IOLBENCH: Benchmarking LLMs on Linguistic Reasoning

Anonymous ACL submission

Abstract

Despite the remarkable advancements of deep neural networks, their ability to perform reasoning tasks remains limited, particularly in domains requiring structured, abstract thought. This paper addresses a critical gap in current evaluation: the conflation of recalled knowledge with genuine algorithmic reasoning. We introduce IOLBENCH, a benchmark derived from International Linguistics Olympiad (IOL) problems, which uses carefully designed, self-contained linguistic puzzles to isolate core reasoning abilities. By removing the need for external knowledge, these tasks directly evaluate a model's capacity for explicit rule induction and abstract system modeling from sparse data.

Through extensive benchmarking of leading LLMs, we find that even the most advanced models struggle significantly with these tasks, revealing critical weaknesses in compositional generalization and rule abstraction. Our analysis shows these limitations are not random but indicate a fundamental difficulty in abstracting and systematically applying inferred principles. By providing a clear signal on these foundational skills, IOLBENCH serves as a crucial tool for measuring progress towards more robust AI systems, with direct implications for building agents that can reason and operate effectively in novel environments.

1 Introduction

The emergence of Large Language Models (LLMs) has spurred intense interest in their capacity for reasoning beyond statistical pattern matching. A critical frontier in this research is evaluating a model's ability for genuine algorithmic reasoning, yet a fundamental challenge persists: disentangling this skill from the vast knowledge recalled from training data. Many existing benchmarks, while valuable, inadvertently conflate these two abilities. To truly assess reasoning, we need tasks that neutralize prior knowledge and force models to deduce underlying principles from first principles.

In this work, we introduce IOLBENCH, a benchmark designed to isolate and evaluate these core reasoning faculties. Derived from the International Linguistics Olympiad (IOL), IOLBENCH leverages self-contained, knowledge-independent linguistic puzzles that act as a "clean room" for reasoning evaluation. Specifically, its design allows us to probe three foundational dimensions of reasoning often entangled in other benchmarks: **explicit rule induction** from sparse data, **abstract system modeling** of complex interacting parts, and generalization under **naturalistic complexity**. By using under-documented and typologically diverse languages (e.g., Rotuman in Figure 1), IOLBENCH provides a stringent and focused evaluation of a model's deductive machinery.

To evaluate state-of-the-art performance, we benchmark a suite of leading LLMs, including the GPT-4 family, the Claude 4 family, Gemini 2.5 family, and Grok 4. Our evaluation is guided by two key research questions aligned with our claims: (1) How proficient are leading LLMs at explicit rule induction and abstract system modeling when prior knowledge is neutralized? (2) What systematic failure modes emerge when models are confronted with the naturalistic complexity of diverse linguistic systems? Our experiments analyze performance across task types, including the text-only and multimodal splits, to identify the precise breaking points in model reasoning.

By introducing IOLBENCH, this work offers a targeted diagnostic tool for the AI community. We provide not just a measure of performance, but a clearer signal on the foundational reasoning capabilities and limitations of current models, establishing a crucial foundation for developing robust and generalizable AI systems.

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 → -VC
-VC***u*** → -VC
-VC***i*** → -V̌C
-VC***a*** → -V***a***C
(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

- $\boxed{\mathsf{N_1\ N_2}}$ ($\mathsf{N_1}$: modified, $\mathsf{N_2}$: modifier);
- $\boxed{\mathsf{N\ A}}$ (also with the meaning 'one who has A N': ***huag 'el'ele*** 'heart + shallow = impatient');
- $\boxed{\mathsf{V\ O}}$ (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.

2 Related Work

Recent efforts to evaluate LLM reasoning have produced a valuable suite of benchmarks. Broad, general-purpose evaluations such as MMLU (Hendrycks et al., 2021) and BIG-Bench (Srivastava et al., 2023; Suzgun et al., 2023) test a wide range of topics from science to social studies. While critical for assessing the breadth of a model's capabilities, these benchmarks inherently struggle with the problem of **knowledge contamination**. Because their tasks are grounded in real-world concepts potentially seen during pre-training, it is difficult to disentangle genuine, first-principles reasoning from the sophisticated retrieval of learned information. Consequently, they do not provide a clean signal on a model's ability to reason in a true zero-knowledge scenario, a core requirement for operating in novel domains.

