IOLBENCH: Benchmarking LLMs on Linguistic Reasoning

Anonymous ACL submission

Abstract

Despite the remarkable advancements and widespread applications of deep neural networks, their ability to perform reasoning tasks remains limited, particularly in domains requiring structured, abstract thought. In this paper, we investigate the linguistic reasoning capabilities of state-of-the-art large language models (LLMs) by introducing IOLBENCH, a novel benchmark derived from International Linguistics Olympiad (IOL) problems. This dataset encompasses diverse problems testing syntax, morphology, phonology, and semantics, all carefully designed to be self-contained and independent of external knowledge. These tasks challenge models to engage in metacognitive linguistic reasoning, requiring the deduction of linguistic rules and patterns from minimal examples.

Through extensive benchmarking of leading LLMs, we find that even the most advanced models struggle to handle the intricacies of linguistic complexity, particularly in areas demanding compositional generalization and rule abstraction. Our analysis highlights both the strengths and persistent limitations of current models in linguistic problem-solving, offering valuable insights into their reasoning capabilities. By introducing IOLBENCH, we aim to foster further research into developing models capable of human-like reasoning, with broader implications for the fields of computational linguistics and artificial intelligence.

1 Introduction

The emergence of Large Language Models (LLMs) has led to growing interest in their capacity to perform reasoning tasks that extend beyond statistical pattern matching, particularly in domains requiring compositional generalization. Compositional reasoning involves deducing abstract rules and structures from a limited set of examples and applying these deductions to novel inputs—a process akin

Problem #5 (20 points). Here are some words and word combinations in Rotuman as well as their English translations:

'el'ele	shallow	*mamasa*	solid
'ele	to be near	*mạtiti*	coldness
'olo	to cut	*mạtit mamasa*	ice
a'öf fau	year's end	*moafmofa*	littered with rubbish
fäeag 'u'u	to use sign language	*niu*	copra
fau	year	*nu'suar tiro*	window
hạfhạfu	rocky	*nu'sura*	door
huag 'el'ele	impatient	*pala*	pierce
huag to'a	courageous	*piri*	to curl
hül hạfu	to blow (of a hurricane)	*poagpoga = palpala*	covered with holes
hün kia	base of the neck	*pogi*	night
huli	to turn over	*puhrạki*	to boil, to bubble up
huni	lower end	*pulu*	glue
is 'ā	sharp-pointed	*kạlu*	bracelet; to encircle
is susu	nipple	*riamrima*	shiny
lala	deep	*rū huga*	stomach-ache
maf tiro	spectacles	*to'a*	hero

(a) Here are the Rotuman names of seven body parts as well as their English translations in arbitrary order. Determine the correct correspondences:

'u'u, isu, kia, leva, mafa, susu, huga
breast, eye, arm/hand, hair, heart, neck, nose

(b) Translate into English:
tiro, poga (noun), *huag lala, hạf puhrạki, maf pogi = maf pala.*

(c) Translate into Rotuman:
round; to cut copra; curly hair; sticky; to flash; rubbish.

(d) Using the material given above you cannot translate 'word' and 'to exhaust' into Rotuman with certainty. What would the theoretically possible translations of these words into Rotuman be?

Problem #5. When two words constitute a phrase, the form of the first word undergoes the following change:

$$-VCV \rightarrow -VC$$
$$-VCu \rightarrow -VC$$
$$-VCi \rightarrow -\bar{V}C$$
$$-VCa \rightarrow -VaC$$

(V: vowel, C: consonant).

The same thing happens when an adjective is formed by doubling a noun or a verb: *'ele* + *'ele* → *'el'ele* 'to be near × 2 = shallow'.

The word order is

- $\boxed{N_1 \; N_2}$ (N_1: modified, N_2: modifier);
- $\boxed{N \; A}$ (also with the meaning 'one who has A N': *huag 'el'ele* 'heart + shallow = impatient');
- $\boxed{V \; O}$ (the resulting compound word can be a noun or a verb: *a'öf fau* 'to exhaust + year = year's end', *hül hạfu* 'to turn over + rock = to blow (of a hurricane)').

(a) *'u'u* — arm/hand, *isu* — nose, *kia* — neck, *leva* — hair, *mafa* — eye, *susu* — breast, *huga* — heart.

(b) *tiro* — glass (stuff),
poga — hole,
huag lala — patient,
hạf puhrạki — volcanic rock,
maf pogi = maf pala — blind.

