\ \texttt{item['source'],} \\ \texttt{item['id\_in\_source'],} \\ \texttt{item['grade'],} \\ \texttt{item['stage'],} \\ \texttt{item['question\_number']} \end{array}\right\}$$

Is everything included or does the data rely on external resources?

Everything is included.

Are there recommended data splits or evaluation measures?

The data is currently split into three: train, valid, test. We attempted a uniform distribution of the data, based on both quantity and quality of the data.

Both the *test* and *valid* splits were sampled via the recipe explained below.

First we do a grade-based separation:

- Grade XII: 175 questions
  - 75 national level
  - 100 state level
- Grade XI: 175 questions
  - 75 national level
  - 100 state level
- Grade X: 200 questions
  - 55 national level
  - 125 state level
  - 20 local level
- Grade IX: 250 questions
  - 115 national level
  - 115 state level
  - 20 local level
- Grade VII: 200 questions
  - 85 national level
  - 85 state level
  - 30 local level
- University Level (*Facultate*): 400 questions (detailed division below)

1. *UMF Timișoara*: 200 questions

- 11 chapters total, 18 questions per chapter, except for the *Nervous System*, which has 20 questions due to higher coverage.

2. *UMF Brașov*: 75 questions

- Derived from 15 questions from each synthesis test.

3. *UMF Cluj*: 125 questions

- *Physiology* (for medical assistant students): 8 questions (1 question per chapter for 5 chapters, plus 3 random questions)
- *Anatomy* (for medical assistant students): 8 questions (same structure as *Physiology*)
- *Physiology* (for medical students): 55 questions (4 questions from each of the first 13 chapters, plus 3 questions from Chapter 14)
- *Anatomy* (for medical students): 54 questions (similar to *Physiology*, but only 2 questions from Chapter 14)

Grade-Stage Yearly Distribution

The tables 9, 10, 11 present the yearly distribution of how many questions to select for each grade, per stage: "-" means no data was available for that year, while "*X*" means nothing was selected.

> **Note:** While each split originally contained 1,400 questions (summing everything mentioned above), the validation and test splits have fewer questions than expected. Although duplicates were identified prior to splitting, an additional round of manual duplicate verification was conducted specifically for the validation and test sets. Newly identified duplicates were moved to the training split, reducing the size of the validation and test splits.

A.3 Data Collection Process

How was the data collected?

Olympiad data: Sourced from public online archives, primarily from *olimpiade.ro* (https://www.olimpiade.ro/). Additional data was retrieved through separate online searches when needed.

College admission books: Obtained from private sources. The collected data consists of PDFs, with

	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
VII	-	-	-	-	-	5	5	7	8	8	12	15	15	-	-	-	-	-	-	-	-
IX	2	2	-	-	4	4	-	5	5	5	8	8	8	-	10	12	-	-	12	15	15
X	-	-	-	-	-	-	-	-	-	-	3	3	4	-	5	7	-	-	8	10	15
XI	-	-	-	-	-	-	-	-	-	-	5	5	7	-	8	8	-	-	12	15	15
XII	-	-	-	-	-	-	-	-	-	-	5	5	7	-	8	8	-	-	12	15	15

Table 9: Number of questions to select in test/validation data for each grade in every year from the **national** stage of the olympiad.

	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
VII	-	-	-	-	-	5	5	7	8	12	13	15	-	-	-	-	-	-	-	-	-
IX	1	1	-	-	1	2	2	3	3	3	4	4	6	8	10	12	12	-	13	15	15
X	-	-	-	-	-	-	-	-	-	-	5	5	6	8	10	12	14	-	20	20	25
XI	-	-	-	-	-	-	-	-	-	-	4	4	6	8	8	12	14	-	14	15	15
XII	-	-	-	-	-	-	-	-	-	-	4	4	6	8	8	12	14	-	14	15	15

Table 10: Number of questions to select in test/validation data for each grade in every year from the **regional** stage of the olympiad.

some containing parsable text and others consisting of images that required additional processing.

Who was involved in the data collection process?

The PDF data was collected by us as well as some medical students.

Over what time-frame was the data collected?

It took roughly one month to collect the data.

How was the data associated with each instance acquired?

The data was initially collected as PDF files. To standardize the format, a Word-to-PDF converter was sometimes used. The PDFs either contained parsable text or had text embedded in images. While the quality of some images was questionable, most of the information was successfully recognized.