A second category of benchmarks isolates reasoning by focusing on formal, symbolic systems. Datasets for mathematical problem-solving like GSM8K (Cobbe et al., 2021), scientific reasoning like SciBench (Wang et al.), and algorithmic thinking like CLRS (Veličković et al., 2022) evaluate a model's ability to execute structured, multi-step logical procedures. These are invaluable for measuring procedural and logical fidelity. However, their primary focus is on rule *application* within a known, formal system (e.g., the axioms of mathematics or the syntax of a programming language). They do not test the antecedent and arguably more fundamental cognitive skill of **explicit rule induction**—the ability to *discover* the rules of an unknown system from sparse, ambiguous examples. Benchmarks like BenchCLAMP (Roy et al., 2023) similarly focus on syntactic and semantic parsing with constrained decoding, evaluating rule application rather than inductive discovery of linguistic patterns.

The benchmarks most closely aligned with rule induction are those testing compositional generalization, such as COGS (Kim and Linzen, 2020). These datasets effectively test a model's ability to learn a grammar from examples and generalize it to novel combinations. However, they typically rely on synthetic grammars with clean, limited rule sets. This synthetic simplicity contrasts sharply with the layered and often exception-riddled nature of real-world languages.

2.1 Linguistics Olympiad Benchmarks

Several recent works have leveraged Linguistics Olympiad problems to evaluate LLM reasoning. PuzzLing (**?**) and modeLing (Chi et al., 2024) represent early efforts in this direction, with modeLing creating 48 new Rosetta Stone problems (Bozhanov and Derzhanski, 2013; Derzhanski and Payne, 2010) across 19 languages to avoid data leakage. Most directly comparable to our work are Linguini (Sánchez et al., 2024), LingOly (Bean et al., 2024), and the recent work by Majmudar and Filatova (2025). Linguini presents 894 ques-

tions across 160 problems from IOL, achieving 24% accuracy with the best models. LingOly contains 1,133 problems from the UK Linguistics Olympiad, and Majmudar and Filatova (2025) extend this benchmark to include puzzle generation tasks, demonstrating that LLMs can outperform humans on certain puzzle types.

IOLBENCH distinguishes itself through several key contributions: (1) it provides comprehensive coverage of International Linguistics Olympiad problems with 1,250 instances spanning two decades (2004-2023), (2) it includes both text-only and multimodal splits to evaluate visual reasoning capabilities, (3) it implements a rigorous three-tier evaluation methodology with expert manual grading for complex explanatory problems, and (4) it introduces a framework for systematic failure analysis across three reasoning dimensions—explicit rule induction, abstract system modeling, and generalization under naturalistic complexity. This dimensional analysis enables more targeted diagnosis of model limitations beyond aggregate accuracy metrics. Additionally, our comparative analysis of reasoning versus non-reasoning models provides insights into how extended reasoning mechanisms impact linguistic problem-solving performance.

3 The IOLBENCH Dataset

IOLBENCH is curated from the International Linguistics Olympiad (IOL), a global competition designed to test abstract problem-solving skills rather than prior linguistic knowledge. The design of IOL problems aligns directly with the core reasoning capabilities this paper seeks to evaluate. First, by focusing on low-resource and typologically diverse languages, the problems are inherently self-contained, thus neutralizing the confound of knowledge contamination and ensuring a clean evaluation of first-principles reasoning. Second, each problem requires participants to deduce underlying grammatical principles from a minimal set of examples, directly testing explicit rule induction. Finally, these principles rarely exist in isolation; they form complex, interacting systems with exceptions and hierarchies, providing a rigorous test of generalization under naturalistic complexity.

