Autonomous Evaluation of LLMs for Truth Maintenance and Reasoning Tasks

Rushang Karia*, Daniel Bramblett*, Daksh Dobhal, Siddharth Srivastava

In Proceedings of ICLR 2025
IJCAI 2025 Workshop on User-Aligned Assessment of Adaptive AI Systems

IJCAI 2025 MONTREAL

AAI Autonomous Agents and Intelligent Robots

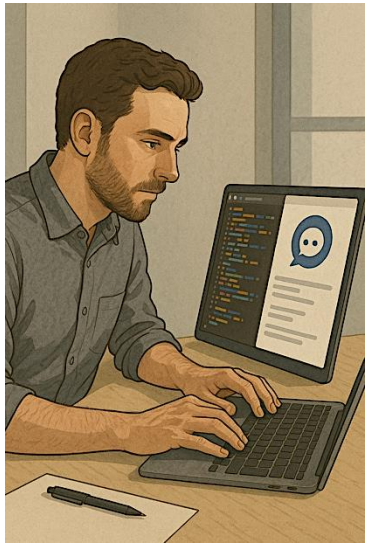
ASU Arizona State University

Motivation: Effective Human-AI Communication

Collaboration with AI requires clear, effective communication.

Promising approach: having the AI between the AI and the human.

- natural language (e.g., describing code)
- formal language (e.g., code, system specifications)



This approach is already being used in vibe coding.

- Generates code from a prompt
- Explains the generated code

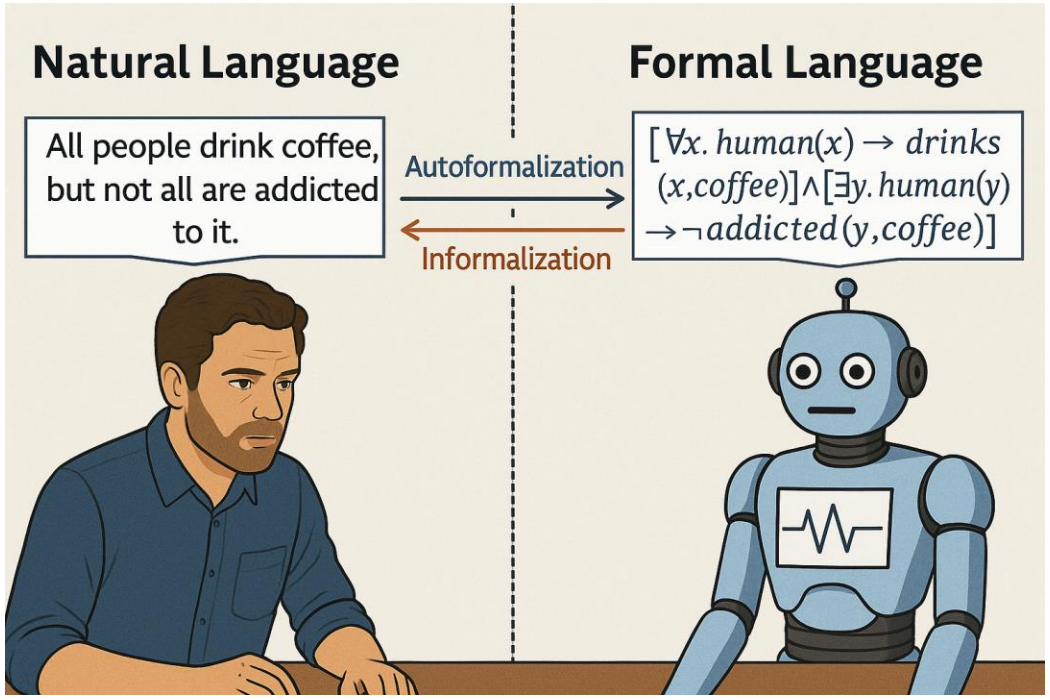
This requires the AI to be semantically accurate doing this translation.

Objective: Assessment of LLM Truth Maintenance

Autoformalization: generating formal language from natural language.

