МАТЕМАТИКА И МАТЕМАТИЧЕСКО ОБРАЗОВАНИЕ, 2026
MATHEMATICS AND EDUCATION IN MATHEMATICS, 2026
Proceedings of the Fifty-Fifth Spring Conference
of the Union of Bulgarian Mathematicians
Tryavna, Bulgaria, April 5–9, 2026

EVALUATING SYSTEMATIC LINGUISTIC REASONING IN LARGE LANGUAGE MODELS VIA LINGUISTICS OLYMPIAD PROBLEMS

Tsvetelina Stefanova, Tsvetozar Georgiev
Department of Computer Systems and Technologies,
Angel Kanchev University of Ruse, Bulgaria
e-mails: tsstefanova@uni-ruse.bg, tgeorgiev@ecs.uni-ruse.bg

Current evaluation benchmarks for large language models often conflate genuine reasoning ability with memorization and statistical pattern matching. As a result, high benchmark scores do not necessarily indicate the capacity to induce abstract rules or to reason systematically from limited data. In this paper, we propose a methodological framework for evaluating linguistic reasoning in large language models based on problems from international and national linguistics olympiads. These problems are designed to be solvable without prior knowledge of the target language, requiring instead the induction of formal linguistic rules from small, self-contained datasets. We formalize linguistics olympiad problems as closed-world reasoning tasks and introduce a structured evaluation protocol that combines answer correctness, reasoning quality, hallucination detection, and generalization analysis.

Keywords: large language models, linguistic reasoning, rule induction, evaluation methodology, linguistics olympiads

ОЦЕНЯВАНЕ НА СИСТЕМАТИЧНОТО ЛИНГВИСТИЧНО РАЗСЪЖДЕНИЕ В ГОЛЕМИ ЕЗИКОВИ МОДЕЛИ ЧРЕЗ ЗАДАЧИ ОТ ЛИНГВИСТИЧНИ ОЛИМПИАДИ

Цветелина Стефанова, Цветозар Георгиев
Катедра „Компютърни системи и технологии“,
Русенски университет „Ангел Кънчев“, България
e-mails: tsstefanova@uni-ruse.bg, tgeorgiev@ecs.uni-ruse.bg

Съществуващите методи за оценяване на големи езикови модели често смесват способността за разсъждение с механично възпроизвеждане на статистически зависимости, което затруднява разграничаването между истинско логическо извеждане и

https://doi.org/10.55630/mem.2026.55.134-149
2020 Mathematics Subject Classification: 68T50, 68T01, 91F20.

запаметяване. В статията се предлага методологична рамка за оценяване на лингвистичното разсъждение в големи езикови модели чрез задачи от международни и национални олимпиади по лингвистика. Тези задачи са конструирани така, че да бъдат решавани без предварителни знания за съответния език и изискват извеждане на абстрактни езикови правила от ограничен набор данни. Оценяването се основава на съчетание от коректност на отговора, качество на аргументацията, наличие на халюцинации и способност за обобщение.

Ключови думи: големи езикови модели, лингвистично разсъждение, извеждане на правила, методология за оценяване, олимпиади по лингвистика

1. Introduction

The rapid evolution of Large Language Models (LLMs) has led to near-human performance on standard Natural Language Processing (NLP) benchmarks. However, a critical question remains: do these models exhibit genuine systematic reasoning, or do they merely rely on approximate retrieval from vast training corpora?

Current evaluation paradigms face a significant challenge known as "epistemic contamination" or data leakage. As noted by Lin et al. [8] in KUMO, publicly released benchmarks are often inadvertently subsumed into training data, rendering them unreliable as tests of novel reasoning. Furthermore, standard benchmarks such as MMLU (Massive Multitask Language Understanding) or GLUE (General Language Understanding Evaluation) conflate linguistic competence (latent knowledge of grammar) with performance (the application of that knowledge), a distinction highlighted by Waldis et al. in the Holmes benchmark [15]. To truly assess an LLM's inductive logic, we require a "closed-world" evaluation environment where success depends solely on in-context learning rather than prior world knowledge.

Linguistics olympiad problems—Rosetta-Stone-style puzzles requiring the decipherment of unseen languages—offer an ideal testbed for this purpose. Unlike translation tasks that benefit from massive pre-training on high-resource languages, linguistics olympiad problems are designed to be solvable without external knowledge, relying instead on meta-linguistic awareness and logical deduction [12].

While recent initiatives have begun to explore this domain, existing methodologies remain fragmented. LingOly [1] and Linguini [13] have established valuable datasets for low-resource languages, demonstrating that LLMs struggle with "language-agnostic" generalization. However, these benchmarks largely rely on exact-match accuracy metrics, which fail to capture the reasoning process. As shown by Bhattacharya et al. [2] in their analysis of counting systems, LLMs often achieve correct answers via pattern matching while failing to induce the underlying compositional operators (e.g., multiplication in base-20 systems), suggesting a disconnect between output correctness and logical faithfulness.