3.1 Dataset Construction and Reproducibility

To construct IOLBENCH, we reviewed the complete IOL archive from 2004 to 2023, digitizing approximately 1,500 problem instances from PDFs into a standardized machine-readable format. This meticulous process, which includes transcription, script normalization, and preservation of all associated materials, is crucial for ensuring the benchmark's reproducibility and is detailed further in the GitHub repository. When problems included visual components, these were preserved, creating a multimodal split. Each problem is paired with its official expert solution, enabling fine-grained evaluation of a model's reasoning process. The zero-shot prompt templates used in our experiments are also available in the Appendix.

3.2 Dataset Preparation Methodology

The initial step involved programmatically scraping problem sets and their corresponding solution guides in English from the official International Linguistics Olympiad (IOL) archives. To ensure high-fidelity text and table extraction at scale, we processed the PDF files using the 'pdfplumber' and 'PyMuPDF' Python libraries. This automated process was followed by a manual verification phase to correct any transcription errors and ensure the integrity of the extracted content. For problems containing essential visual elements, image snippets were extracted directly from the source PDFs. These images were designated to be passed to multimodal models via their respective APIs. In the textual representation of these problems, a placeholder note, '[Image attached]', was inserted to indicate the presence of an accompanying visual.

We partition IOLBENCH into 1198 text-only problems and 52 multimodal problems that require consuming or generating visual information.

3.3 Evaluation Metrics

Each problem is categorized into one of three types for evaluation:

1. **Type 1 (Short Answer, 666 problems):** Requires a single word or phrase, evaluated via string matching against the gold-standard answer.

2. **Type 2 (Long Answer, 501 problems):** Involves longer textual outputs (e.g., translations), evaluated using BLEU and an LLM-based judge on a 0-2 scale.

3. **Type 3 (Explanatory Reasoning, 31 problems):** Requires complex explanatory reasoning, graded manually by an expert evaluator on the same 0-2 scale. To ensure scoring

consistency, a subset of these problems was graded by a second expert, achieving a high inter-annotator agreement (Cohen's Kappa $\kappa = 0.758$).

4 A Framework for Evaluating Abstract Reasoning

Beyond model accuracy scores, we analyze model failures through the lens of our three core reasoning dimensions. Below, we define each dimension, establish how we derived it, and provide a canonical example from IOLBENCH that illustrates the required skill.

4.1 Explicit Rule Induction from Sparse Data

This foundational skill is the ability to observe a pattern in minimal examples and form a hypothesis for a local rule. The capacity for inductive reasoning—generalizing from specific instances to broader principles—is a necessity of both human and artificial general intelligence (Chater et al., 2006). Our analysis will examine cases where models fail at this initial stage of discovery.

The Persian problem (IOL 2007, Problem 1) serves as a clear illustration. The solver is given eight example phrases of the language to its translation and must deduce the semantic difference between two words for "under," *zir-e* and *paːin-e*.

> **Data Snippet:**
> 1. *gorbe zir-e raxtexaːb* — The cat under the bed
> 2. *xane paːin-e kuh* — The house under the mountain
> ...
> **Question:** Translate *tʃahaːrpaːye paːin-e miz* (The stool [?] the table)

Success on this problem is in part based on observing the data and hypothesizing a grammatical rule: *zir-e* denotes direct physical contact, while *paːin-e* denotes a general lower elevation. This act of hypothesizing one core principle from a few contrasting examples is explicit rule induction.

4.2 Abstract System Modeling

This skill involves synthesizing multiple, distinct rules into a coherent, interacting system. The ability to form such internal "mental models" has long been considered a central component of complex cognition and problem-solving in humans (Johnson-Laird, 1983). A failure here indicates a model can identify individual patterns but cannot grasp how they fit together.

The Ancient Greek problem (IOL 2007, Problem 2) is a perfect test of this skill. A solver must match phrases to translations by building a mini-grammar for noun phrases.

English	Tupinambá	Guaraní Mbya
rock	itá	itá
head	akanga	akã
to bring	erur	eru
you want	erepotár	erepotá
to heal	pueráb	kuerá

Table 1: Example data from the Tupi-Guarani problem.