(c) round — *kạlkạlu*; to cut copra — *'ol niu*; curly hair — *leav pirpiri*; sticky — *pulpulu*; to flash — *rima*; rubbish — *mofa*.

(d) • word: *fäega* (or *fäeaga, fäeagu*).
• to exhaust: *a'ofi* (or *a'öfi, a'öfö, a'öfu, a'öfü, a'ofü*).

Figure 1: An example problem from IOL 2012 involving the Austronesian language Rotuman, spoken by roughly 9000 people in Fiji. The problem involves lexical matching and translations, and the solution explains how these can be deduced from the provided dictionary and common-sense linguistic reasoning.

to human problem-solving. Linguistic problem-solving, especially in areas such as syntax, morphology, and phonology, offers a unique and rigorous benchmark for evaluating this capability. These tasks often involve complex linguistic phenomena, typologically diverse languages, and minimal data, requiring models to generalize effectively without relying on prior knowledge or memorized patterns.

In this work, we introduce IOLBENCH, a novel dataset derived from the International Linguistics Olympiad (IOL), a competition designed to test advanced linguistic reasoning. Spanning over two decades of IOL problems, IOLBENCH covers a wide range of linguistic phenomena, including phonological rule inference, morphological paradigm discovery, syntactic structure analysis, and semantic reasoning. These tasks are intentionally crafted to challenge models to hypothesize, test, and generalize abstract linguistic rules from limited data. By leveraging under-documented and typologically diverse languages (eg: Rotuman in Figure 1), IOLBENCH minimizes reliance on pre-trained linguistic biases, providing a stringent evaluation of reasoning capabilities.

To evaluate state-of-the-art performance, we benchmark several leading LLMs, including OpenAI's GPT-4 model family, Anthropic's Claude models, and Google's Gemini model, focusing on two key research questions: (1) To what extent can LLMs handle complex linguistic reasoning tasks requiring abstraction and generalization? (2) What specific strengths and weaknesses do these models exhibit in tackling linguistic problem-solving challenges? Our experiments analyze model performance across task types, identifying gaps in areas such as phonology and morphology, where systematic reasoning and hierarchical pattern recognition are critical. We further show performance on the text-only and multimodal splits of IOLBENCH, demonstrating models are far less performant on visuo-linguistic problems.

By introducing IOLBENCH and conducting a detailed evaluation of LLM performance, this work bridges the fields of computational linguistics and artificial intelligence, providing a foundation for developing models that better emulate the structured reasoning processes underlying human linguistic problem-solving.

2 Related Work

There have been several recent efforts to evaluate the reasoning capabilities of LLMs through the development of domain-specific benchmarks, including MathBench (Liu et al., 2024) for mathematical problem-solving, SciBench (Wang et al.) for scientific reasoning, and datasets based on competitive programming tasks that test logical and algorithmic thinking (Veličković et al., 2022; Shi et al., 2024). These benchmarks provide valuable insights into the ability of LLMs to generalize abstract patterns and solve problems in structured domains. However, the domain of linguistic reasoning remains underexplored, despite its centrality to the applications and theoretical foundations of LLMs.

Language presents a uniquely challenging testbed for reasoning due to its compositional, hierarchical, and rule-governed nature. While models like GPT-4 have demonstrated proficiency in tasks requiring common-sense reasoning, textual coherence, and natural language generation, their capacity to navigate the intricacies of linguistic structure—particularly in typologically diverse or low-resource languages—remains insufficiently understood (Dziri et al., 2024). This gap is particularly pronounced in areas such as morphology (e.g., rule-based inflectional systems) and syntax (e.g., hierarchical phrase structure), where solutions require reasoning over structured data rather than retrieving patterns from pre-trained distributions.

Existing benchmarks, such as BIG-Bench (Suzgun et al., 2023), include a handful of linguistics-inspired tasks but lack the depth and diversity required to evaluate models on truly complex linguistic reasoning. For instance, these tasks often focus on well-documented languages or simplified scenarios, which fail to capture the linguistic diversity and data sparsity challenges posed by real-world language problems. Other datasets, such as NLP-Bench (Song et al.), focus on syntactic parsing or semantic role labeling but do not test a model's ability to deduce and generalize linguistic rules from minimal examples—a hallmark of human linguistic reasoning.