Informalization: generating natural language from formal language.

Truth Maintenance: do these translations maintain semantic truth?



Challenges With Current Approaches to LLM Assessment

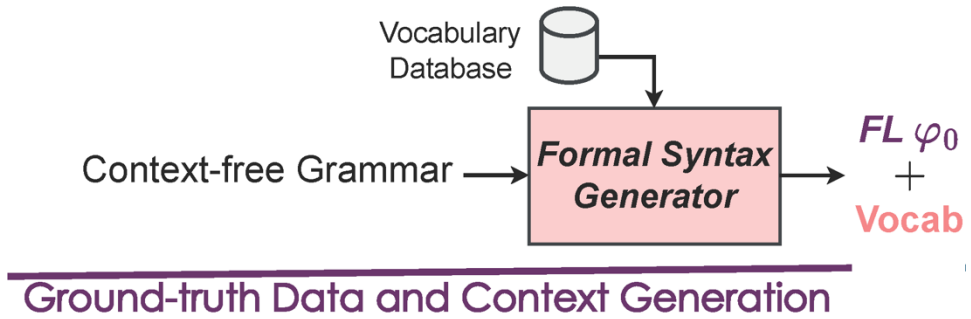
1. Benchmark Contamination Problem: Risk of models training on evaluation data.
2. Difficult and expensive for expert annotators to construct new, high-quality datasets.
3. Incomplete set of ground truths (e.g., HumanEval) and imperfect existing autonomous evaluations metrics (e.g., BLEU) provide an inaccurate assessment of LLM capabilities.

BLEU("the weather is sunny and warm", "the weather is **not** sunny and warm") = 0.673

BLEU("the weather in Montreal is sunny and warm", "the weather in Montreal is **not** sunny and warm") = 0.767

References: BLEU [Papineni et al., *ACL* 2002], HumanEval [Chen et al., arXiv:2107.03374, 2021].

AutoEval Process Example



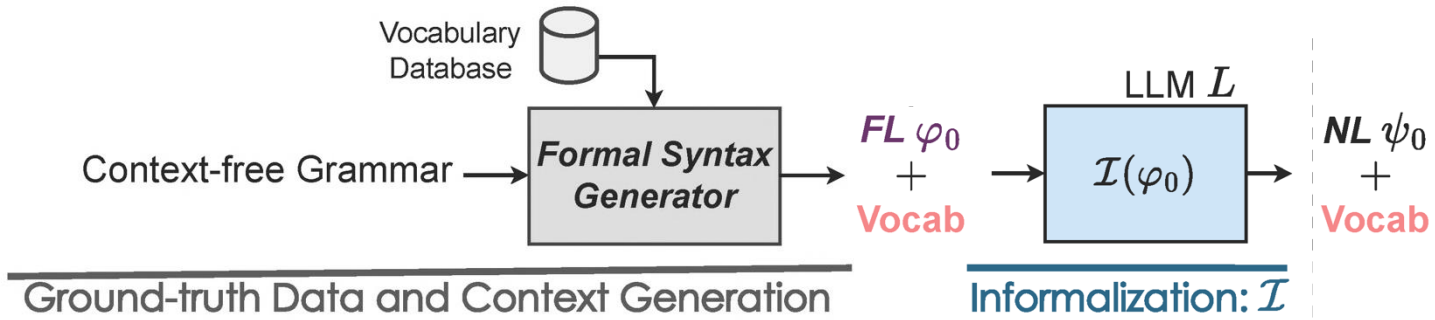
Propositional Logic Context-Free Grammar

$$S \rightarrow (S \wedge S) | (S \vee S)$$
$$S \rightarrow \neg S$$
$$S \rightarrow \neg v | v$$

Formal Language String + Vocab

$\varphi_0 = p_1 \wedge p_2 \wedge p_1$
p_1: it is raining
p_2: it was sunny yesterday

