Autonomous Evaluation of LLMS for Truth Maintenance and Reasoning Tasks

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

Autonomous Agents and Intelligent Robots

Arizona State University

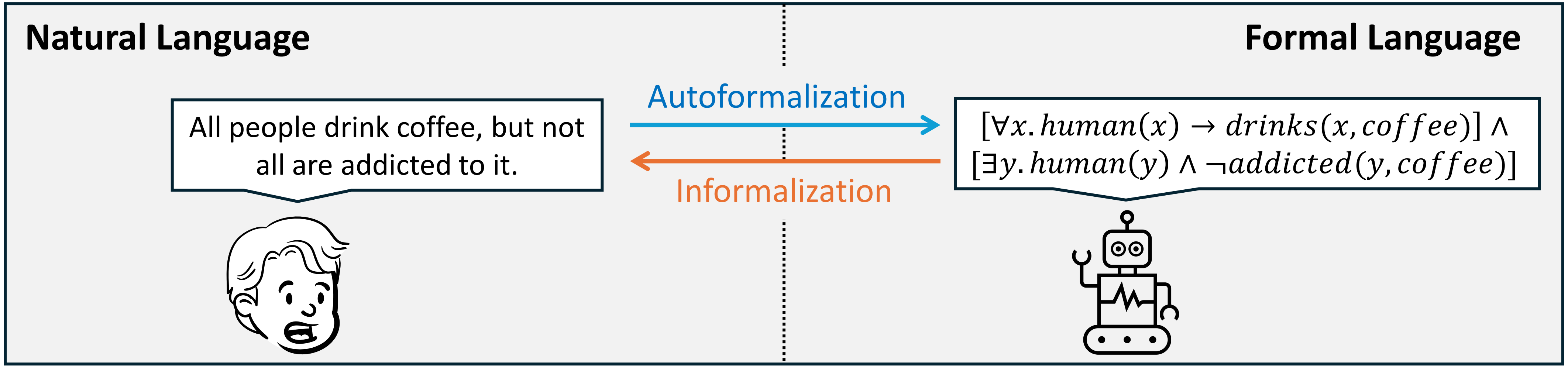
ICLR

Objective: Assessment of LLM Truth Maintenance

Autoformalization: generating formal language (e.g., code, system specifications) from natural language.

Informalization: generating natural language (e.g., describing code) from formal language.

Truth Maintenance: do the translations maintain truth.

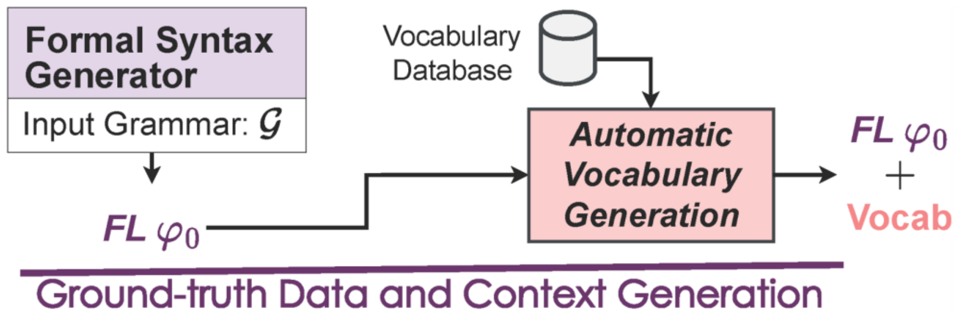


Challenges With Current Approaches to LLM Assessment

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.