In this paper, we introduce a rigorous, research-grade framework for evaluating systematic linguistic reasoning. Building on the agentic insights of LingBench++ [7] and the reasoning critiques of CriticBench [9], we propose a multi-dimensional evaluation protocol. Our methodology goes beyond surface-level accuracy to evaluate Reasoning

Transparency, Hallucination Rates, and the Generalization Gap between seen and unseen linguistic structures. By formalizing these metrics, we aim to provide a standardized instrument for measuring the transition of LLMs from stochastic parrots to systematic reasoners.

2. Related Work

Current evaluation paradigms for Large Language Models can be broadly categorized into knowledge-intensive benchmarks, general reasoning suites, and specialized linguistic challenges. Our proposed framework addresses the gaps at the intersection of these fields.

2.1 From Knowledge Retrieval to Rule Induction

Standard benchmarks such as MMLU and GLUE primarily evaluate an LLM's ability to retrieve information seen during pre-training. However, distinguishing between memorization (approximate retrieval) and generalization (rule application) remains a fundamental challenge. [12] was among the first to propose Rosetta-Stone-style puzzles as a proxy for low-shot learning, establishing that these tasks require "meta-linguistic awareness" rather than semantic knowledge. More recently, the authors of LingOly [1] expanded this domain by compiling over 90 low-resource languages (e.g., Nivkh, Burushaski). Their findings demonstrate that model performance degrades significantly on "unseen" languages, suggesting that high performance on major languages is often an artifact of data contamination rather than genuine inductive capability. Similarly, Linguini [13] and IOLBENCH [4] have formalized these problems into "language-agnostic" benchmarks, yet they largely retain Exact Match (EM) or Character n-gram F-score (chrF) accuracy as the sole metric for success. While LingOly and Linguini focus on low-resource generalization, other datasets have targeted the structural diversity of these problems. MODELING [3] allows for finer-grained evaluation of specific linguistic reasoning skills, though it focuses on easier, author-constructed problems. Conversely, Majmudar & Filatova [10] explore the generative capabilities of models, demonstrating that while LLMs can solve complex puzzles, they still struggle to generate novel, solvable linguistic problems, highlighting a gap between passive solving and active design.

2.2 The Evaluation Gap: Accuracy vs. Faithfulness

A critical limitation of current benchmarks is the assumption that a correct answer implies correct reasoning. CriticBench [9] reveals that LLMs frequently generate "hallucinated reasoning"—sound-looking but logically invalid explanations that fortuitously arrive at the correct label. This phenomenon is particularly acute in compositional tasks. Bhattacharya et al. [2] demonstrated in their study of counting systems that while LLMs can memorize number sequences (e.g., "uno, dos, tres"), they fail to induce the underlying mathematical operators (e.g., Base-20 multiplication) unless explicitly prompted with symbols. This finding underscores the necessity of our proposed Reasoning Quality metric: an evaluation framework must verify the derivation of the rule, not just the final output.

2.3 Theoretical Grounding: Competence vs. Performance

Our methodology is grounded in the distinction between Linguistic Competence (latent grammatical knowledge) and Performance (the ability to use it), as explored in

Holmes [15]. While Holmes uses probing classifiers to assess latent knowledge, our framework focuses on the application of that knowledge in a few-shot setting. Furthermore, we draw on KUMO [8] to justify our "Generative Evaluation" approach. Lin et al. argue that static benchmarks become contaminated the moment they are released; thus, a robust evaluation protocol must rely on "closed-world" problems where the rules are self-contained within the prompt, mimicking the "Unseen" conditions of valid psychometric testing.

2.4 Agentic and Multi-Step Reasoning

Finally, the transition from prompt-based solvers to agentic architectures represents the state-of-the-art in this domain. LingBench++ [7] introduced a "Check-of-Thought" framework where separate agents generate and verify linguistic hypotheses. While effective, LingBench++ focuses on measuring the performance boost provided by agents. In contrast, our paper proposes using similar multi-step protocols as a measurement instrument—using the divergence between "Direct-Shot" and "Agentic-Verified" performance to quantify a model's Self-Correction Capability, a key indicator of robust reasoning.

3. Methodology

3.1 Problem Selection and Scope

To evaluate systematic linguistic reasoning in LLMs, we choose problems exclusively from official linguistics olympiads, including the International Linguistics Olympiad (IOL), the North American Computational Linguistics Olympiad (NACLO), the UK Linguistics Olympiad (UKLO), and selected national competitions. These sources are chosen because their problems are carefully designed, verified by expert linguists, and accompanied by official solutions. Only problems with publicly available solutions are included, ensuring reproducibility and verifiability.

