GeoLogicQA - A Benchmark for Evaluating Logical Reasoning in Georgian For Large Language Models

Irakli Koberidze
Dep. of Computer Science
Tbilisi State University
Tbilisi Georgia
irakli.koberidze@tsu.ge

Archil Elizbarashvili
Dep. of Computer Science
Tbilisi State University
Tbilisi Georgia
archil.elizbarashvili@tsu.ge

Magda Tsintsadze
Dep. of Computer Science
Tbilisi State University
Tbilisi Georgia
magda.tsintsadze@tsu.ge

Abstract

Advancements in LLMs have largely overlooked low-resource languages (LRLs), creating a gap in evaluation benchmarks. To address this for Georgian, a Kartvelian language, we introduce GeoLogicQA. This novel, manually-curated benchmark assesses LLMs' logical and inferential reasoning through 100 questions. Questions cover syllogistic deduction, inferential reading comprehension, common-sense reasoning, and arithmetic, adapted from challenging sources (Kangaroo Mathematics Competition) and validated by native Georgian speakers for linguistic nuances. Initial evaluations of state-of-the-art LLMs (Gemini 2.5 Flash, DeepSeek-V3, Grok-3, GPT-4o) show an average accuracy of 64% to 83%, significantly exceeding the human baseline of 47%. While demonstrating strong reasoning potential, error analysis reveals persistent challenges in multi-step combinatorial and highly constrained inferential tasks. GeoLogicQA is a public resource for tracking progress and diagnosing weaknesses in Georgian LLMs. We plan to expand the benchmark and establish a public leader-board to foster continuous improvement.

1 Introduction

The rapid evolution of Large Language Models (LLMs), like GPT-4 and Llama 3, has revolutionized AI, excelling in natural language generation and problem-solving. This progress is largely due to vast computational resources and datasets in high-resource languages (HRLs), primarily English (OpenAI et al., 2024). Consequently, low-resource languages (LRLs) face a significant disparity in LLM development and evaluation, lacking appropriate benchmarks to assess their true capabilities.

Georgian, a Kartvelian agglutinative language, exemplifies this resource gap in NLP. Existing Georgian NLP resources are insufficient for evaluating the deeper cognitive abilities of modern generative LLMs, failing to test complex logical reasoning, inferential comprehension, or common-sense understanding. The critical question for Georgian NLP has evolved from "Can a model process Georgian text?" to "Can a model think in Georgian?", requiring evaluation beyond just a pattern recognition.

This work introduces GeoLogicQA, a novel, manually-curated evaluation benchmark designed to assess the logical and inferential reasoning abilities of LLMs in Georgian language. It offers a diverse set of multiple-choice questions covering syllogistic deduction, inferential reading comprehension, common-sense reasoning, and arithmetic problem-solving, aiming to diagnose model weaknesses. Our primary goal is to provide a rigorous, publicly accessible resource for tracking progress and foster more robust Georgian LLMs. Initial evaluations of models like ChatGPT, DeepSeek, Gemini, and Grok on our 100-question benchmark show an average performance below 70% accuracy, highlighting significant challenges in complex Georgian reasoning.

The paper is structured as follows: Section 2 reviews existing LLM evaluation benchmarks and Georgian NLP resources. Section 3 details GeoLogicQA's design principles, task categories, and data curation. Section 4 outlines the experimental setup, presents baseline results from LLM evaluations on GeoLogicQA, and analyzes performance and error patterns, followed by section 5, were implications, limitations, and future work is discussed. The work is concluded by section 6, which summarizes

Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languages associated with RANLP 2025, pages 121–130, Varna, Bulgaria, Sep 13, 2025.
https://doi.org/10.26615/978-954-452-100-4-013

the contributions and emphasizes the broader impact of the resource.

2 Background and Related Work

The landscape of Natural Language Processing has been significantly shaped by the development of sophisticated evaluation benchmarks that measure the capabilities of large language models. These benchmarks serve as crucial instruments for comparing models, identifying their strengths, and diagnosing their limitations (Wang et al., 2019). Our work on GeoLogicQA is situated within this broader context, while simultaneously addressing the unique challenges presented by low-resource languages.