AutoEval Process Example



Formal Language String + Vocab

$\varphi_0 = p_1 \wedge p_2 \wedge p_1$

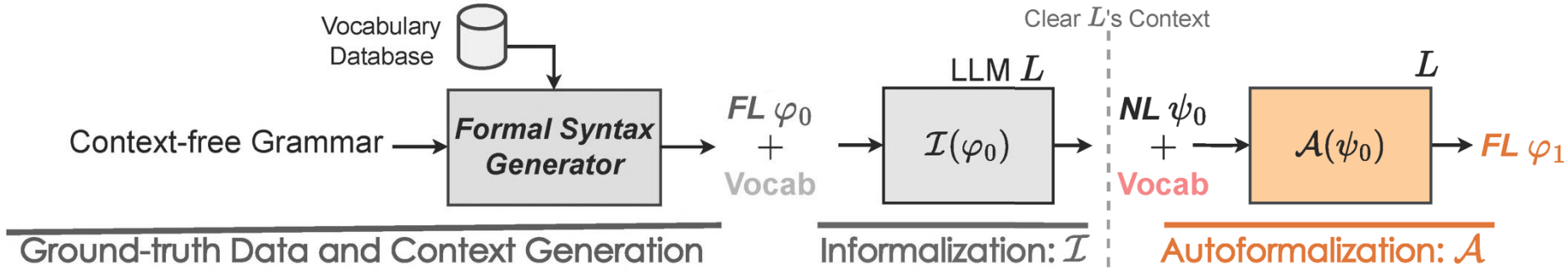
p_1: it is raining

p_2: it was sunny yesterday

Informalization Using LLM L

$\psi_0 =$ The sun was bright the day before whilst it is raining heavily today.

AutoEval Process Example



Natural Language String + Vocab

$\psi_0 =$ The sun was bright the day before whilst it is raining heavily today.

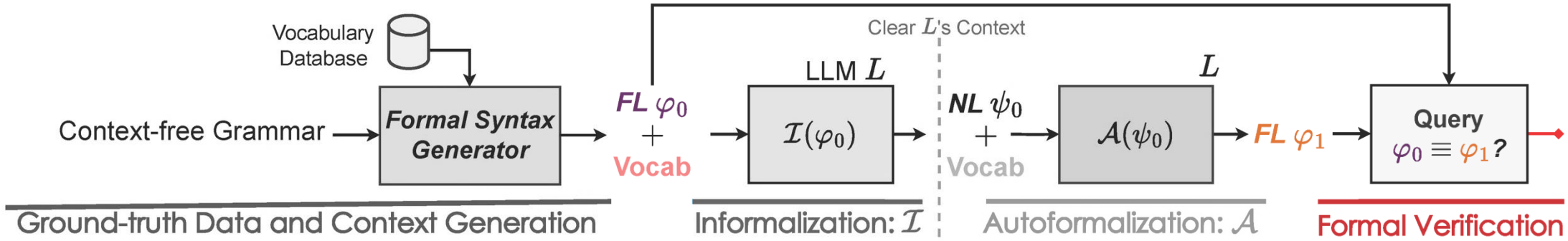
p_1: it is raining

p_2: it was sunny yesterday

Autoformalization Using LLM L

$$\varphi_1 = p_1 \wedge p_2$$

AutoEval Process Example



Original Formal Language String + Generated Formal Language String

Formal Verification

$\varphi_0 = p_1 \wedge p_2 \wedge p_1$
$\varphi_1 = p_1 \wedge p_2$

$\varphi_0 \equiv \varphi_1$ (use semantic formal verifier (e.g., Prover9 for FOL))

- Done for propositional logic, first-order logic, and regular expressions.
- **Can be extended to any formal language that has a semantic equivalence checker.**

References: Prover9 [McCune, 2010].

Theoretical Results

We prove the upper bound on false positive assessment rate.

- p_I: probability that an LLM L accurately informalizes formal language expression φ.
- p_A: probability that L accurately autoformalize natural language expression ψ.
- p_h : probability that if L hallucinates during both informalization and autoformalization that it will produce formal language equivalent to the original.

