DARG: Dynamic Evaluation of Large Language Models via Adaptive Reasoning Graph

Zhehao Zhang, Jiaao Chen, Diyi Yang

DARTMOUTH

Georgia Tech College of Computing

Stanford University

Current LLM Evaluation Landscape

Mostly rely on **static benchmarks.**

Current LLM Evaluation Landscape

Mostly rely on **static benchmarks.**

Examples: GSM8K, BBQ, BigBench, etc.

Current LLM Evaluation Landscape

Mostly rely on **static benchmarks.**

Examples: GSM8K, BBQ, BigBench, etc.

Limitations:

- Vulnerability to data contamination
- Lack of adaptability to evolving LLM capabilities
- …

We need to evaluate LLMs dynamically!

Category Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Llama 3.1 70B	GPT 3.5 Turbo	Llama 3.1 405B	GPT-4 Omni	Claude 3.5 Sonnet
General MMLU Chat (0-shot, CoT)	73.0	72.3 (0-shot, non-CoT)	86.0	69.8	88.6	88.7	88.3
MMLU PRO (5-shot, CoT)	48.3	-	66.4	49.2	73.3	74.0	77.0
IFEval	80.4	73.6	87.5	69.9	88.6	85.6	88.0
Code HumanEval (0-shot)	72.6	54.3	80.5	68.0	89.0	90.2	92.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	86.0	82.0	88.6	87.8	90.5
Math GSM8K (8-shot, CoT)	84.5	76.7	95.1	81.6	96.8	96.1	96.4 (0-shot)
MATH (0-sho, CoT)	51.9	44.3	68.0	43.1	73.8	76.6	71.1
Reasoning ARC Challenge (0-shot)	83.4	87.6	94.8	83.7	96.9	96.7	96.7
GPQA (0-shot, CoT)	32.8	-	46.7	30.8	51.1	53.6	59.4
Tool use BFCL	76.1	-	84.8	85.9	88.5	80.5	90.2
Nexus (0-shot)	38.5	30.0	56.7	37.2	58.7	56.1	45.7

Does those numbers reflect their abilities?

Need for Dynamic Evaluation

- Adapt to LLM evolving capabilities

Need for Dynamic Evaluation

- Adapt to LLM evolving capabilities
- Generate evaluation data with controlled complexity

Need for Dynamic Evaluation

- Adapt to LLM evolving capabilities
- Generate evaluation data with controlled complexity
- Less concerns of data contamination issues

Prior works on dynamic evaluation

- Template-based methods (e.g., DyVal [1])
  - Limited to specific tasks (math, logic)
  - Lack diversity

[1] Zhu, Kaijie, et al. "Dyval: Graph-informed dynamic evaluation of large language models." ICLR 2024.

Prior works on dynamic evaluation

- Template-based methods (e.g., DyVal [1])
  - Limited to specific tasks (math, logic)
  - Lack diversity
- LLM-based perturbation (e.g., DyVal 2 [2], Benchmark Self-Evolving [3])
  - Low controllability
  - Suffer from LLM instability
  - Difficult to verify quality and correctness

[1] Zhu, Kaijie, et al. "Dyval: Graph-informed dynamic evaluation of large language models." ICLR 2024.
[2] Zhu, Kaijie, et al. "Dyval 2: Dynamic evaluation of large language models by meta probing agents." ICML 2024.
[3] Wang, Siyuan, et al. "Benchmark Self-Evolving: A Multi-Agent Framework for Dynamic LLM Evaluation." arXiv 2024.

DARG: Dynamic Evaluation of Large Language Models via Adaptive Reasoning Graph

Key Features:

- Controlled complexity
- Maintained diversity
- Validated labels

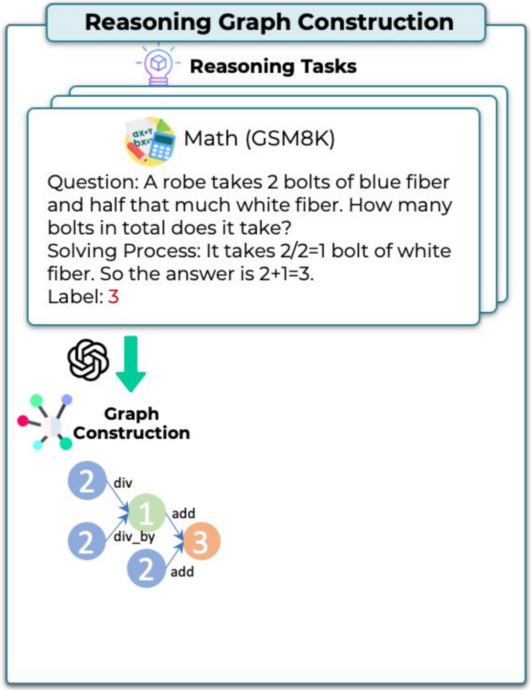
DARG: Dynamic Evaluation of Large Language Models via Adaptive Reasoning Graph