2.1 Benchmark Creation for LRLs

Benchmark is a set of standardized tests that assess LLM performance across various tasks. Creating benchmarks for LRLs involves several steps that ensure the evaluation is meaningful, fair, and generalizable across models. The challenges faced by Georgian NLP are not unique, many low-resource languages worldwide contend with similar limitations in terms of data availability and evaluation infrastructure. Consequently, there has been a growing global movement within the NLP community to address this disparity by creating dedicated benchmarks for LRLs. Efforts often involve:

- Typical strategies that translate high-resource-language (HRL) benchmarks into low-resource languages (LRLs) often lose culturally-specific context or meaning, limiting faithful assessment of reasoning capabilities in the target LRL (Ghafoor et al., 2021). Further, studies such as (Alhanai et al., 2024) show that direct translations of benchmarks (e.g., Winogrande, MMLU into low-resource African languages) underperform until cultural adjustments are incorporated—highlighting that simple translation fails to preserve the nuanced reasoning demands of the original tasks.

- Collaborative efforts involving native speakers and linguists are crucial for curating high-quality, culturally relevant datasets; empirical evidence shows that native-written corpora enhance lexical diversity and cultural content (Cahyawijaya et al., 2023), while participatory and community-centric approaches ensure linguistic authenticity and foster richer dataset design (Ousidhoum et al., 2025).

- Developing benchmarks tailored to specific linguistic phenomena or reasoning types that are particularly challenging for a given LRL. This approach helps diagnose unique model weaknesses (Goyal and Dan, 2025; Sánchez et al., 2024; Bean et al., 2024).

By developing GeoLogicQA,[1] a manually-curated benchmark for Georgian logical reasoning, we contribute to the crucial effort of creating equitable and culturally-relevant evaluations for LLMs in low-resource languages.

2.2 Evaluation Benchmarks in HRLs

In high-resource languages, particularly English, a rich ecosystem of evaluation benchmarks exists, each targeting different facets of language understanding and reasoning. Prominent examples include:

GLUE (General Language Understanding Evaluation) and its successor, SuperGLUE: A collection of diverse natural language understanding tasks, such as sentiment analysis, textual entailment, question answering, and paraphrase detection (Wang et al., 2019). They assess a model's ability to capture semantic and syntactic nuances across various linguistic phenomena. While foundational, GLUE and SuperGLUE primarily evaluate general language understanding rather than complex, multi-step logical reasoning.

MMLU (Massive Multitask Language Understanding): A significant advancement in evaluating LLMs by testing knowledge and reasoning across 57 diverse subjects, including humanities, social sciences, STEM, and professional disciplines (Hendrycks et al., 2021). It is designed to be challenging, often requiring zero-shot or few-shot inference, and assesses a model's ability to apply pre-trained knowledge to novel problems. MMLU's focus on a wide array of academic and professional subjects makes it a strong indicator of a model's

[1] https://github.com/irakli97/GeoLogicQA.

general intelligence and reasoning capabilities beyond simple pattern matching.

Big-Bench (Beyond the Imitation Game Benchmark): Includes over 200 tasks, many of which are specifically designed to push the boundaries of LLM capabilities, encompassing logical reasoning, common-sense reasoning, mathematical problem-solving, and creative writing (Srivastava et al., 2023). Big-Bench Hard (BBH), a subset of the most challenging Big-Bench tasks, explicitly targets reasoning abilities that are difficult for current LLMs, often involving multi-hop deduction, complex causal relationships, or counterfactual reasoning. These benchmarks provide a robust framework for assessing higher-order cognitive functions in LLMs.

These HRL benchmarks have been instrumental in driving the rapid progress of LLMs by providing standardized, rigorous, and publicly accessible evaluation tools. They allow researchers to track performance, pinpoint weaknesses, and develop more sophisticated models.