> **Data Snippet:**
> A. *ho toːn hyioːn dulos*
> B. *hoi toːn duloːn cyrioi*
> ...
> **Translations (Unordered):**
> 1. the slave of the sons
> 2. the masters of the slaves
> ...

Solving this requires building a multi-part system model by correctly inducing and integrating several interlocking rules: (1) a rule for number distinguishing *ho* (sg.) from *hoi* (pl.); (2) a rule for possession distinguishing *tu* (of the sg.) from *toːn* (of the pl.); and (3) a rule for word order. A solver cannot succeed by finding one trick; they must model how the number, possession, and word-order subsystems interact to form a coherent whole.

4.3 Generalization under Naturalistic Complexity

This dimension tests the ability of an inferred system to handle the messy, irregular, and layered nature of real-world phenomena. The algebraic capacity to compose known rules is a key feature of general intelligence (Fodor and Pylyshyn, 1988), however, real-world systems often contain exceptions and interactions not present in clean, formal systems. This skill evaluates the robustness of a model's reasoning when moving beyond idealized scenarios.

The Tupi-Guarani problem (IOL 2011, Problem 3) exemplifies this challenge. It requires inducing a series of sound changes between the related Tupinambá and Guarani Mbya languages.

The task involves filling in blank words by applying the inferred rules. This requires not just building a system model of sound changes, but generalizing this model to new words that introduce the complexity of real language evolution. The sound changes are not perfectly regular; they interact with syllable structure and vowel quality. This forces the solver to apply their inferred system to the noisy, irregular data characteristic of a natural, evolved system.

Model	Type 1	Type 2		Type 3
	Acc.	LLM	Bleu	Manual
GPT-5	**61.12**	41.91	20.76	20.21
Grok-4	38.11	32.91	**22.86**	25.21
Claude 4 sonnet	36.84	46.07	27.18	16.17
Claude 4 opus	41.90	46.45	20.91	28.88
GPT-4	24.48	37.70	10.81	20.25
GPT-4o	26.87	39.25	7.30	27.29
GPT-4o mini	20.62	31.17	6.20	19.83
Gemini 2.5 pro	39.76	**50.47**	37.05	**29.79**
Gemini 2.5 flash	19.76	35.47	17.75	19.79

Table 2: 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 4 Sonnet	33.10	24.20
Grok-4	27.10	12.22
Claude 4 Opus	44.80	**43.10**
GPT-4o	10.00	17.50
Gemini 2.5 Pro	**45.20**	39.10

Table 3: Performances (in %) of different LLMs for each problem category for multimodal-problems in IOLBENCH.

Problem Snippet (Lalana Chinantec):

1. kalakwa: kwɨ: li:ʔ – The beautiful corn grew.
2. mɨladʒö mo:h kya – My pineapples have turned out well.
3. li:ʔ kalane kwɨ: kwa: kya – My tall corn yellowed beautifully.

...

Question: What does the word li:ʔ mean?

Model Response (Gemini 2.5 Pro (thinking)):

...This suggests li:ʔ can mean both "grew" and "yellowed" - likely it means "yellowed" and in sentence 1 it's part of a compound meaning.
Final Analysis: Part a: li:ʔ means "yellowed"

Figure 2: A failure of explicit rule induction. The model gets confused by the polysemy in the English translation and fails to deduce the correct meaning of li:ʔ from the syntactic context.

5 Experiments and Results

5.1 Experimental Setup

Models Evaluated. We evaluate a suite of state-of-the-art LLMs to assess the current frontier of linguistic reasoning. The models include: OpenAI's GPT-4 family (GPT-4, GPT-4o, GPT-4o mini), OpenAI's o1 model, Anthropic's Claude 4 family (Sonnet and Opus), xAI's Grok-4, and Google's Gemini 2.5 family (Pro and Flash).

Prompting Strategy. All experiments are conducted using a zero-shot prompting strategy. Each problem from IOLBENCH is presented to the model with instructions to solve it based solely on the provided information. This approach is critical for enforcing the knowledge-independent nature of our evaluation. The complete prompt templates used for each problem type are available in Appendix.