Key Components:

- Reasoning Graph Construction
- Graph Perturbation
- New sample generation
    - Graph-to-text Decoding
    - Data Verification

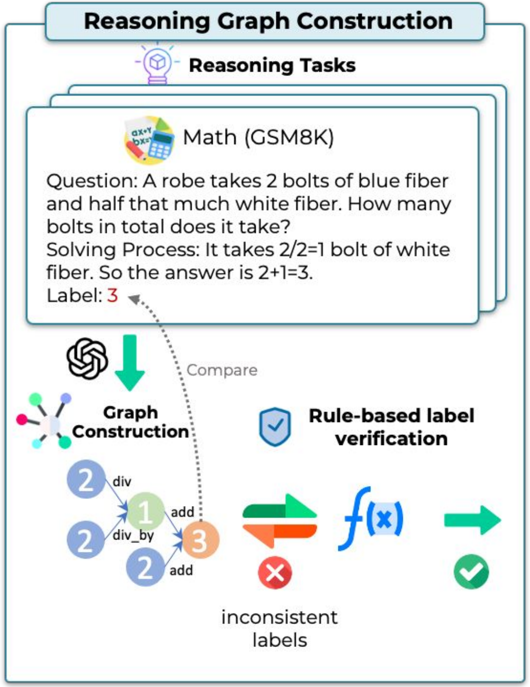
Reasoning Graph Construction

- Extract underlying reasoning structure from benchmark data
    - Use LLMs with in-context learning for graph construction
- Example reasoning graph: The computational graph for a math problem



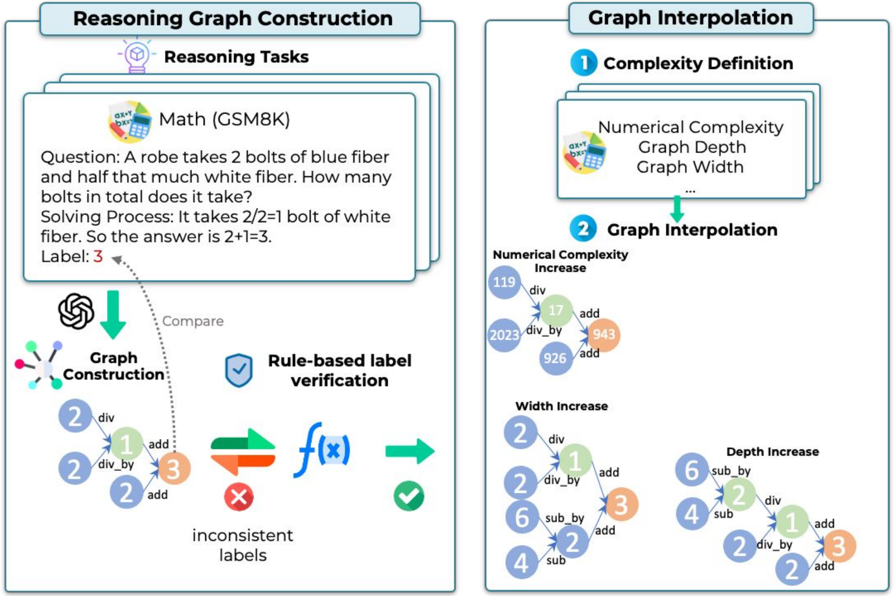
Reasoning Graph Construction

- Extract underlying reasoning structure from benchmark data
    - Use LLMs with in-context learning for graph construction
- Verify graph accuracy using rule-based label computation