The International Linguistics Olympiad (IOL) provides an ideal framework for addressing this gap. As one of the foremost competitions in linguistics, the IOL challenges participants to solve problems in a wide range of typologically diverse languages, often focusing on under-documented or low-resource languages. These problems require

contestants to deduce complex grammatical rules, analyze morphophonemic patterns, and identify syntactic or semantic structures based on minimal data. The tasks are explicitly designed to test meta-linguistic reasoning and require abstract problem-solving skills rather than prior knowledge of the languages involved. As such, the IOL represents a uniquely challenging and unbiased benchmark for evaluating LLMs on tasks that align closely with human linguistic reasoning.

This work seeks to bridge the gap by introducing IOLBENCH, a dataset derived from over 20 years of IOL problems. Unlike existing linguistics datasets, IOLBENCH focuses on reasoning over typologically diverse languages, such as the endangered Aymara language. It also tests models' ability to generalize rules across unseen data and to engage with tasks that combine multiple levels of linguistic analysis. By leveraging the rich, diverse, and challenging linguistic tasks provided by the IOL, we aim to address this gap and advance our understanding of LLMs' ability to emulate the abstract reasoning processes characteristic of human linguistic cognition.

3 Methodology

IOLBENCH is a high-quality curated dataset of linguistic reasoning problems drawn from the International Linguistics Olympiad (IOL), a global competition that has convened annually since 2003. The IOL is distinguished by its focus on problem-solving abilities that are grounded in linguistic pattern recognition and hypothesis formation, rather than on encyclopedic knowledge of particular languages. Each problem requires participants to deduce underlying grammatical principles—pertaining to phonology, morphology, syntax, semantics, or orthography—from a minimal set of annotated examples. The reasoning process involves extracting abstract generalizations and applying inferred rules to novel test items.

The tasks are purposefully designed to be language-agnostic and frequently focus on low-resource languages where participants cannot rely on preexisting lexical or grammatical familiarity. Problems thus serve as a stringent evaluation framework for linguistic inference, and shallow statistical associations or memorized world knowledge are insufficient for successful performance. Instead, effective problem-solving requires modeling complex morphosyntactic systems, deducing

phonological processes, and unraveling semantic and pragmatic relations with no pretraining bias toward these particular languages families.

Across this dataset, one observes a diverse typology of linguistic phenomena. For instance, phonological problems might involve identifying underlying phonemes, morphological tasks frequently center on discerning affixation patterns (e.g. verbal agreement paradigms), syntactic challenges may focus on constituent structure, semantic and lexicographical subtasks often demand recognizing compositional meanings within the provided language samples. All IOL problems are designed to allow inference of the solution from first principles, encouraging a reasoning-oriented approach to language analysis.

3.1 Dataset Construction

To construct IOLBENCH, we conducted a comprehensive review of the IOL archive, encompassing all main contests and supplementary sample materials from 2003 to 2024. This yielded 25 distinct sets of problems, each containing six core problems with multiple subparts, ultimately resulting in a total of approximately 1,500 problem instances. The digitization process involved transcription from PDFs into machine-readable text, standardized formatting of example sets, and normalization of non-Latin scripts via transliteration tables. Data consistency checks ensured that all problems retained their original logical structure and that associated materials—such as morphological paradigms, orthographic charts, and glossaries—were preserved. When problems included visual or tabular components, these were reformatted as structured textual representations, ensuring that the dataset remains fully accessible to text-based computational models.

Each problem instance in IOLBENCH is paired with its official solution, which was originally authored by expert linguists. These expert solutions are integral to the dataset's utility: they not only verify the correctness of inferred linguistic patterns, but also outline a sequence of deductive steps and intermediate hypotheses. This alignment of problems and solutions thus supports fine-grained evaluation of a model's reasoning process, enabling analyses of whether models can recapitulate the reasoning chains that human solvers employ.

We partition IOLBENCH into the text-only problems (1198) which are expressed entirely in text (or text-convertible forms such as simply struc-

Model	Type 1	Type 2		Type 3
	Acc.	LLM	Bleu	Manual
Claude 3.5s	**38.11**	**59.91**	**22.86**	25.21
Claude 3h	19.84	43.01	17.18	26.17
Claude 3o	36.62	48.73	10.91	28.88
GPT-4	24.48	37.70	10.91	20.25
GPT-4o	26.87	39.25	7.30	27.29
GPT-4om	20.62	31.17	6.20	19.83
GPT-4o1	28.46	46.88	10.41	22.08
Gemini 1.5p	19.76	40.47	7.75	**29.79**

Table 1: Performances (in %) of different LLMs for each problem category for text-problems in IOLBENCH. The best results for each category are bolded.