For PDFs with parsable text, Python libraries were used for data extraction, with occasional manual verification and refactoring. For PDFs containing images, Gemini 1.5 Flash was employed to extract the data. Random sampling was performed to verify the accuracy of the extracted data.

Does the dataset contain all possible instances?

No. Some olympiads, although we know for sure existed, were not found on the internet. Additionally, there is more data collected in PDF format that has not yet been parsed into actual instances.

If the dataset is a sample, then what is the population?

The population includes additional college admissions and olympiads from Romania that can be found and parsed. It can also contain closely related national contests that feature choice-based questions, which could be included.

Is there information missing from the dataset and why?

Questions that included images/figures were removed as this is not a multi-modal dataset (at the moment).

Are there any known errors, sources of noise, or redundancies in the data?

There are several potential sources of error and redundancy in the data:

- *Parsing issues*: Questions with options represented as tables might have been parsed incorrectly. Some parsing errors may result in typos (e.g., words broken into two segments) or missing words at the end of an option. Many of these errors have been manually corrected, especially in the test split, which should be free of such issues.

- *Image noise*: The images for college admissions can present noise, but Gemini 1.5 Flash processed them relatively well. Some hallucinations may still exist, although we manually searched for them.

- *Duplicates*: Some questions and options are duplicated across different problem sets or

	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
VII	X	-	-	-	-	X	X	-	X	X	X	X	X	15	15	-	-	-	-	-	-
IX	X	-	-	-	-	X	-	-	X	X	X	X	X	15	15	-	-	-	-	-	-
X	X	-	-	-	-	X	-	-	X	X	X	-	X	10	10	-	-	-	-	-	-
XI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
XII	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 11: Number of questions to select in test/validation data for each grade in every year from the **local** stage of the olympiad.

even within the same source. We have marked the obvious duplicates, but repetition of questions and answer options could still occur.

- *Answer errors*: Some answers might be wrong due to parsing errors or LLM hallucinations. Although we have manually checked every parsed answer, human error is still a possibility. Additionally, there could be mistakes in the original answer sheets, where wrong answers may have been transcribed. Despite thorough checks (as the collected data is from national contests with official sources), it is possible that a few incorrect answers might have slipped through.

- *Image dependent questions*: We have tried to filter out any question that was dependent on a figure, as we do not intend for the dataset at the moment to be multi-modal, but some questions might have slipped through. This is possible only for the olympiad questions.

A.4 Data Pre-processing

What pre-processing/cleaning was done?

After extraction, several pre-processing and cleaning steps were applied to standardize and structure the data:

1. Extracted the question number from the question text and placed it in a separate field.

2. Standardized option identifiers to uppercase letters.

3. Ensured all options followed the structure: `"[identifier]. [text]"`, where `[identifier]` is either a letter (*A–D*, or *A-E* for five-option lists) or a number (*1–4* for group-choice questions).

4. Replaced multiple spaces with a single space.

5. Replaced newline characters with spaces.

6. Standardized quotes by replacing Romanian quotation marks with English ones.

7. Normalized diacritics to proper Romanian characters (e.g., `ș, ț, â, ă`).

8. Manually corrected grammar issues and typos.

9. Removed trailing characters such as commas, dots, spaces, and semicolons from option texts.

10. Made Gemini 1.5 Flash act as a grammar correcting tool to help us further find typos. Manually checked the output of it as the LLM has a tendency to replace words besides the typos. (Also used Gemma-2-9B when Gemini 1.5 Flash was unavailable).

Was the "raw" data saved in addition to the preprocessed/cleaned data?

The PDF files are saved privately.

Is the pre-processing software available?

No.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

This dataset successfully provides specialized (Romanian) biology terms that can be used for training or knowledge evaluation.

B Prompts

User Prompts Used for Benchmarking

Single Choice

> %question-text%
>
> You received a biology question in Romanian with multiple options. The biology question is collected from either national high school olympiads or admission exams for medical universities. Only one answer is correct.
>
> You will output only the letter of the right answer. Do not give any explanations.
>
> The letter of the right answer is:

Group Choice

%question-text%

You received a biology question in Romanian with multiple numbered options. The question is from national high school olympiads or medical university admission exams.

To answer:
1. Identify correct options.
2. If only option 4 is correct, the answer must be D.
3. If only options 1,3 are correct, the answer must be B.
4. If only options 2,4 are correct, the answer must be C.
5. If only options 1,2,3 are correct, the answer must be A.
6. If all options are correct, the answer must be E.