To prevent model performance from reflecting memorization rather than reasoning, tasks satisfy two strict inclusion criteria:

1. **Unseen / Anti-Contamination:** Target languages must fall below a "digital poverty" threshold, favoring low-resource, extinct, or isolated languages (e.g., Nivkh, Burushaski, Sumerian). This ensures that models cannot rely on pre-training exposure to solve the problems [1, 5].
2. **Closed-World / Self-Containment:** Each problem is solvable solely from the problem statement; any reference to external knowledge constitutes hallucination [9].

Each problem is manually annotated by its dominant linguistic phenomenon. The benchmark covers six primary domains: morphology, phonology, syntax, semantics and pragmatics, writing systems and orthography, and historical/comparative linguistics. Problems may involve multiple interacting phenomena and are labeled accordingly.

To avoid bias toward specific linguistic structures, the dataset is typologically stratified across three cognitive domains [2]:

- **Morphological Induction:** segmentation of agglutinative or polysynthetic words.

- **Syntactic Alignment:** inference of non-English word orders (e.g., VSO, OVS) or ergative-absolutive systems.
- **Compositional Systems:** recursive numerical or kinship systems requiring multi-step arithmetic or rule application.

Problems are further stratified into three difficulty levels—*Introductory*, *Intermediate*, and *Advanced*—based on olympiad round, number of inference steps, and ambiguity. Introductory problems involve single rules and few examples, intermediate problems require combining several rules, and advanced problems demand high-level abstraction and reasoning under incomplete data.

3.2 Dataset Structure

All problems are stored in a standardized schema (Table 1), including:

- **Problem ID:** unique identifier encoding olympiad, year, problem number.
- **Source Metadata:** olympiad name, year, round, original language.
- **Phenomenon Labels:** one or more linguistic categories.
- **Difficulty Level:** Introductory, Intermediate, or Advanced.
- **Problem Statement:** verbatim text including tables, scripts, examples.
- **Input–Output Pairs:** structured representations of linguistic examples.
- **Question Targets:** explicit outputs required from the model (e.g., translations, forms, rules).
- **Gold Solution:** official solution with final answers and reasoning steps.
- **Evaluation Type:** exact-match, partial credit, or rubric-based evaluation.

The dataset is stored in a machine-readable format (JSON) while preserving human-readable statements for qualitative analysis. Problems are grouped hierarchically by domain and difficulty. To minimize contamination, recent problems are reserved for evaluation, and a challenge problem set with rare languages or unusual systems is designed to test generalization.

3.3 Prompting and Inference Protocol

Models are evaluated using a fixed prompting protocol to ensure comparability. Each prompt includes the original problem statement preceded by a standardized instruction:

> "Solve the following linguistics problem. Explain your reasoning step by step and provide the final answer clearly."

No additional examples are provided. Non-ASCII scripts are preserved. Inference parameters (temperature, max output length) are fixed.

To differentiate pattern matching from structured reasoning, two computational conditions are defined [7, 11, 16]:

1. **Baseline / Monolithic Chain-of-Thought:** single-pass "Field Linguist" prompt enforcing a linear chain-of-thought.
2. **Agentic Check-of-Thought Loop:** decomposed into specialized agents (Grammar Induction, Consistency Checker, Solver) for iterative verification.

Table 1: Benchmark Dataset Schema

Field	Description	Type
Problem ID	Unique identifier encoding olympiad, year, and problem number	String
Olympiad Source	IOL, NACLO, UKLO, or national olympiad	Categorical
Year	Year of publication	Integer
Round	Contest round (e.g., preliminary, final)	Categorical
Linguistic Domain	Morphology, phonology, syntax, semantics, orthography, historical	Multi-label categorical
Difficulty Level	Introductory / Intermediate / Advanced	Ordinal
Language(s)	Target language(s) or scripts involved	String / List
Problem Statement	Original problem text (verbatim)	Text
Data Examples	Input–output linguistic examples	Structured text
Question Targets	Required outputs (translations, forms, rules)	Text
Gold Answer	Official final answer	Text
Gold Reasoning	Official solution explanation	Text
Evaluation Type	Exact-match / Partial / Rubric-based	Categorical

3.4 Evaluation Metrics

LLM performance is measured along three dimensions: answer accuracy, reasoning quality, and generalization/hallucination. Additional metrics from literature (A_{surf}, R_{faith}, $SLRS$) are provided for conceptual grounding [9, 13].

3.4.1 Problem-Level Accuracy

Problem-Level Accuracy serves as the primary quantitative metric for assessing model performance, measuring the correspondence between the model's generated answers and the ground-truth solutions provided in the linguistics olympiad datasets. Unlike reasoning metrics, this measure focuses strictly on the correctness of the final output (the "performance" aspect of the competence-performance distinction). Let $P = \{p_1, \ldots, p_N\}$ denote the benchmark problem set. For each problem p_i, the model generates output y_i compared against the gold solution y_i^*.