2.3 NLP Resources for Georgian

Despite the global advancements in NLP, the Georgian language faces significant challenges due to its low-resource status (Pakray et al., 2025). The availability of high-quality training data and advanced NLP tools for Georgian is notably limited compared to HRLs. Existing resources primarily include several initiatives focused on compiling Georgian text corpora from various sources, such as Wikipedia, news articles, and literary works. These corpora are valuable for foundational tasks like language modeling and morphological analysis (Doborjginidze and Lobzhanidze, 2016). Limited parallel corpora exist for machine translation between Georgian and other languages, supporting cross-lingual transfer. Also, the tools for Part-of-Speech tagging, lemmatization, and dependency parsing have been developed, aiding in basic linguistic analysis (Giorkhelidze, 2017). However, these existing resources predominantly cater to traditional NLP tasks and surface-level linguistic analysis. They largely fall short in providing the challenging, reasoning-focused datasets necessary to evaluate the deep language understanding and inferential capabilities of modern generative LLMs. The scarcity of structured, annotated data designed for complex logical inference means that current Georgian NLP lacks the benchmarks required to gauge how well LLMs can process and reason with Georgian text beyond simple recognition or translation. There is a marked absence of standardized datasets that demand multi-step reasoning, logical deduction, or nuanced common-sense inference in Georgian, creating a significant gap in the evaluation framework for advanced Georgian LLMs.

3 The GeoLogicQA Benchmark: Design and Curation

The GeoLogicQA benchmark is meticulously designed to provide a robust evaluation of Large Language Models' (LLMs) logical and inferential reasoning capabilities specifically within the Georgian language. This section details the core design principles, the diverse task categories included, and the rigorous data collection and validation processes employed to ensure the benchmark's quality and validity. Because of the unique linguistic characteristics of the Georgian language, including its agglutinative nature and prevalent polysemy (Ma et al., 2020), we had to apply careful curation when designing the GeoLogicQA benchmark to overcome these challenges.

3.1 Design Principles

GeoLogicQA's design is underpinned by several core principles aimed at comprehensively assessing LLMs' reasoning in a low-resource language context.

3.1.1 Focus on Logic and Inference

GeoLogicQA explicitly prioritizes tasks that demand genuine logical and inferential reasoning, moving beyond simple keyword matching, surface-level pattern recognition, or statistical correlations. The fundamental aim is to ascertain whether an LLM can truly "think in Georgian," grasping complex relationships and deriving non-explicit conclusions, rather than merely processing and reproducing text. This necessitates questions that require multi-step reasoning, an understanding of causality, and the ability to synthesize information from various premises.

3.1.2 Linguistic and Cultural Nuance

GeoLogicQA deeply integrates Georgian linguistic and cultural nuances. Questions weren't just translated; they were crafted to be natural, culturally relevant, and contextually appropriate for native Georgian speakers. This means scenarios, idioms, and common knowledge referenced in the questions genuinely resonate within the Georgian context, avoiding awkward translations that could distort meaning or reasoning challenges.

Crucially, the design process addressed polysemy in Georgian, where words and phrases can have multiple meanings. For example, "მანძილის დაფარვა" can mean "to cover distance" or "to cover something with a lid." To prevent misinterpretations by LLMs due to linguistic misunderstanding rather than a lack of reasoning, question designers carefully constructed sentences and scenarios. They provided unambiguous contextual cues, ensuring only the intended meaning was conveyed. This precise phrasing was paramount to isolating and testing true reasoning rather than surface-level recognition.

3.1.3 Task Diversity

GeoLogicQA incorporates a diverse range of reasoning task categories to provide a comprehensive assessment of LLMs' cognitive abilities. These categories include syllogistic and deductive reasoning, reading comprehension with inference, common-sense reasoning, and arithmetic reasoning. This diversity ensures that the benchmark evaluates a broad spectrum of reasoning skills, preventing LLMs from excelling based on proficiency in only one type of task.

3.1.4 Quality Assurance

The creation and validation of questions for GeoLogicQA followed a rigorous, multi-stage quality assurance process. Each question was meticulously reviewed to ensure it was unambiguous, logically sound, and genuinely tested complex reasoning rather than simple recall or pattern matching. This iterative process involved expert review and refinement to eliminate any potential for misinterpretation or an unintended correct answer, guaranteeing the integrity of the evaluation.

3.2 Task Categories and Examples

GeoLogicQA comprises four distinct task categories, each designed to probe specific facets of logical and inferential reasoning. The following examples illustrate the type of questions included in each category:

3.2.1 Category 1: Syllogistic & Deductive Reasoning

Description: Tasks requiring deriving a logically sound conclusion from a set of premises. These questions often test a model's ability to follow chains of inference and identify valid deductions.