5.2 Quantitative Analysis

The overall performance of the evaluated models on the text-only and multimodal splits of IOLBENCH is summarized in Table 2 and Table 3, respectively. These results provide a high-level overview of current capabilities and highlight the significant challenge our benchmark presents.

As the tables indicate, overall performance remains low across all models, confirming that the abstract reasoning required by IOLBENCH is a significant challenge for current architectures. Gemini 2.5 Pro demonstrates the strongest performance on text-based tasks, while Claude 4 Opus excels on the more difficult multimodal problems. However, to understand the nature of these limitations, we now turn to a qualitative analysis of model failures.

5.3 Qualitative Error Analysis: A Dimensional Breakdown

Quantitative scores reveal that models struggle; a qualitative analysis is necessary to understand the dimensionality of failures. We now analyze specific model failures through the lens of our three reasoning dimensions, using representative examples to diagnose the breaking points in the models' reasoning processes.

5.3.1 Failure in Explicit Rule Induction

The most fundamental errors occur when models fail to identify simple, local patterns, demonstrating an inability to perform the discovery step of reasoning. In these cases, models often hallucinate non-existent rules or overlook clear evidence in the data. The Lalana Chinantec problem (IOL 2008, Problem 5) provides an example of this failure. The task requires decomposing sentences to find the meaning of individual morphemes.

The model's reasoning in Figure 2 demonstrates a failure of rule induction. By comparing sentences 1 and 3, a human solver can deduce that verbs can appear before or after the noun phrase. In sentence 1, li:ʔ is the verb 'grew', while in sentence 3, the verb is 'kwa:' ('yellowed'). The model, however, incorrectly assumes a fixed word order and

Problem Snippet (Guaraní Verbs):
`japurahei – We are singing`
`noñe'i – He is not talking`
`ndajajupirima – We were not waking up`
...
Question: Translate "We will be eating" into Guaraní.

Model Response (Claude 4 Opus (Thinking)):
`...Morphological Analysis: Looking at negation patterns... The future appears to use mo'ã as an auxiliary or suffix.`
`Part b - Guaraní translations: ... "We will be eating" = jakarumo'ãta (ja-karu-mo'ã-ta)`

Figure 3: A failure of abstract system modeling. The model correctly identifies individual morphemes but fails to synthesize the correct interaction for future tense, incorrectly combining a future auxiliary with a future suffix.

gets confused, ultimately misidentifying `li:ʔ` as 'yellowed'. It failed to ground its reasoning in the syntactic evidence provided, instead using a plausible but incorrect semantic association.

5.3.2 Failure in Abstract System Modeling

A more complex failure occurs when a model correctly identifies some individual rules but cannot synthesize them into a coherent, interacting system. The Guaraní Verbs problem (IOL 2004, Problem 16) highlights this weakness. Solving it requires building a complete model of the verb paradigm from multiple, interlocking morphemes.

As shown in Figure 3, the model successfully performs rule induction on the person prefixes (*ja-* for "we") and the verb stem (*karu* for "eat"). However, it fails to build a correct system model for tense. It observes both a future suffix *-ta* and a future auxiliary *mo'ã* (used in negative constructions) and incorrectly concludes that they should be combined. The correct answer, *jakaruta*, requires understanding that these are part of separate subsystems (affirmative vs. negative future tense). The model identified the parts but could not assemble them correctly, demonstrating a critical failure in synthesizing rules into an accurate system.

5.3.3 Failure of Generalization under Naturalistic Complexity

The most advanced reasoning failures occur when a model's inferred system, while perhaps internally consistent, doesn't generalize well to complex linguistic structure. The Icelandic Kinship problem (IOL 2005, Problem 14) demonstrates this, requiring a model to apply the rules of Icelandic naming to a multi-generational family tree.

Problem Snippet (Icelandic Kinship):
`...Guðrún Eriksdóttir and Jakob Christiansson had three kids... Here is the list of all the people in the party: Christian Kristínsson, Daniel Guðrunarson Jakobsson, ... Kristín Guðrunardóttir Aronsdóttir...`
Question: Which of the sons/daughters of Guðrún and Jakob has no grandsons/granddaughters?