Model	Type 1	Type 2
Claude 3.5s	16.67	**30.00**
Claude 3o	16.67	10.00
GPT-4	16.67	0.00
GPT-4o	**33.33**	20.00
GPT-4om	16.67	10.00
Gemini 1.5p	**33.33**	20.00

Table 2: Performances (in %) of different LLMs for each problem category for multimodal-problems in IOLBENCH. The best results for each category are bolded.

tured tables), and the multimodal split (52 problems), which requires consuming or generating visual information to solve the problem. For the text split, we evaluate all the listed models, whereas for the multimodal split we evaluate only the models which support visuals inputs/outputs.

3.2 Models Evaluated

We employed IOLBENCH to benchmark state-of-the-art LLM. The goal was to systematically investigate their capacity for abstract linguistic reasoning, as opposed to tasks dependent on rote memorization or external knowledge retrieval. The models evaluated are: OpenAI Models (GPT-4-O1, GPT-4, GPT-4o, GPT-4o-mini(m)), Anthropic Models (Claude 3.5 sonnet(s), haiku (h), opus (o)) and Gemini 1.5p.

3.3 Evaluation Metrics

Each problem in IOLBENCH is categorized into one of three evaluation types, each with corresponding scoring metrics.

1. **Type 1:** 666 problems. For tasks requiring the production of a single word or a short phrase as the solution, correctness is assessed via string matching against the gold-standard answer, with relaxations for minor spelling differences.

2. **Type 2:** 501 problems. For tasks involving longer textual outputs such as translations, evaluation employs both a lexical similarity metric (BLEU) and a LLM-based scoring mechanism. The LLM-based judge assigns a three-tiered score:
   - **0 points:** The response is entirely incorrect compared to the provided solution.
   - **1 point:** The response partially aligns with the reference solution but exhibits notable omissions or errors.
   - **2 points:** The response fully matches the expected solution in both semantic and structural terms.

3. **Type 3:** 31 problems. For more complex tasks requiring explanatory reasoning, we manually grade the solution using an expert human evaluator and the above tiered system.

This multi-faceted scoring framework allows for granular assessment of not only the correctness and completeness of final answers, but also the quality of the reasoning process that leads to those answers, thereby enabling a thorough evaluation of the model's linguistic inference capabilities.

3.4 Results

Table 1 summarizes model performance on IOLBENCH by linguistic domain. Claude 3.5 Sonnet significantly outperforms other models on the text split of IOLBENCH for Categories 1 and 2, while Gemini-1.5 performs the best for Category 3. Table 2 shows that the multimodal problems are significantly more challenging, especially those of Type 2 where several models get zero accuracy.

4 Conclusion

Our benchmarking of LLMs on IOLBENCH, derived from the International Linguistics Olympiad, highlights both their strengths and limitations in linguistic reasoning. While GPT-4 outperformed other models overall, challenges persist in tasks involving morphology and phonology, where abstract rule induction and generalization are critical. These findings underscore the need for improved datasets that capture greater linguistic diversity and for advanced prompting strategies to enhance reasoning capabilities. Future work will focus on expanding IOLBENCH and exploring tailored training approaches to address these gaps in linguistic problem-solving.

5 Limitations

In future work, we plan to look at a larger number of models, including small language models and open language models. We also plan to explore fine-tuning of LLMs to improve their performance on IOLBENCH. This work is meant to be a comprehensive novel exploration in benchmarking LLMs on problems from the International Linguistics Olympiad.

References

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. 2024. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36.

Hongwei Liu, Zilong Zheng, Yuxuan Qiao, Haodong Duan, Zhiwei Fei, Fengzhe Zhou, Wenwei Zhang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*.

Quan Shi, Michael Tang, Karthik Narasimhan, and Shunyu Yao. 2024. Can language models solve olympiad programming? *arXiv preprint arXiv:2404.10952*.

Linxin Song, Jieyu Zhang, Lechao Cheng, Pengyuan Zhou, Tianyi Zhou, and Irene Li. Nlpbench: Evaluating large language models on solving nlp problems. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051.

Petar Veličković, Adrià Puigdomènech Badia, David Budden, Razvan Pascanu, Andrea Banino, Misha Dashevskiy, Raia Hadsell, and Charles Blundell. 2022. The clrs algorithmic reasoning benchmark. In *International Conference on Machine Learning*, pages 22084–22102. PMLR.

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. In *Forty-first International Conference on Machine Learning*.