Example: Georgian: "ყოველ მონეტას აქვს ორი მხარე, 'გერბი' და 'საფასური'. მაგიდაზე ძევს ხუთი მონეტა, ხუთივე ზემოთ იყურება 'გერბით'. ყოველ ბიჯზე უნდა ამოვატრიალოთ ზუსტად სამი მონეტა. იპოვეთ ბიჯების ის უმცირესი რაოდენობა, რომლის შემდეგაც ხუთივე მონეტა ზემოთ იქნება 'საფასურით'." English: "Five coins are lying on a table with the "heads" side up. At each step you must turn over exactly three of the coins. What is the least number of steps required to have all the coins lying with the "tails" side up?"

3.2.2 Category 2: Reading Comprehension with Inference

Description: Presenting a short paragraph or scenario and asking a question where the answer is not explicitly stated but must be inferred from the provided text, requiring deeper understanding and synthesis of information.

Example: Georgian: "კინოთეატრში ერთ რიგში ზის 23 ცხოველი. თითოეული ცხოველი არის ან თახვი ან კენგურუ. თითოეულ ცხოველს ჰყავს სულ მცირე ერთი მეზობელი, რომელიც კენგურუა. ყველაზე მეტი რამდენი თახვი შეიძლება იყოს რიგში?" English: "There are 23 animals sitting in a row at the cinema. Each animal is either a beaver or a kangaroo. Everyone has at least one neighbour who is a kangaroo. What is the largest possible number of beavers in the row?"

3.2.3 Category 3: Common-Sense Reasoning

Description: Questions relying on implicit, everyday knowledge about the world and practical understanding of cause-and-effect rela-

tionships, adapted for a Georgian cultural context.

Example: Georgian: "აისბერგს კუბის ფორმა აქვს. მისი მოცულობის 90% არის წყლის ზედაპირის ქვემოთ. წყლის ზედაპირის ზემოთ ჩანს კუბის მხოლოდ სამი წიბოს ნაწილი. ამ ნაწილების სიგრძეებია: 24 მ, 25 მ და 27 მ. იპოვეთ კუბის წიბოს სიგრძე." English: "An iceberg has the shape of a cube. Exactly 90% of its volume is hidden below the surface of the water. Three edges of the cube are partially visible over the water. The visible parts of these edges are 24m, 25m and 27m. How long is an edge of the cube?"

3.2.4 Category 4: Arithmetic Reasoning

Description: Word problems that require extracting numerical quantities, understanding relationships, and performing basic to moderately complex calculations within a narrative context.

Example: Georgian: "იპოვეთ $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$ ნამრავლის ბოლო ორი ციფრის ჯამი." English: "What is the sum of the last two digits of the product $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$?"

3.3 Data Collection and Validation

The quality and challenge of GeoLogicQA are rooted in its careful data collection and rigorous validation processes.

3.3.1 Source

The questions used in GeoLogicQA were adapted from the annual Kangaroo Mathematics Competitions, organized by the "Association Kangourou sans Frontières (AKSF)" (https://www.aksf.org) and officially translated into Georgian by its representatives in Georgia. We primarily selected problems from the 9th to 12th-grade levels to ensure a high level of cognitive demand and complexity, making them suitable for evaluating advanced reasoning capabilities in LLMs.

Crucial modifications were made to ensure AI interpretability without altering the core reasoning challenge. This involved standardizing mathematical notations (e.g., using ^ for powers), adding parentheses for clarity, and converting essential visual information from image-based questions into descriptive text. This last step was only done when the underlying logical reasoning could be fully preserved without the visual component, avoiding the need for computer vision. We obtained explicit permission from AKSF for ethical data sourcing.