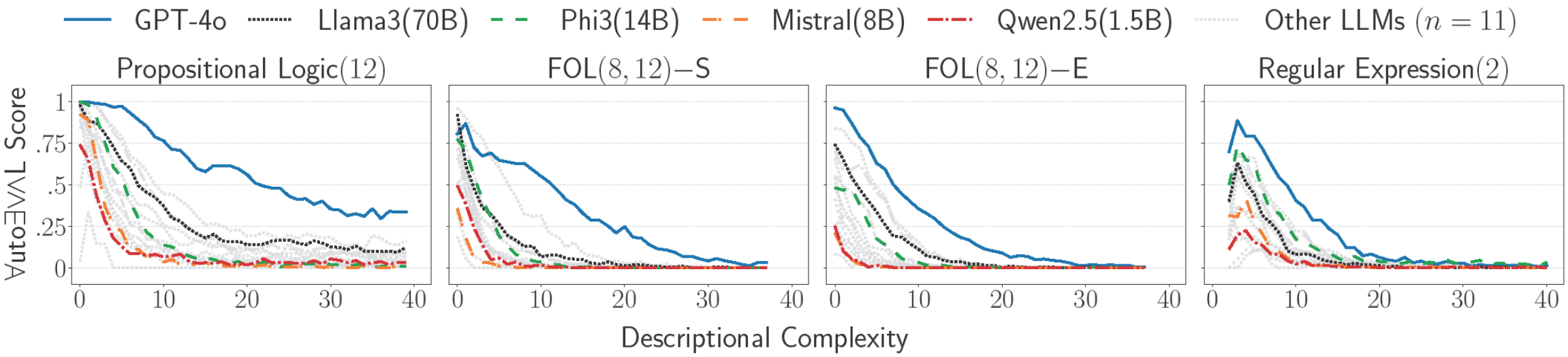
$$\varphi_0 \overset{I_L(\varphi_0)}{\longrightarrow} \psi_0 \overset{A_L(\psi_0)}{\longrightarrow} \varphi_1 \longrightarrow \cdots \overset{I_L(\varphi_{n-1})}{\longrightarrow} \psi_{n-1} \overset{A_L(\psi_{n-1})}{\longrightarrow} \varphi_n$$

$$\varphi_0 \equiv \varphi_1 \equiv \cdots \equiv \varphi_{n-1} \equiv \varphi_n$$

$$P(False\ Positive\ Assessment) = (1 - p_I)^n (1 - p_A)^n p_h^n$$

1. As LLMs improve, $p_I \to 1$, $p_A \to 1$, and $p_h \to 0$.
   - Meaning, $P(False\ Positive\ Assessment) \to 0$.
2. Probability can be reduced by iteratively applying this process.