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

AutoEval Process Example



Propositional Logic Context-Free Grammar

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

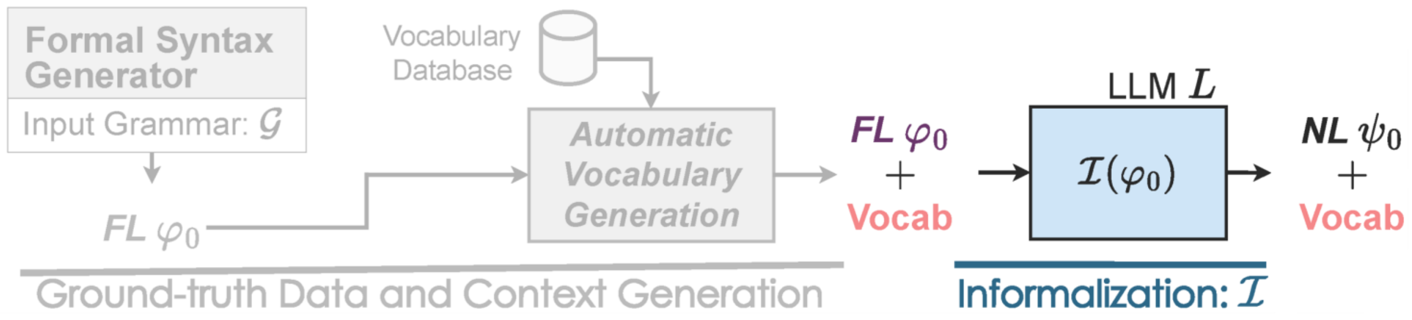
Formal Language String + Vocab

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

p_1: it is raining

p_2: it was sunny yesterday

AutoEval Process Example



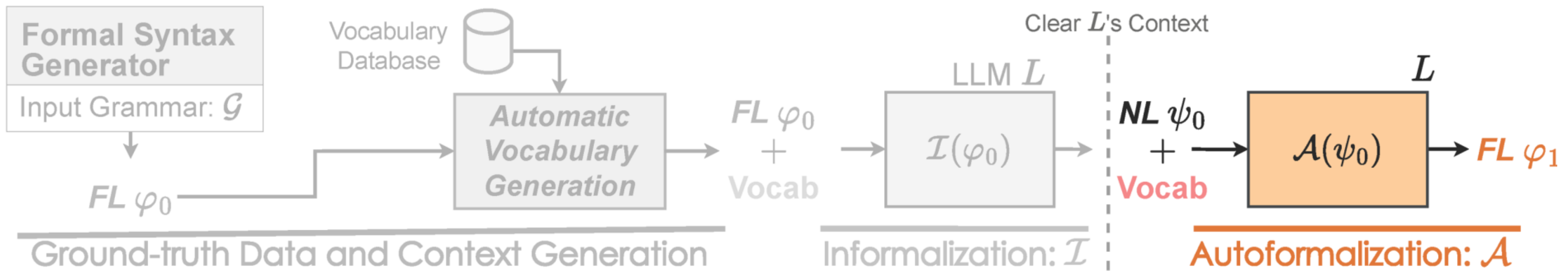
Formal Language String + Vocab

$\varphi_0 = p_1 \land p_2 \land p_1$
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

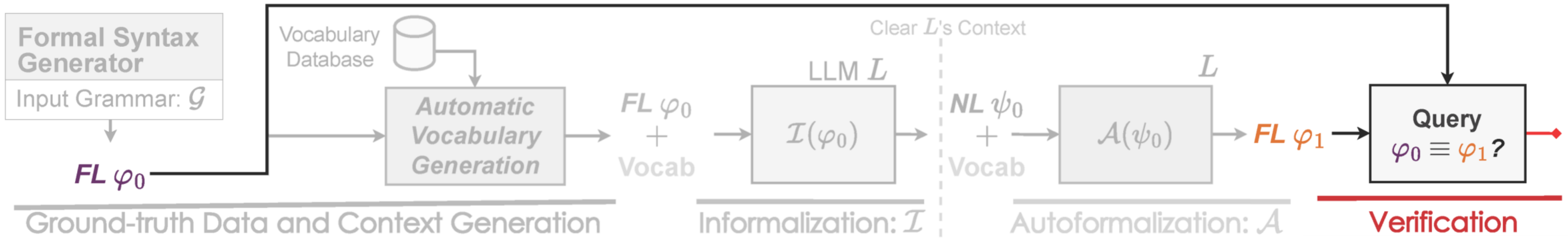
p_2: it was sunny yesterday

→

Autoformalization Using LLM L

$$\varphi_1 = p_1 \wedge p_2$$

AutoEval Process Example



Original Formal Language String + Generated Formal Language String

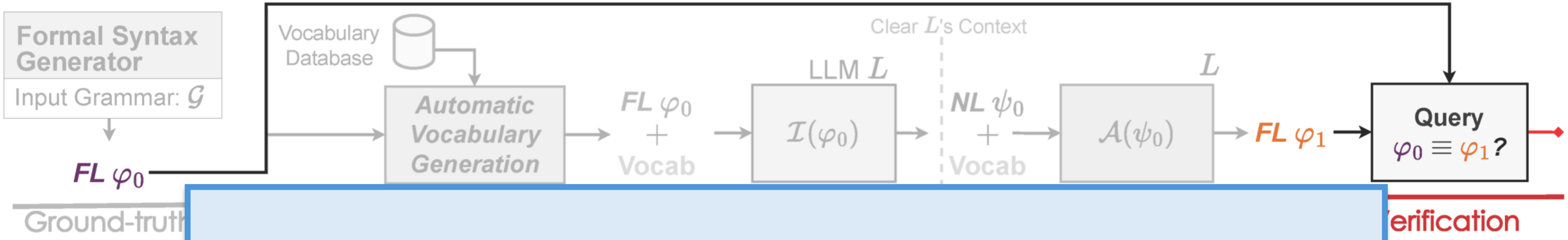
Formal Verification

$\varphi_0 = p_1 \land p_2 \land p_1$
$\varphi_1 = p_1 \land p_2$

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

References: Prover9 [McCune, 2010].