3.3.2 Validation Process

A rigorous multi-step verification process was implemented to ensure the quality, clarity, and correctness of each question and its intended answer:

- **Initial Drafting and Adaptation:** The core research team was responsible for the initial drafting and adaptation of questions from the source materials, ensuring adherence to the design principles.
- **Expert Review by Native Speakers:** Each adapted question underwent rigorous review by a panel of at least two independent native Georgian speakers. This panel critically evaluated each question for linguistic clarity, potential ambiguities (particularly addressing polysemy), naturalness of expression, and the unequivocal correctness of the designated answer. This step was paramount in refining the questions for precision and ensuring they truly assessed reasoning in Georgian, eliminating any linguistic pitfalls that might mislead an LLM.
- **Pilot Testing on Human Subjects:** A subset of the questions was pilot tested on human subjects to establish a human performance baseline. These selected questions are notoriously challenging for human students, as evidenced by the published average of 47% correct answers by students on the Kangaroo Mathematics Competition questions.[2] It is important to note that this relatively high percentage reflects the fact that participants in the upper grades of the competition are typically students with a strong interest and background in mathematics. This human baseline provides a vital context for evaluating LLM performance, highlighting the benchmark's inherent difficulty even for human solvers.

[2]Source: https://kenguru.ge/olympiad.

3.3.3 Statistics

The benchmark comprises of 100 questions.

4 Experimental Setup and Baseline Results

This section details the experimental methodology employed to evaluate large language models (LLMs) on the GeoLogicQA benchmark and presents the baseline results. We describe the specific models chosen, the evaluation protocol, and an in-depth analysis of their performance, including an error breakdown to highlight common challenges.

4.1 LLM Models

For the evaluation of logical and inferential reasoning capabilities in Georgian, a selection of advanced large language models was chosen. The models evaluated were **GPT-4o, Gemini 2.5 Flash, DeepSeek-V3, and Grok-3**. These models represent a diverse set of current state-of-the-art LLMs, offering a robust comparison of their performance on complex reasoning tasks in a low-resource language.

4.2 Evaluation Protocol

To ensure a consistent and fair assessment of each model's inherent reasoning abilities, a standardized evaluation protocol was strictly adhered to.

4.2.1 Prompting Strategy

For all evaluations, a **zero-shot prompting strategy** was employed. The full question text in Georgian, as presented in the GeoLogicQA benchmark, was directly submitted to each model without any additional instructions, examples, or specific formatting cues. This approach was chosen to assess the models' inherent reasoning capabilities in Georgian without external scaffolding. This method provides a direct measure of how well models understand and respond to novel, complex questions solely based on their pre-trained knowledge and reasoning faculties.

4.2.2 Metric

The performance of each LLM was quantified by its **accuracy**, defined as the percentage of correctly answered questions out of the total 100 questions in the GeoLogicQA benchmark. A correct answer was determined by an exact match with the ground truth solution. This binary metric provides a clear and unambiguous measure of successful reasoning.

4.3 Results

The baseline performance of the evaluated LLMs on the GeoLogicQA benchmark is presented in Table 1. For comparison, we include a human baseline derived from the performance of 9th to 12th-grade students on the adapted questions from the annual Kangaroo Mathematics Competition in Georgia.[3]

Model	Accuracy (%)
Gemini 2.5 Flash	83.00
DeepSeek-V3	74.00
Grok-3	67.00
GPT-4o	64.00
Human Baseline	47.0

Table 1: Baseline performance of evaluated LLMs and human subjects on the GeoLogicQA benchmark.

4.4 Analysis and Error Breakdown

The results demonstrate a clear hierarchy in performance among the evaluated LLMs, with **Gemini 2.5 Flash** emerging as the top-performing model, achieving an accuracy of 83.00%. Following closely were **DeepSeek-V3** (74.00%), **Grok-3** (67.00%), and **GPT-4o** (64.00%). A significant observation is that all evaluated LLMs substantially surpassed the **human baseline performance of 47.0%**. This indicates that current advanced LLMs possess a considerable advantage over human subjects on these specific types of logical and inferential reasoning tasks in Georgian, despite the benchmark's design to challenge models in a low-resource linguistic context.

While all models performed well above the human baseline, an analysis of specific errors reveals common challenging categories and unique failure modes. Syllogistic and Deductive Reasoning questions, as well as complex Arithmetic Reasoning tasks, often proved to be the most difficult for the models, aligning

[3]Data adapted from the official Georgian Kangaroo Competition statistics, available at https://www.kenguru.ge/posts/7700b50e-a89e-41c2-a6c8-24cab065b424.

with the inherent complexity of these problem types.