Results: Truth Maintenance in Popular LLMs

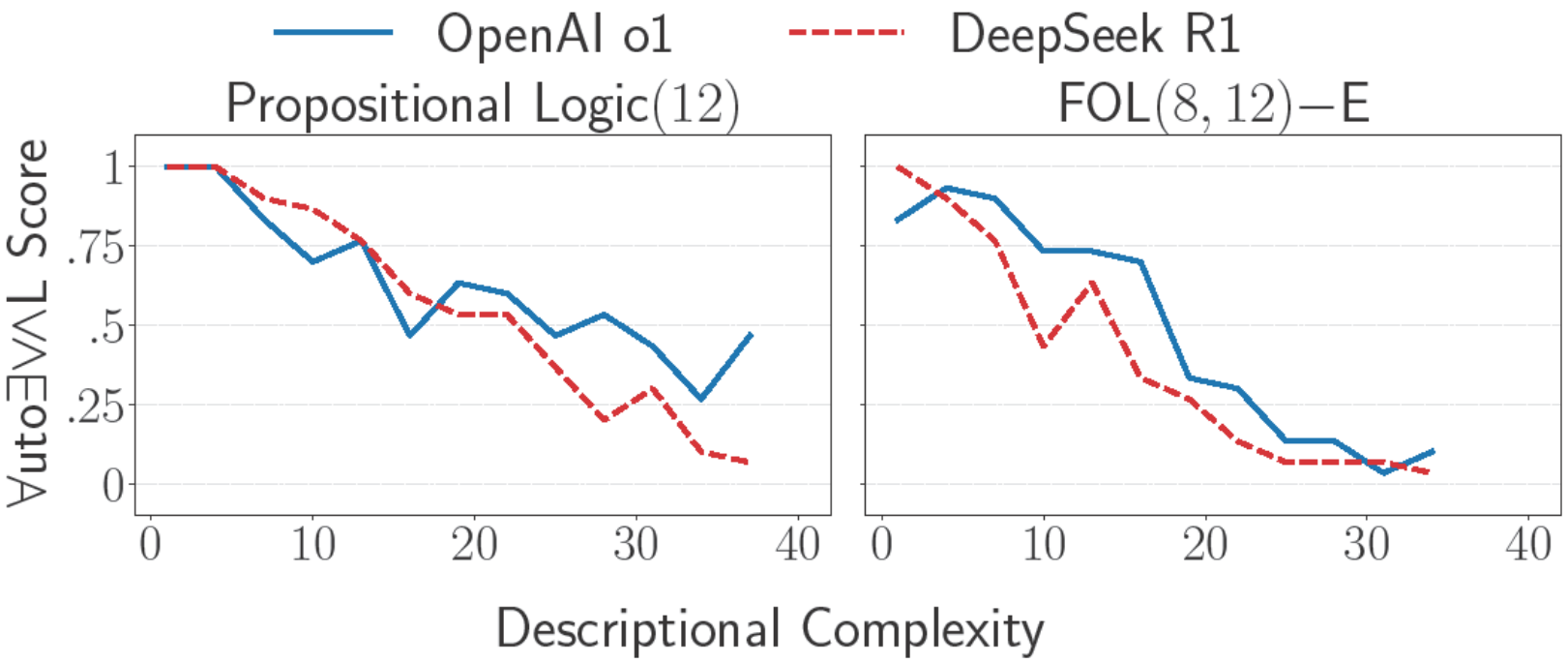


Evaluated 16 state-of-the-art, open and closed sourced LLMs.

- 3 types of formal language: propositional logic, first order logic, and regular expressions.
- 5 autogenerated datasets with approximately 85,000 unique evaluation examples.

All LLMs are less than 50% accurate on maintaining truth while translating formal language with 20 or more operators