For problems with a single deterministic answer, correctness is defined through an indicator function:

$$A_i = \mathbb{I}(y_i = y_i^*) \tag{1}$$

where $\mathbb{I}(\cdot)$ equals 1 when the prediction exactly matches the reference and 0 otherwise. This exact-match criterion ensures objective scoring for discrete outputs such as translations or morphological forms.

For multi-part problems, scoring is decomposed to account for structured responses:

$$A_i = \sum_{j=1}^{k_i} w_{ij} \cdot \mathbb{I}(y_{ij} = y_{ij}^*) \tag{2}$$

where k_i is the number of components and w_{ij} is normalized weight. This formulation captures partial correctness without overvaluing trivial sub-items.

Interpretation of accuracy must remain cautious. Correct outputs may arise from superficial pattern matching rather than genuine rule induction, meaning accuracy reflects behavioral agreement rather than validated reasoning. To reduce evaluation noise, normalization procedures are applied before comparison, resolving orthographic variation, formatting differences, or equivalent symbolic representations.

Aggregate benchmark accuracy is computed as:

$$A_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} A_i \tag{3}$$

and serves as a descriptive baseline for performance reporting and stratified analysis across problem types or difficulty levels. Within the broader framework, Problem-Level Accuracy underpins composite scoring and generalization analysis, providing the measurable outcome layer that anchors subsequent multidimensional evaluation of linguistic reasoning.

3.4.2 Reasoning Quality Score

Problem-Level Accuracy does not reveal whether correct outputs are supported by valid linguistic reasoning. In linguistics olympiad problems, correct answers may result from superficial pattern matching or partial memorization. To address this limitation, we introduce a Reasoning Quality Score (R_i) that evaluates the extent to which models explicitly induce and apply rules from the provided data.

Explanations generated by models are evaluated by expert annotators using a standardized scoring rubric (Table 2), yielding a score $R_i \in [0, 1]$ for each problem instance. Let M denote the number of annotators. The final reasoning score is computed as:

$$R_i = \frac{1}{M} \sum_{m=1}^{M} r_{im}, \tag{4}$$

where r_{im} is the score assigned by annotator m to problem p_i.

The rubric assesses logical coherence, empirical grounding, completeness of rule systems, and consistency of application. High scores reflect explicit hypothesis formation, systematic testing, and principled abstraction, while lower scores indicate descriptive, incomplete, or unsupported reasoning.

Annotators follow detailed guidelines and participate in calibration sessions to ensure scoring consistency across domains and difficulty levels. The Reasoning Quality Score enables identification of cases in which high accuracy is achieved through invalid reasoning and supports comparisons between prompting strategies and model architectures.

Although based on human evaluation, the formalized criteria provide a foundation for future automated assessment methods.

3.4.3 Composite Problem Score

To integrate surface-level correctness and reasoning validity into a unified evaluation signal, we define a Composite Problem Score (S_i) for each problem instance p_i:

$$S_i = \alpha A_i + (1 - \alpha) R_i, \quad \alpha = 0.7 \tag{5}$$

where A_i denotes Problem-Level Accuracy and R_i denotes the Reasoning Quality

Table 2: Scoring Rubric for Reasoning Quality

Score	Descriptor	Criteria
1.0	Fully correct	Correctly infers all relevant linguistic rules; reasoning is logically coherent, complete, and strictly grounded in the provided data; final answer follows unambiguously from the explanation.
0.75	Largely correct	Core linguistic rules correctly identified, minor omissions or imprecision; reasoning mostly coherent and leads to a correct solution.
0.50	Partially correct	Some correct insights or sub-rules identified, overall rule system incomplete or inconsistently applied.
0.25	Flawed	Substantial errors or unsupported assumptions; explanation poorly aligned with problem statement.
0.00	Invalid / Absent	No meaningful reasoning, explanation incorrect or irrelevant.

Score.

This formulation reflects that correct outputs are necessary but insufficient indicators of systematic reasoning. The weighting parameter $\alpha = 0.7$ prioritizes answer correctness while preserving a substantial contribution from explanatory quality.

The composite score penalizes superficial pattern matching and distinguishes partial understanding from complete failure. Dataset-level performance is summarized as:

$$S_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{6}$$

where N is the number of problems in the dataset.

Aggregate scores are interpreted together with domain-specific and difficulty-stratified results. The metric also supports cross-condition analyses and sensitivity studies in which α is varied to examine alternative evaluation priorities.

3.4.4 Generalization and Hallucination

Systematic linguistic reasoning requires both reliable generalization and adherence to closed-world constraints. We therefore introduce complementary metrics for generalization and hallucination.