To illustrate, consider the following examples of observed errors:

- In a **Syllogistic & Deductive Reasoning** question about flipping five coins, where "Five coins are lying on a table with the "heads" side up. At each step you must turn over exactly three of the coins. What is the least number of steps required to have all the coins lying with the "tails" side up?" **all evaluated models failed** to provide the correct minimum number of steps. This suggests a fundamental challenge in multi-step combinatorial reasoning.

- For an **Arithmetic Reasoning** question that asked to "What is the sum of the last two digits of the product $1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$?" **Gemini 2.5 Flash and DeepSeek-V3 correctly identified the answer**, while **GPT-4o and Grok-3 were incorrect**. This highlights varying levels of numerical reasoning and attention to detail among the models for specific arithmetic properties.

- A **Reading Comprehension with Inference** question posed a scenario: "In a cinema row, there are 23 animals sitting. Each animal is either a beaver or a kangaroo. Every animal has at least one neighbor who is a kangaroo. What is the maximum number of beavers there can be in the row?" Interestingly, **Grok-3 was the only model to correctly answer this question**, whereas Gemini 2.5 Flash, GPT-4o, and DeepSeek-V3 all failed. This particular instance points to Grok-3's potentially stronger ability to handle complex conditional constraints and infer maximum possibilities in a constrained environment.

- Another commonly challenging problem involved determining a specific digit in a product of "six consecutive numbers" forming a 12-digit number of the form 'abb cdd cdd abb'. In this **Arithmetic Reasoning** task, **Gemini 2.5 Flash and DeepSeek-V3 provided the correct answer**, while **GPT-4o and Grok-3 did not**. This error pattern indicates that some models struggle more with reverse engineering numerical properties or identifying specific digit values within large products based on structural constraints.

These examples underscore that while LLMs show robust performance on average, specific types of logical puzzles and intricate numerical challenges continue to pose significant hurdles, revealing areas for future model improvement in handling complex reasoning in Georgian.

5 Discussion

5.1 Key Takeaways

State-of-the-art Large Language Models (LLMs) consistently outperformed the human baseline of 47.0% on GeoLogicQA, a benchmark for logical and inferential reasoning in Georgian. Gemini 2.5 Flash led with 83.00% accuracy, followed by DeepSeek-V3 (74.00%), Grok-3 (67.00%), and GPT-4o (64.00%). This demonstrates a significant advantage for LLMs in structured logical and arithmetic problems, even in a low-resource language like Georgian. This performance gap underscores the rapid advancements in LLM reasoning. While LLMs excel at precise, multi-step deduction, they still struggle with complex multi-step combinatorial problems and nuanced inferential reading comprehension requiring the synthesis of multiple constraints. The varied performance across problem types highlights that no single model is universally superior, emphasizing the need for continued refinement in intricate logical deductions within low-resource language contexts.

5.2 Future Work

Building upon GeoLogicQA's initial release, our future work will focus on several key directions to expand its utility and impact:

- **Benchmark Expansion:** We plan to significantly expand the GeoLogicQA dataset by curating hundreds of additional questions. This expansion will not only increase statistical robustness but also allow for the inclusion of new task

categories. Potential additions include questions testing understanding of figurative language, detection of logical fallacies in natural arguments, and more complex causal reasoning scenarios that require deeper narrative comprehension.

- **Public Leaderboard and Community Contributions:** To foster continuous progress and facilitate comparative research, we intend to establish a publicly accessible leaderboard. This platform will allow researchers to submit their models' performance on GeoLogicQA, tracking advancements in Georgian LLM reasoning over time. Furthermore, we will actively encourage community contributions to the benchmark, inviting native Georgian speakers, linguists, and AI researchers to propose new questions and reasoning challenges. This collaborative approach will ensure the benchmark remains dynamic, comprehensive, and reflective of the evolving needs of the Georgian NLP community.

6 Conclusion

This paper introduces GeoLogicQA, the first dedicated benchmark for evaluating logical and inferential reasoning capabilities of Large Language Models in the Georgian language. Through meticulous manual curation and rigorous validation, GeoLogicQA provides a challenging set of 100 questions spanning syllogistic deduction, inferential reading comprehension, common-sense reasoning, and arithmetic problem-solving. Our baseline evaluations demonstrate that contemporary LLMs, notably Gemini 2.5 Flash, DeepSeek-V3, Grok-3, and GPT-4o, significantly outperform human subjects on these complex Georgian reasoning tasks, highlighting the advanced logical capabilities of current models even in low-resource linguistic contexts.