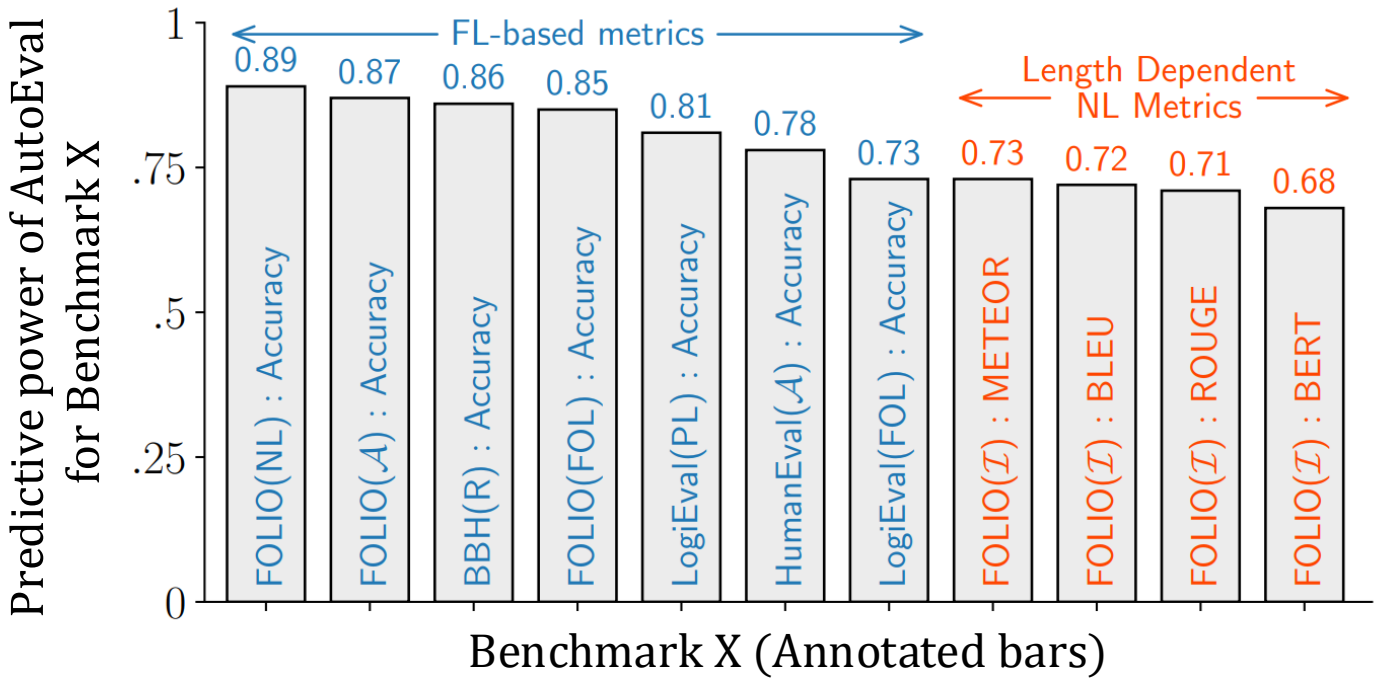
Results: Truth Maintenance in Popular LRMs



State-of-the-art large reasoning models are at most 50% accurate on maintaining truth while translating logic with 25 or more operators.

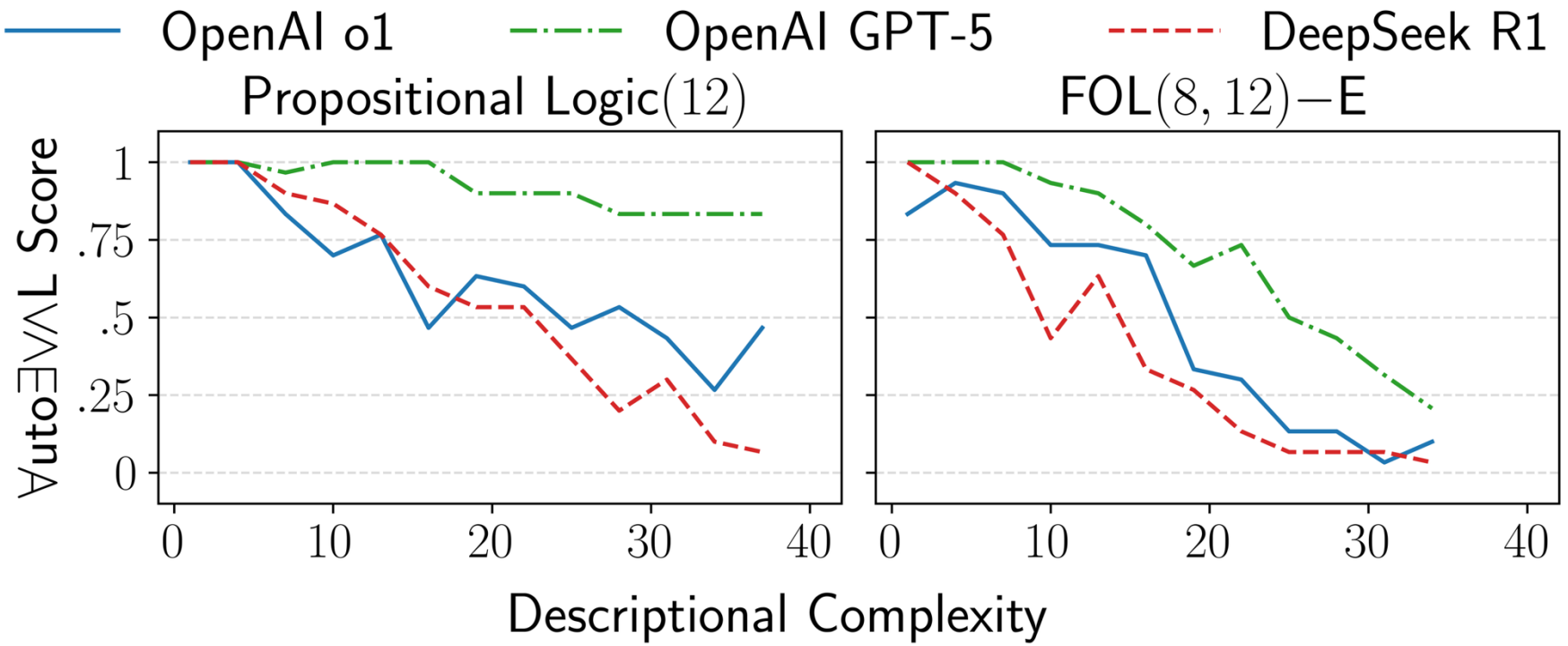
Results: AutoEval Predicts Performance on Other Tasks