Model Response (Grok-4 (thinking)):
`"2": "Kristín Guðrunardóttir Aronsdóttir"`

Figure 4: A failure to generalize to naturalistic complexity. The model incorrectly identifies Kristín as a child of Guðrún, failing to parse the multi-part surnames.

In Figure 4, the model fails a complex inferential task. To answer correctly, one must first identify all of Guðrún and Jakob's children from the list, then identify all of their children (the grandchildren), and finally see which of the original children has no descendants in the third generation. The correct answer is Ragnheiður Jakobsdóttir. The model incorrectly identifies Kristín Guðrunardóttir Aronsdóttir. This error reveals that the model's internal system for Icelandic names was not generalizable. It likely identified "Guðrunardóttir" (daughter of Guðrún) however Kristín also has a second matronymic, "Aronsdóttir," indicating her mother is someone named Aron, not Guðrún. The model's derived rules fail to generalize under naturalistic complexity.

5.4 Case Study: Reasoning vs. Non-Reasoning Models

To understand the impact of explicit reasoning capabilities, we conducted a focused comparison between GPT-4o, a standard non-reasoning model, and Gemini 2.5 Pro, which incorporates extended reasoning mechanisms. This analysis examines how systematic, step-by-step reasoning affects performance on tasks requiring explicit rule induction and abstract system modeling.

5.4.1 Ekari Numerals - Base System Discovery

This problem is problem 1 in the IOL 2021 contest.

The Ekari numeral problem requires discovering a complex vigesimal (base-20) and sexagesimal (base-60) system with subtractive constructions. When asked to translate `moeto waroewo` into digits (correct answer: 480):

GPT-4o provided a direct answer of `127` with minimal justification, failing to identify the multiplicative structure of `moeto` (60) times `waroewo`

(3).

Gemini 2.5 Pro demonstrated extensive exploration: "I've been analyzing Ekari numerals...I now hypothesize that the system is based on 20, with addition and subtraction used to construct other numbers...moeto appears to be 60...I'm correlating their values to establish the underlying structure." Despite this systematic approach yielding 180 (still incorrect), the model correctly identified moeto as 60 and attempted multiplicative composition, showing partial success in explicit rule induction. The error stemmed from misidentifying waroewo as 3 rather than 8.

This comparison reveals that reasoning models make progress on the foundational skill of pattern discovery but still struggle with the sparse data problem when multiple hypotheses remain viable.

5.4.2 Kilivila Grammar - Word Order and Agreement

This problem is problem 3 in the IOL 2021 contest.

For sentence 14 (Navila vivila biyamata tomwaya mtona?), the correct translation is "How many women will this old man look after?"

GPT-4o produced: "How many women looked after this child?", misidentifying both the tense marker (*bi-* indicates future, not past) and the noun *tomwaya* (old man, not child). These errors suggest failure at the most basic level of rule induction.

Gemini 2.5 Pro showed grammatical analysis: "I've identified the use of prefixes like bi- for future tense and le- for past tense, and I'm beginning to understand the role of suffixes, particularly -si, in plural subject agreement." The model correctly identified the VOS word order and tense system, producing accurate translations across multiple sentences.

This example illustrates that reasoning capabilities significantly improve performance on problems requiring system-level grammatical modeling, particularly when multiple morphological and syntactic rules must be integrated.

5.4.3 Rikbaktsa Verbs - Morphological Synthesis

This problem is problem 5 in the IOL 2021 contest.

When translating aharapyryk (correct: "I hugged you"), GPT-4o provided "I will hug you", correctly identifying most components but failing on tense. Gemini 2.5 Pro analyzed: "After thoroughly examining word order, verb tenses, and agreement patterns...aharapyryk uses aha- prefix suggesting past reference to second person plural object" and concluded "I hugged you_{pl}", achieving near-perfect accuracy with only a minor singular/plural distinction error.