The creation and public release of GeoLogicQA address a critical gap in the evaluation infrastructure for Georgian Natural Language Processing, moving beyond superficial linguistic analysis to probe deeper cognitive abilities. This benchmark will serve as a vital resource for the research community, enabling systematic tracking of progress, identifying specific areas for model improvement, and fostering the development of more robust and intelligent LLMs for Georgian. As we continue to expand and refine GeoLogicQA, we emphasize the urgent and ongoing need for community-driven resource creation to ensure equitable and comprehensive AI development across the world's diverse linguistic landscape, ultimately paving the way for truly multilingual and reasoning-capable AI systems.

Limitations of GeoLogicQA

While a valuable step, GeoLogicQA has limitations. Firstly, its modest size of 100 questions limits statistical confidence compared to larger benchmarks, hindering comprehensive analysis across diverse logical challenges. Secondly, GeoLogicQA primarily focuses on structured logical, inferential, and arithmetic reasoning, lacking coverage of broader human-like reasoning. It currently omits common-sense reasoning (e.g., social understanding, ethical dilemmas, logical fallacy detection) and deep understanding of Georgian cultural nuances like idioms or proverbs. Finally, its reliance on adapted Math competition questions, though ensuring high cognitive demand, constrains the scope to formalized problems with single correct answers. This may not fully capture the breadth of real-world, open-ended, ambiguous, or creative reasoning challenges.

References

Alhanai, T., Kasumovic, A., Ghassemi, M., Zitzelberger, A., Lundin, J., and Chabot-Couture, G. (2024). Bridging the gap: Enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments.

Bean, A. M., Hellsten, S., Mayne, H., Magomere, J., Chi, E. A., Chi, R., Hale, S. A., and Kirk, H. R. (2024). Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages.

Cahyawijaya, S., Lovenia, H., Koto, F., Adhista, D., Dave, E., Oktavianti, S., Akbar, S. M., Lee, J., Shadieq, N., Cenggoro, T. W., Linuwih, H. W., Wilie, B., Muridan, G. P., Winata, G. I., Moeljadi, D., Aji, A. F., Purwarianti, A., and Fung, P. (2023). Nusawrites: Constructing high-quality corpora for underrepresented and extremely low-resource languages.

Doborjginidze, N. and Lobzhanidze, I. (2016). Corpus of the georgian language. In *Proceedings of the XVII EURALEX International Congress*, pages 328–335.

Ghafoor, A., Imran, A. S., Daudpota, S. M., Kastrati, Z., Abdullah, Batra, R., and Wani, M. A. (2021). The impact of translating resource-rich datasets to low-resource languages through multi-lingual text processing. *IEEE Access*, 9:124478–124490.

Giorkhelidze, G. (2017). Software tools for initial processing of georgian texts. In *SENS-2017 Conference*. Accessed: 2025-07-18.

Goyal, S. and Dan, S. (2025). Iolbench: Benchmarking llms on linguistic reasoning.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2021). Measuring massive multitask language understanding.

Ma, R., Jin, L., Liu, Q., Chen, L., and Yu, K. (2020). Addressing the polysemy problem in language modeling with attentional multi-sense embeddings. pages 8129–8133.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. (2024). Gpt-4 technical report.

Ousidhoum, N., Beloucif, M., and Mohammad, S. M. (2025). Building better: Avoiding pitfalls in developing language resources when data is scarce.

Pakray, P., Gelbukh, A., and Bandyopadhyay, S. (2025). Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S.,

Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C. F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K. D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L. O., Metz, L., Şenel, L. K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M. J. R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S. T., Shieber, S. M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V. U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z., and Wu, Z. (2023). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.

Sánchez, E., Alastruey, B., Ropers, C., Stenetorp, P., Artetxe, M., and Costa-jussà, M. R. (2024). Linguini: A benchmark for language-agnostic linguistic reasoning.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. (2019). Glue: A multi-task benchmark and analysis platform for natural language understanding.