The predictive power of benchmark A for benchmark B: probability that an LLM that ranks better in A also ranks better in B [$\Pr(L_1 \geq_B L_2 | L_1 \geq_A L_2)$].



A LLM's performance on AutoEval is predictive of its performance on reasoning tasks.

New Results: GPT-5



Newer models are improving on this task, but there is still significant room for improvement

Autonomous Evaluation of LLMs for Truth Maintenance and Reasoning Tasks

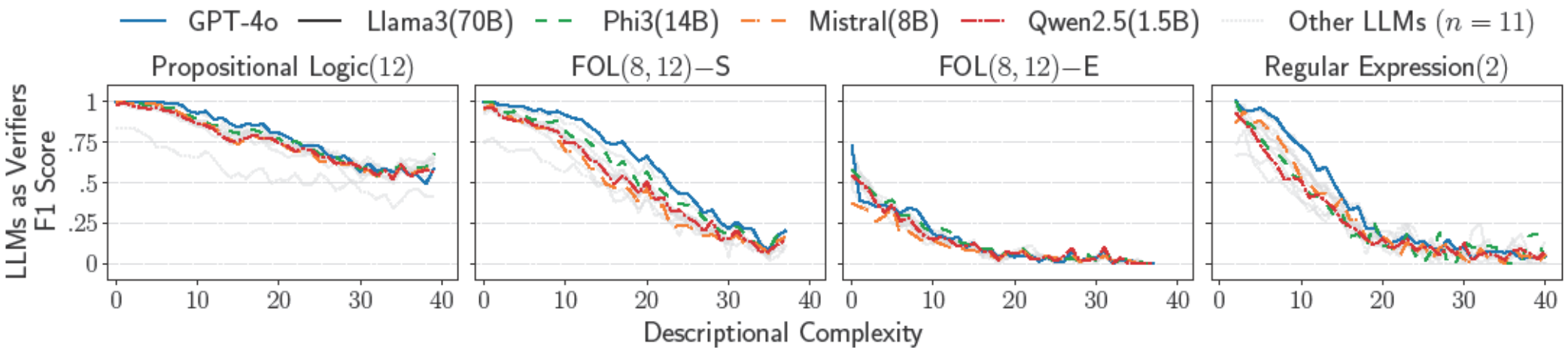
Rushang Karia*, Daniel Bramblett*, Daksh Dobhal, Siddharth Srivastava

IJCAI 2025 Workshop on User-Aligned Assessment of Adaptive AI Systems

Additional Results: LLMs as Equivalence Verifiers

Evaluated an LLM's to correctly evaluate whether the original and AutoEval process produced formal language were equivalent.

- Measured the F1 score compared to the formal verifier.



All LLMs, regardless of size, perform similarly as equivalence verifiers.