For the translation "we will fall," GPT-4o produced tsipikharapyryk (incorrect), while Gemini 2.5 Pro generated tsiknanaha (closer to the correct tsimynarahanaha), demonstrating better compositional generalization.

5.4.4 Key Findings

Our comparative analysis reveals three critical insights:

1. Reasoning improves explicit rule induction. Gemini 2.5 Pro consistently formulates and tests explicit hypotheses about linguistic patterns, while GPT-4o often produces answers without evident systematic analysis. However, both models still struggle when data is highly sparse or ambiguous.

2. Extended reasoning substantially benefits abstract system modeling. On problems requiring synthesis of multiple interacting rules (Kilivila grammar), Gemini 2.5 Pro shows marked improvement. The explicit reasoning trace demonstrates the model building coherent internal representations of linguistic systems.

3. Reasoning does not fully overcome naturalistic complexity. Despite systematic analysis, Gemini 2.5 Pro still fails on problems with high exception rates or multi-layered rule interactions, suggesting that current reasoning mechanisms, while beneficial, do not fundamentally solve the generalization problem under naturalistic conditions.

These findings indicate that while extended reasoning capabilities represent progress toward more robust linguistic analysis, significant architectural gaps remain in achieving human-level inductive reasoning from sparse, complex data.

6 Discussion and Conclusion

In this work, we introduced IOLBENCH, a benchmark designed not just to measure performance on linguistic puzzles, but to isolate and diagnose fundamental reasoning capabilities in Large Language Models. Our experiments reveal that even state-of-the-art models exhibit significant weaknesses when confronted with tasks requiring gen-

uine, knowledge-independent reasoning. The quantitative results show low aggregate scores, but our qualitative analysis provides the critical insight on three key dimensions of reasoning where LLMs fail. We find that models struggle with the foundational skill of explicit rule induction, often fail to synthesize discovered rules into a coherent abstract system model, and struggle to form generalizable reasoning strategies against naturalistic complexity.

These findings have profound implications for the development of advanced AI. The goal of creating autonomous agents that can reason, plan, and act in novel environments is a central theme in current research (Bubeck et al., 2023). Such an agent's success is predicated on its ability to do precisely what IOLBENCH tests: deduce the "rules of the world" from sparse observations. Our results suggest that current architectures, despite their vast knowledge, lack the robust deductive machinery to do this reliably. A model that cannot induce the simple rules of Rotuman morphology is unlikely to reliably infer the complex causal rules of a new scientific domain or an unfamiliar interactive environment. The failures we document are therefore not just linguistic, but are indicative of a deeper architectural gap that must be addressed to build more capable and trustworthy AI systems.

Limitations

Our analysis provides a clear diagnostic, but it is not exhaustive. The current work focuses on a representative set of leading proprietary models; future work should expand this evaluation to a wider range of open-source models and varied architectures to broaden our understanding. Furthermore, while we focused on zero-shot evaluation to ensure a clean test of reasoning, exploring the efficacy of few-shot prompting and fine-tuning on a subset of IOLBENCH could reveal important insights into how these reasoning skills might be learned. Additionally, expanding the IOLBENCH dataset across more questions, especially multimodal ones, can yield more insights on model behavior. We release IOLBENCH and its code to the community to facilitate this research, providing a challenging and precise tool to measure our collective progress towards true artificial general intelligence.

Acknowledgments

We thank the International Linguistics Olympiad organization for making problem sets publicly available, and we thank the reviewers for their constructive feedback.

References

Andrew M Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Chi, Scott A Hale, and Hannah Rose Kirk. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages. *Advances in Neural Information Processing Systems*, 37.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8, Sofia, Bulgaria. Association for Computational Linguistics.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Nick Chater, Joshua B Tenenbaum, and Alan Yuille. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, 10(7):287–291.

Nathan A Chi, Teodor Malchev, Riley Kong, Ryan A Chi, Lucas Huang, Ethan A Chi, R Thomas McCoy, and Dragomir Radev. 2024. Modeling: A novel dataset for testing linguistic reasoning in language models. *arXiv preprint arXiv:2406.17038*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Ivan Derzhanski and Thomas Payne. 2010. *The Linguistic Olympiads: academic competitions in linguistics for secondary school students*, page 213–226. Cambridge University Press.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Philip Nicholas Johnson-Laird. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. 6. Harvard University Press.