Do not give any explanations.

The right answer is:

Multiple Choice

%question-text%

You received a biology question in Romanian with multiple options. The question is from national high school olympiads or medical university admission exams. One or multiple answers are correct.

You will output the letter(s) of all the correct answers. Do not give any explanations.

The letters of the right answers, as compact as possible, are:

System Prompts Used for Benchmarking

We include only five-shot prompts; one- and three-shot follow the same format with fewer questions. The displayed prompts use translated questions, but LLMs receive the original Romanian versions.

Single Choice - Five Shot

Here are some examples of biology questions in Romanian with multiple options and the correct format for answering them:

Question: The prokaryotic cell:
A. characterizes viruses, bacteria, and blue-green algae
B. contains peptidoglycan in the composition of the cell membrane
C. does not have a cell wall
D. the nuclear material is a circular double-stranded DNA molecule
Answer: D
—
Question: The mesosomes of prokaryotes:
A. have a role in respiration
B. are made up of rRNA and proteins
C. are invaginations of the plasma membrane in the form of lamellae
D. have a role in photosynthesis
Answer: A
—
Question: The sciatic nerve:
A. is a cranial nerve
B. contains only motor fibers
C. contains both sensory and motor fibers
D. originates in the medulla oblongata
Answer: C
—
Question: Contain hydrolytic enzymes with a role in intracellular digestion:
A. ribosomes
B. lysosomes
C. centrosome
D. centrioles
Answer: B
—
Question: Photosynthetic plastids are:
A. oleoplasts
B. leucoplasts
C. rhodoplasts
D. amyloplasts
Answer: C

Group Choice - Five Shot

Here are some examples of biology questions in Romanian with multiple numbered options and the correct format for answering them:

Question: Organic substances with a structural role include:
1. lipids
2. carbohydrates
3. proteins
4. nucleic acids
Explanation: 1,3 are correct; 2,4 are not
Answer: B
—
Question: The fundamental substance is present in the structure of:
1. mitochondria
2. chloroplasts
3. the nucleus
4. vacuoles
Explanation: 1,2,3 are correct; 4 is not
Answer: A
—
Question: The nucleolus:
1. is surrounded by its own membrane
2. is the densest part of the nucleus
3. is the site of mRNA synthesis
4. its volume depends on the physiological state of the cell
Explanation: 2,4 are correct; 1,3 are not
Answer: C
—
Question: The granum of chloroplasts:
1. is found freely in the stroma
2. contains DNA, RNA, proteins, and metals
3. is surrounded by a double porous membrane
4. contains photosynthetic pigments
Explanation: 4 is correct; 1,2,3 are not
Answer: D
— # Question: The interphase:
1. represents the time interval between two successive cell divisions
2. is characterized by DNA, RNA, and protein synthesis
3. is the most metabolically active stage
4. precedes the division phase of the cell cycle
Explanation: 1,2,3,4 are correct
Answer: E

Multiple Choice - Five Shot

Here are some examples of biology questions in Romanian with multiple options and the correct format for answering them:

Question: The heart:
A. has the mitral valve between the right atrium and right ventricle
B. is equipped with trabeculae in the atria
C. is a parenchymatous organ due to its strong ventricular musculature
D. is equipped with 2 valves
E. contains the His bundle, which plays a role in automatism with a discharge frequency of 25 impulses/min
Answer: E
—
Question: The right atrium is characterized by:
A. containing the sinoatrial node
B. having trabeculae inside
C. receiving the inferior venae cavae
D. having a systole duration of 0.1s
E. being the site where pulmonary veins open
Answer: ACD
—
Question: The following associations are correct:
A. chordae tendineae - atrioventricular valves
B. sinoatrial node - interatrial septum
C. cardiac cycle - 0.8s at a heart rate of 100 beats/min
D. venous pressure at the level of the right atrium is 10 mmHg
E. tricuspid valve - right atrioventricular orifice
Answer: AE
—
Question: Arteries that originate directly from the subclavian artery include:
A. external carotid
B. vertebral
C. brachial
D. internal thoracic
E. anterior intercostal
Answer: BD
—
Question: The pulmonary veins:
A. are two in number
B. open into the left atrium, which contains the sinoatrial node
C. are part of the small circulation, which begins in the right ventricle
D. bring oxygenated blood to the heart from the alveolar-capillary membrane, which has an average thickness of 0.6 microns
E. like the venae cavae, bring venous blood into the atria
Answer: CD