AutoEval Process Example



Theoretical Results: processes for error reduction and upper bound on false positive rates (see the paper).

Original Formal

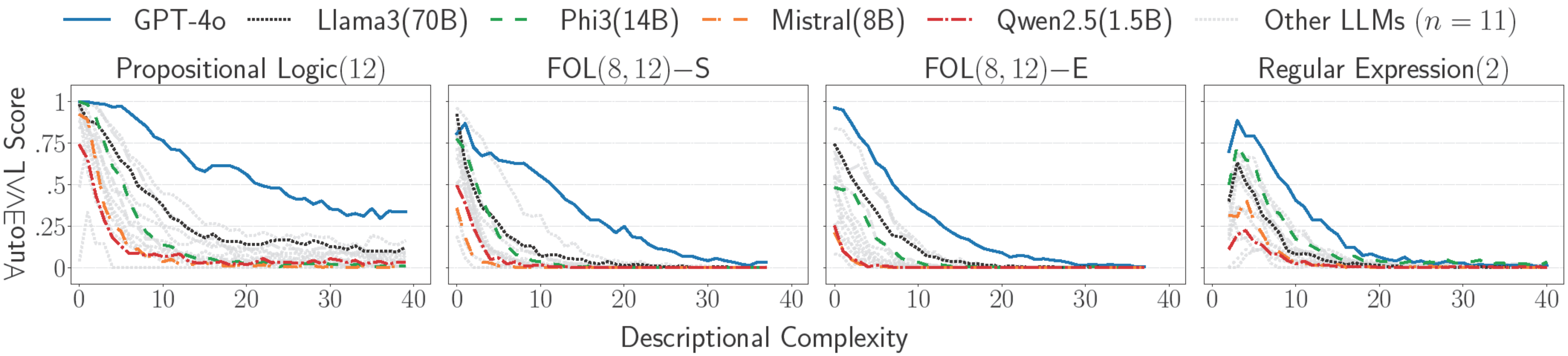
Generated Form

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

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

References: Prover9 [McCune, 2010].