Generalization Gap

Let P_s denote the set of structurally familiar (seen) instances within problems, and P_u denote held-out or structurally novel (unseen) instances. The Generalization Gap is defined as:

$$\Delta G = S_{\text{avg}}(P_s) - S_{\text{avg}}(P_u) \tag{7}$$

where $S_{\text{avg}}(\cdot)$ represents the mean Composite Problem Score over the corresponding subset.

Small values of ΔG indicate stable abstraction and rule transfer, whereas large values suggest reliance on memorization or shallow heuristics. Domain-specific gaps ΔG_d can also identify particularly challenging linguistic phenomena.

Hallucination Rate

While generalization measures extrapolative competence, the Hallucination Rate quantifies violations of the closed-world assumption. For each problem p_i, a binary indicator h_i is defined:

$$h_i = \begin{cases} 1 & \text{if external or unverifiable information is present,} \\ 0 & \text{otherwise.} \end{cases}$$

The Hallucination Rate is:

$$H = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(h_i = 1) \tag{8}$$

Hallucinations include references to external linguistic knowledge not inferable from the problem statement. Detection is performed by trained annotators using standardized procedures. Low values of H indicate epistemic discipline, while high values reflect dependence on memorized priors. Low hallucination rates alone, however, do not guarantee high reasoning quality.

Joint Interpretation

Generalization and hallucination metrics are interpreted jointly to characterize model behavior. Four typical regimes can be distinguished:

1. Low ΔG, Low H: Robust rule induction and disciplined reasoning.
2. Low ΔG, High H: Apparent generalization driven by external knowledge.
3. High ΔG, Low H: Faithful but brittle reasoning with limited abstraction.
4. High ΔG, High H: Reliance on heuristics and memorization.

This typology enables fine-grained diagnosis of model reasoning strategies and ensures that high performance reflects genuine inductive reasoning rather than opportunistic use of background knowledge.

3.4.5 Additional Metrics

In addition to the core metrics above, we incorporate complementary measures from prior computational linguistics and explainability research. These metrics function as diagnostic tools rather than primary benchmarks, enabling comparison with established evaluation traditions and providing additional interpretative signals.

One such measure is **Surface Form Accuracy ($A_{\mathbf{surf}}$)**, computed using character-level n-gram overlap (chrF++ [13]). This metric captures graded similarity between model outputs and gold references, particularly in problems involving transliteration, inflection, or multi-token outputs where strict exact matching may be overly rigid. While not a substitute for problem-level correctness, it reflects structural correspondence at a subword level.

A second auxiliary measure is **Reasoning Faithfulness ($R_{\mathbf{faith}}$)**, assessed through rubric-guided evaluation of intermediate reasoning traces. This score reflects the extent to which explanations remain grounded in the provided data and logically support the final answer, helping distinguish valid inference from post hoc rationalization.

To combine these auxiliary dimensions, we define the **Strict Linguistic Reasoning Score (SLRS)**:

$$SLRS = (1 - H) \cdot [\alpha A_{\text{surf}} + (1 - \alpha) R_{\text{faith}}], \quad \alpha = 0.6 \tag{9}$$

where H denotes the hallucination rate. This formulation ensures that high scores require structural similarity, faithful reasoning, and absence of unsupported inferences.

Overall, these optional metrics enrich diagnostic analysis and support fine-grained model comparison, while the primary evaluation remains grounded in accuracy, reasoning quality, and generalization performance.

3.5 Human Baselines and Inter-Rater Reliability

To contextualize model performance, we establish human baselines using solutions produced by trained linguistics students and former olympiad participants. These individuals possess explicit experience with rule-induction tasks and serve as an upper-bound reference for systematic reasoning under the same informational constraints imposed on the models [11, 12]. Human solvers are provided with the identical problem statements and are instructed to produce both final answers and step-by-step reasoning, mirroring the model evaluation protocol.

Multiple expert annotators independently evaluate a stratified subset of model outputs covering all linguistic domains and difficulty levels. Each annotator assigns a reasoning quality score according to the rubric in Table 2, without access to the model identity or other evaluators' scores. Inter-rater agreement is quantified using Cohen's κ for pairwise agreement and Krippendorff's α for multi-annotator settings. High agreement scores indicate that the rubric yields stable and reproducible judgments, mitigating subjectivity in reasoning evaluation.

In addition, human performance can be compared to model outputs in terms of error types and reasoning completeness, providing insight into which linguistic phenomena are inherently challenging versus those where models systematically fail. This dual analysis enables a nuanced understanding of the reasoning gap between humans and LLMs [8].