Najoung Kim and Tal Linzen. 2020. COGS: A compositional generalization challenge based on semantic

interpretation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9087–9105, Online. Association for Computational Linguistics.

Neh Majmudar and Elena Filatova. 2025. Can llms solve and generate linguistic olympiad puzzles? *arXiv preprint arXiv:2509.21820*.

Subhro Roy, Samuel Thomson, Tongfei Chen, Richard Shin, Adam Pauls, Jason Eisner, and Benjamin Van Durme. 2023. Benchclamp: A benchmark for evaluating language models on syntactic and semantic parsing. *Advances in Neural Information Processing Systems*, 36:49814–49829.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R Costa-jussà. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning. *arXiv preprint arXiv:2409.12126*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, and 1 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and 1 others. 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*.

A Prompt Template for Zero-Shot Evaluation

All experiments were conducted using a zero-shot prompting strategy. The models were given a persona and a structured set of instructions to guide their reasoning and formatting. The complete prompt template used for querying the large language models is provided below. The specific problem text was inserted into the $< problem >$ tag, and the required JSON structure was provided in the tag $< answer_format >$ for each problem.

```
As a linguistics expert with deep knowledge
of computational linguistics, historical
linguistics, and language structure analysis,
solve the following problem systematically.
Analyze the linguistic patterns, apply
your knowledge of phonology, morphology,
syntax, and semantics as appropriate. Show
your reasoning clearly and provide accurate
solutions.
At the end, provide your solution in the exact
JSON format specified below. Include only the
answers in the final JSON - no explanations
or additional text after the JSON format.
<answer_format>
{
"Problem X": {
"Part a": {
...
}
}
}
</answer_format>
<problem>
[Problem text is inserted here]
</problem>
```

B Prompt Template for LLM-as-Judge Evaluation

The LLM-as-judge evaluation employed a structured grading rubric with three question types: Type 1 (exact match, 0-1 points), Type 2 (phrase answer, 0-2 points), and Type 3 (answer with explanation, 0-2 points). The judge model was instructed to strictly adhere to pre-assigned classifications and extract precise answers from model responses for comparison against ground truth. The complete prompt template is provided below.

```
Act as an expert linguistics grader. You will
grade ONE problem at a time using the provided
rubric.

RUBRIC:
- Type 1: Exact match required (0 or 1 point)
- Type 2: Phrase answer with 3+ words
(0=completely wrong, 1=partially correct,
2=completely correct)
- Type 3: Answer with explanation required
(0=both wrong, 1=one component correct,
2=both correct)

CRITICAL CLASSIFICATION RULE:
- USE ONLY THE PROVIDED CLASSIFICATIONS. DO
NOT CHANGE QUESTION TYPES.
- If the input says "Type 1", you MUST classify
it as "Type 1" in your output.
- If the input says "Type 2", you MUST classify
it as "Type 2" in your output.
- You are NOT allowed to reclassify questions
based on your judgment.

GRADING INSTRUCTIONS:
1. Extract the specific answer from the LLM's
response for each question
2. Compare against the ground truth answer
3. Use the EXACT classification provided in
the input
4. Grade strictly according to the rubric and
the PROVIDED question type
5. If LLM provides no clear answer, mark as
"No answer provided"

REQUIRED JSON OUTPUT FORMAT:
{
"problem_id": 1,
"parts": {
"a": {
"questions": {
"15": {
"classification": "Type 1",
"llm_answer": "extracted answer text",
"real_answer": "ground truth answer",
"grade": 1,
"max_points": 1
}
}
}
},
"summary": {
"total_score": 1,
"total_possible": 2,
"accuracy": 0.5
}
}

INPUT DATA:
Problem: {problem_content}
LLM Response: {llm_response}
Ground Truth: {ground_truth}
Classifications: {classifications}
```