Results: Truth Maintenance in Popular LLMs

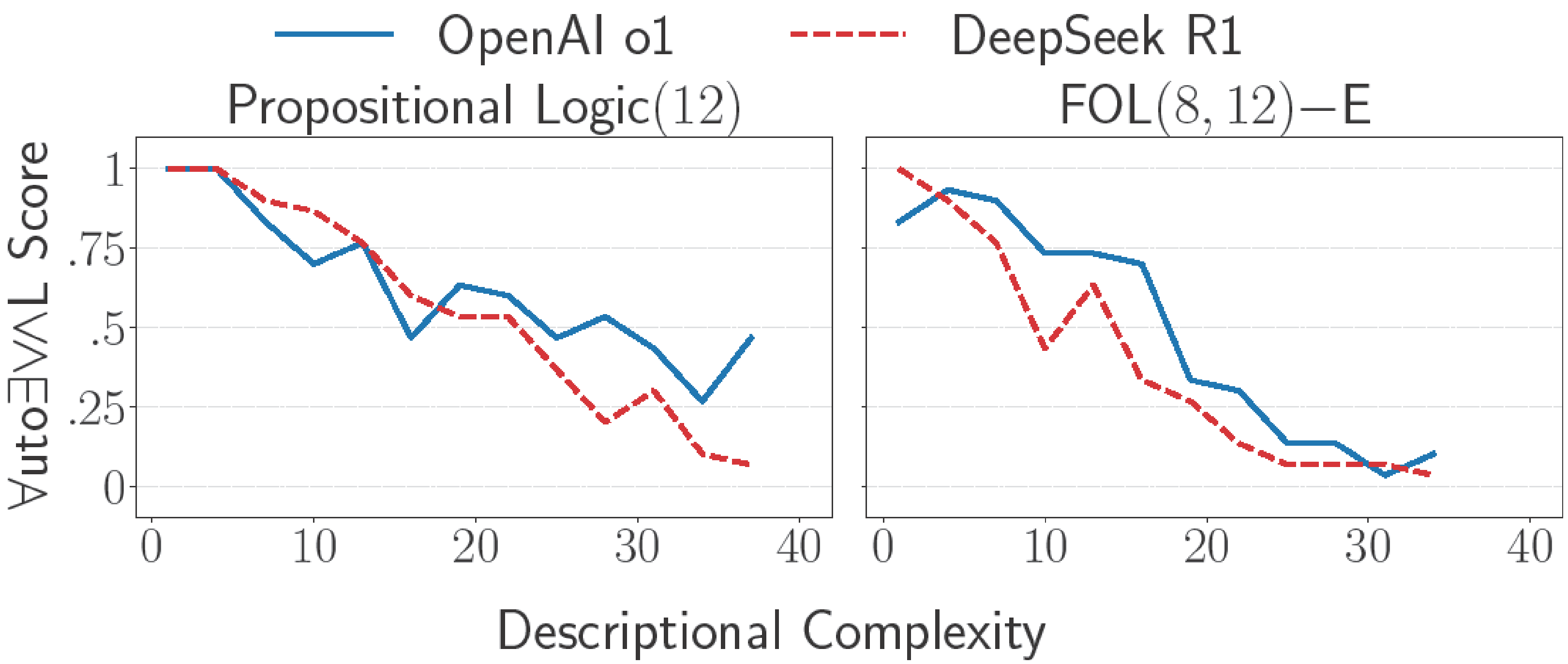


Evaluated 16 state-of-the-art, open and closed sourced LLMs.

- 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

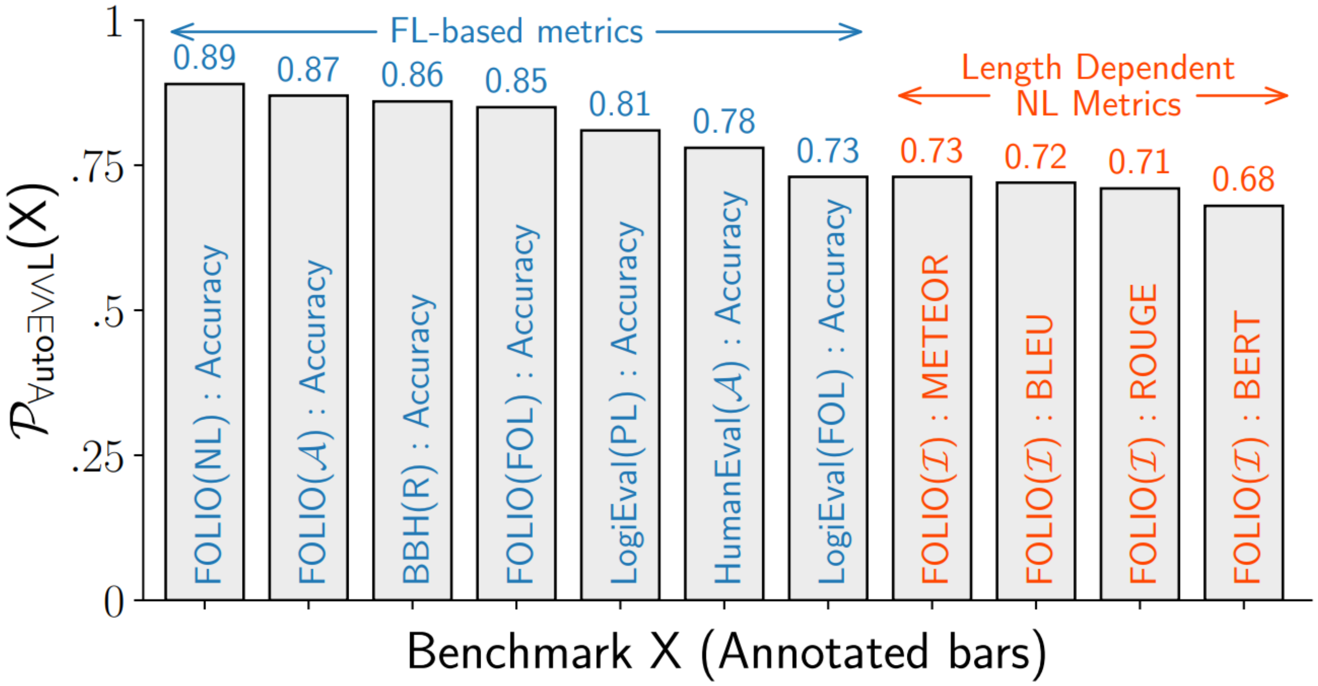


State-of-the-art large reasoning models are at most 50% accurate on maintaining truth while translating formal language with 25 or more operators.

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. Formally:

$$\text{Predictive Power of A for B} = \Pr(L_1 \geq_B L_2 | L_1 \geq_A L_2)$$



A LLM's performance on AutoEval is predictive of its performance on other formal-language-based tasks (e.g., reasoning).

Autonomous Evaluation of LLMS for Truth Maintenance and Reasoning Tasks

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

To know more about our lab, please visit: **https://aair-lab.github.io/projects/autoeval**
The code for this project can also be found at: **https://github.com/AAIR-lab/autoeval**