3.6 Error Analysis

Beyond aggregate metrics, a detailed qualitative error analysis should be conducted to identify systematic failure modes in LLM reasoning. Errors are manually categorized into four primary types:

1. **Incorrect Rule Induction:** The model proposes coherent but empirically unsupported linguistic rules (e.g., misgeneralizing a morphological suffix).
2. **Failure to Apply a Correct Rule:** The model identifies the relevant pattern but applies it inconsistently or incompletely, leading to partially correct answers.
3. **Misinterpretation of Problem Constraints:** Errors arise from overlooking explicit restrictions, misreading examples, or misprocessing linguistic problem formatting.
4. **Language-Specific Misinterpretation:** Failures stem from unfamiliar scripts, morphological complexity, or typological features absent from high-resource training data [6, 14].

Representative examples from each category are to be presented to illustrate the

model's reasoning limitations. This analysis serves two purposes: it distinguishes between superficial pattern matching and deeper linguistic reasoning failures, and it identifies which domains (e.g., morphology, syntax, compositional systems) pose the greatest challenge.

Error analysis also informs potential improvements in prompting strategies, agentic verification loops, and model fine-tuning [7].

3.7 Statistical Analysis

Quantitative results are reported using descriptive statistics, including mean composite scores, standard deviations, and confidence intervals, computed separately for each linguistic domain and difficulty level. For comparisons between model variants under matched problem conditions, non-parametric paired tests are employed, such as the Wilcoxon signed-rank test, which do not assume normality and are suitable for ordinal or bounded metrics like reasoning quality [9].

For multiple comparisons, p-values are adjusted using Bonferroni or Holm-Bonferroni corrections to control family-wise error. Effect sizes (e.g., Cohen's d or Cliff's delta) are reported alongside significance to quantify the magnitude of observed differences. Statistical analysis also examines correlations between problem difficulty, linguistic domain, and error types, highlighting systematic patterns in model reasoning capabilities.

Where possible, visualization of performance distributions, error frequencies, and reasoning scores provides an intuitive complement to numerical analysis, enhancing interpretability for both human and computational audiences.

3.8 Reproducibility, Limitations, and Ethical Considerations

Reproducibility is a central design goal. All benchmark problems, prompts, scoring rubrics, and annotation guidelines are released alongside evaluation code. Model versions, inference parameters, and random seeds are explicitly documented to allow exact replication of reported results [4]. Anonymized human annotation data and inter-rater agreement statistics are provided to support independent verification.

Despite these precautions, several limitations persist:

- **Training Data Contamination:** Some olympiad problems or solutions may be publicly available online, making full elimination of pre-training leakage challenging.
- **Reasoning vs. Pattern Matching:** The rubric-based evaluation improves interpretability, but distinguishing genuine multi-step reasoning from sophisticated pattern matching is inherently imperfect [8, 11].
- **Coverage Bias:** Certain linguistic traditions or problem styles may be underrepresented, biasing evaluation toward particular reasoning strategies.

Ethically, the benchmark avoids personal or sensitive data. Care is taken to include typologically diverse languages, scripts, and low-resource systems, mitigating language and cultural biases. The methodology is intended to complement existing NLP evaluation paradigms, providing a linguistically grounded perspective on systematic reasoning in LLMs [1, 2].

4. Benchmark Usage and Evaluation Protocol

This section specifies how the proposed benchmark and evaluation framework are intended to be used in future studies. As this work presents a methodological proposal rather than an empirical evaluation, the focus is on defining standardized usage scenarios, interpretation guidelines, and reporting conventions. The benchmark is conceived as a diagnostic instrument for linguistic reasoning, not as a leaderboard-oriented performance test.

4.1 Intended Evaluation Scenarios

The proposed framework is designed to support multiple evaluation scenarios that correspond to distinct research questions about systematic linguistic reasoning in large language models.

First, the benchmark enables controlled comparisons between different prompting and reasoning protocols. Researchers may contrast direct, single-pass prompting with structured multi-step or agentic approaches, as defined in Section 3.3. Because all problems satisfy the closed-world requirement, differences in performance can be attributed to reasoning strategy rather than access to external knowledge.

Second, the benchmark supports cross-model comparisons under fixed conditions. By holding the dataset, scoring metrics, and evaluation protocol constant, researchers can examine how architectural differences or training regimes affect rule induction and generalization, independent of memorized linguistic content.

Third, the framework is suitable for generalization studies. By separating training examples from structurally similar but unseen test items within each problem, researchers can analyze systematic generalization using the Generalization Gap metric defined in Section 3.4.

Crucially, the benchmark is not intended to measure broad linguistic competence. Instead, it isolates a narrow but theoretically meaningful capability: the induction and application of abstract linguistic rules from sparse, artificialized input.

4.2 Evaluation Procedure

To ensure reproducibility and comparability across studies, we define a standardized evaluation procedure governing problem presentation, model interaction, and score computation. Each evaluation run consists of four stages.

Problem Presentation. Each model is presented with a single linguistic problem instance comprising a set of training examples and one or more held-out test queries. No external tools, retrieval mechanisms, or linguistic resources are permitted. All information required to solve the problem must be contained within the prompt.

Reasoning Trace Generation. Models are instructed to produce both a final answer and an explicit reasoning trace. The reasoning trace is required for the computation of the Reasoning Quality Score and for detecting closed-world violations. Suppressing intermediate reasoning is therefore incompatible with the proposed evaluation framework.

Hallucination Detection. The reasoning trace is examined for references to real-world linguistic facts, typological generalizations, or external language knowledge. Any such reference constitutes a violation of the closed-world assumption and is recorded when computing the Hallucination Rate.

Score Computation. For each problem instance, Problem-Level Accuracy, Reasoning Quality Score, and the Composite Problem Score are computed as defined in Section 3.4. These scores jointly characterize both outcome correctness and the validity of the reasoning process.

4.3 Interpretation of Evaluation Metrics

The evaluation metrics proposed in Section 3.4 are intended to be interpreted jointly, as each captures a different aspect of model behavior.

Problem-Level Accuracy measures whether the model produces the correct output form, but does not provide insight into how the answer was derived. As shown in prior work, correct outputs may result from coincidental pattern matching or memorization rather than systematic reasoning.

The Reasoning Quality Score evaluates the internal coherence and data-consistency of the model's inferred rule system. High scores indicate explicit identification and consistent application of abstract rules, while low scores reflect heuristic or logically invalid reasoning, even when the final answer is correct.

The Composite Problem Score integrates these two dimensions, ensuring that surface correctness and reasoning validity are not evaluated in isolation. This composite measure is intended as the primary problem-level indicator within the framework.

Importantly, scores should be reported separately for different linguistic domains (morphological, syntactic, compositional), as aggregate values may obscure domain-specific reasoning failures.

4.4 Generalization and Hallucination Analysis

Beyond problem-level performance, the framework explicitly targets two systemic behaviors: generalization and hallucination.

The Generalization Gap quantifies the discrepancy between performance on structurally familiar instances and unseen test items. A large gap suggests reliance on surface-level associations rather than abstraction of underlying rules, particularly in compositional systems such as counting or kinship structures.

The Hallucination Rate captures the frequency with which models violate the closed-world assumption by introducing external linguistic knowledge. High hallucination rates indicate that apparent success may be driven by memorized typological facts rather than in-context learning.

Analyzing these metrics together allows researchers to distinguish between models that fail due to insufficient inductive capacity and those that succeed only by bypassing the intended reasoning constraints.

4.5 Reporting and Comparability Guidelines

To support meaningful comparison across studies, we recommend that future work using the proposed benchmark adhere to a standardized reporting format. At minimum,

researchers should report: (i) Problem-Level Accuracy, (ii) Reasoning Quality Score, (iii) Composite Problem Score, (iv) Generalization Gap, and (v) Hallucination Rate, disaggregated by linguistic domain.

Comparisons across fundamentally different evaluation setups—such as single-pass prompting versus agentic reasoning pipelines—should be accompanied by explicit discussion of the additional computational and structural assumptions involved.

These guidelines are intended to ensure that the benchmark functions as an interpretable evaluation instrument rather than as a single-number performance metric.

5. Conclusions and Future Work

In this work, we have proposed a benchmark and evaluation framework for assessing systematic linguistic reasoning in large language models. The methodology emphasizes problems derived from official linguistics olympiads, the preservation of a closed-world problem structure, and careful stratification across linguistic phenomena and problem difficulty. By integrating both problem-level correctness and reasoning quality metrics, the framework enables a nuanced assessment of model capabilities beyond surface-level accuracy.

Crucially, our benchmark is designed as a diagnostic tool rather than a competition leaderboard. It supports evaluation scenarios that isolate the effects of prompting strategies, architectural differences, and generalization ability under controlled conditions. The incorporation of explicit reasoning traces, hallucination detection, and domain-specific analysis ensures that performance reflects genuine inductive reasoning rather than memorization or pattern matching.

Looking forward, several avenues for future work emerge. First, the benchmark can be extended to include additional languages, scripts, and typological phenomena, particularly from under-represented linguistic families, to further stress-test compositional reasoning and cross-linguistic generalization. Second, agentic or multi-step evaluation protocols can be refined, potentially incorporating verification loops, analogical prompting, or self-correction mechanisms to study their impact on reasoning quality. Third, while the current framework relies on human-annotated reasoning scores, the development of automated or semi-automated reasoning evaluation tools could enhance scalability and reproducibility.

Finally, the benchmark opens opportunities for cross-disciplinary research, linking computational linguistics, cognitive science, and language pedagogy. By providing a standardized, linguistically grounded instrument for evaluating reasoning, this work lays the foundation for future studies that rigorously probe the capabilities and limitations of large language models in systematic, symbolic-like problem solving.

Acknowledgements

This research is supported by the Scientific Research Fund of "Angel Kanchev" University of Ruse, Bulgaria.

References

[1] A. Bean, S. Hellsten, H. Mayne, J. Magomere, E. A. Chi, R. Chi, S. A. Hale, and H. R. Kirk. 2024. LINGOLY: a benchmark of olympiad-level linguistic

reasoning puzzles in low-resource and extinct languages, *Proceedings of the 38th International Conference on Neural Information Processing Systems (NIPS '24)*, Vol. 37. Curran Associates Inc., Red Hook, NY, USA, Article 825, pp. 26224–26237, 2024. https://dl.acm.org/doi/10.5555/3737916.3738741

[2] A. R. Bhattacharya, I. Papadimitriou, K. Davidson, D. Alvarez-Melis, Investigating the interaction of linguistic and mathematical reasoning in language models using multilingual number puzzles, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28322–28332, 2025, https://aclanthology.org/2025.emnlp-main.1438.pdf

[3] N. A. Chi, T. Malchev, R. Kong, R. A. Chi, L. Huang, E. A. Chi, R. T. McCoy, D. Radev, *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2024)*, pp. 113–119, 2024, https://aclanthology.org/2024.sigtyp-1.14.pdf

[4] S. Goyal, S. Dan, IOLBENCH: Benchmarking LLMs on Linguistic Reasoning, *arXiv preprint arXiv:2501.04249*, 2025.

[5] J. Khouja, L. Yang, R. Kearns, K. Korgul, V. Neacsu, S. Hellsten, A. Bean, H. Mayne, A. Mahdi, LINGOLY-TOO: Disentangling Reasoning from Knowledge with Templatised Orthographic Obfuscation, *arXiv preprint arXiv:2503.02972*, 2025.

[6] J. Hwang, K. Tanmay, S.-J. Lee, A. Agrawal, P. P. Liang, H. Palangi, K. Ayush, Learn Globally, Speak Locally: Bridging the Gaps in Multilingual Reasoning, *arXiv preprint arXiv:2507.05418*, 2025.

[7] D.-C. Lian, R.-S. Huang, P.-E. Chen, C. Lim, Y.-K. Lin, G.-Y. Tseng, T.-C. Yang, Z.-Y. Lin, P.-C. Chen, S.-K. Hsieh, LingBench++: A Linguistically-Informed Benchmark and Reasoning Framework for Multi-Step and Cross-Cultural Inference with LLMs, *arXiv preprint arXiv:2507.16809*, 2025.

[8] H. Lin, X. Wang, R. Yan, B. Huang, H. Ye, J. Zhu, Z. Wang, J. Zou, J. Ma, Y. Liang, Generative Evaluation of Complex Reasoning in Large Language Models (KUMO), *arXiv preprint arXiv:2504.02810*, 2025.

[9] Z. Lin, Z. Gou, T. Liang, R. Luo, H. Liu, Y. Yang, CRITICBENCH: Benchmarking LLMs for Critique-Correct Reasoning, *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1552–1587, 2024, https://aclanthology.org/2024.findings-acl.91.pdf

[10] N. Majmudar, E. Filatova, Can LLMs Solve and Generate Linguistic Olympiad Puzzles?, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19163–19200, 2025, https://aclanthology.org/2025.emnlp-main.969.pdf

[11] R. Ramji, K. Ramji, Inductive Linguistic Reasoning with Large Language Models, *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22783–22810, 2025, https://aclanthology.org/2025.findings-acl.1171.pdf

[12] G. G. Sahin, Y. Kementchedjhieva, P. Rust, I. Gurevych, PuzzLing Machines: A Challenge on Learning From Small Data, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1241–1254, 2020, https://aclanthology.org/2020.acl-main.115.pdf

[13] E. Sánchez, B. Alastruey, C. Ropers, P. Stenetorp, M. Artetxe, M. R. Costa-jussà, Linguini: A Benchmark for Language-Agnostic Linguistic Reasoning,

arXiv preprint arXiv:2409.12126, 2024.

[14] Y.-F. Shih, Z.-L. Lin, S.-K. Hsieh, Reasoning Over the Glyphs: Evaluation of LLM's Decipherment of Rare Scripts, *arXiv preprint arXiv:2501.17785*, 2025.

[15] A. Waldis, Y. Perlitz, L. Choshen, Y. Hou, I. Gurevych, Holmes: A Benchmark to Assess the Linguistic Competence of Language Models, *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 1616–1647, 2024, https://aclanthology.org/2024.tacl-1.88.pdf

[16] H. Zhu, Y. Liang, W. Xu, H. Xu, Evaluating Large Language Models for In-Context Learning of Linguistic Patterns in Unseen Low Resource Languages, *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025)*, pp. 414–426, 2025, https://aclanthology.org/2025.loreslm-1.31.pdf