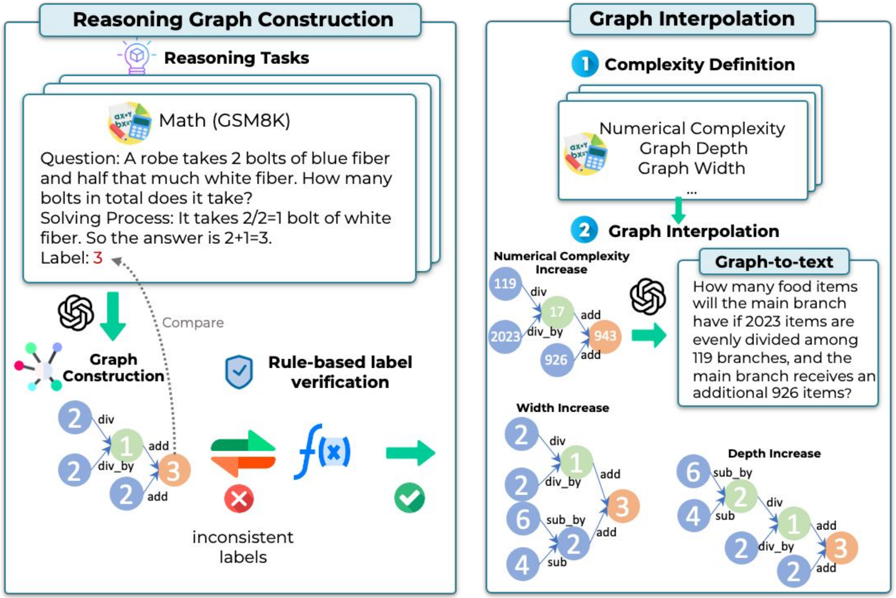
Reasoning Graph Perturbation/Interpolation

- Systematically modify graph structure based on complexity levels
  - Example: for math problem:
    - Numerical complexity (e.g., larger numbers)
    - Graph depth
    - Graph width



New Sample Generation

- Graph-to-text decoding using LLMs through in-context learning
  - Maintain consistent language style with original data (Easy: LLMs are good at style mimicking)
  - Encode reasoning graph structure in generated text (non-trivial, the generated new test sample's reasoning graph may be changed)



New Sample Generation

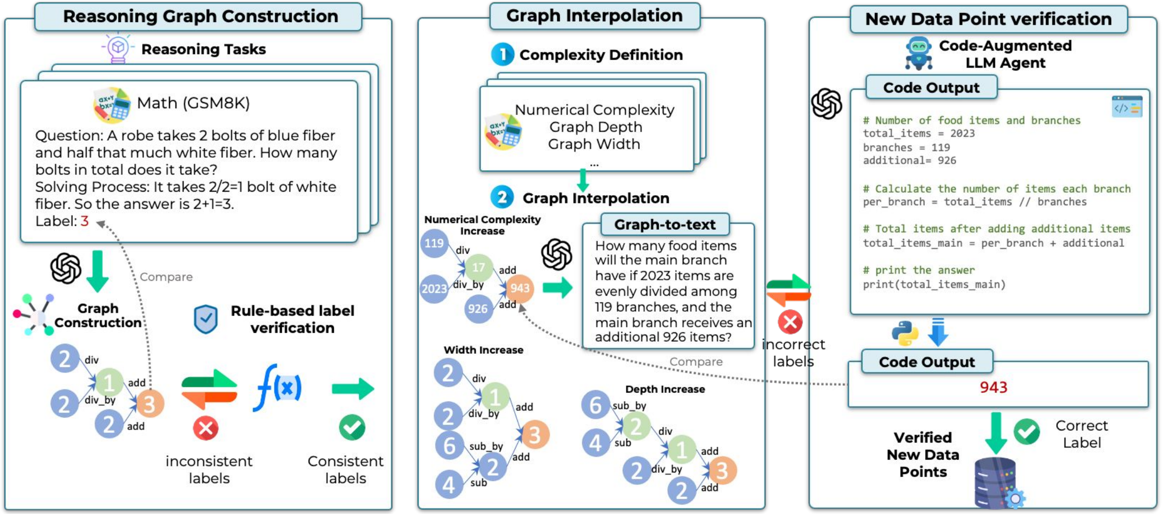
- Graph-to-text decoding using LLMs through in-context learning
    - Maintain consistent language style with original data (Easy: LLMs are good at style mimicking)
    - Encode reasoning graph structure in generated text (non-trivial, the generated new test sample's reasoning graph may be changed)

How to solve this?

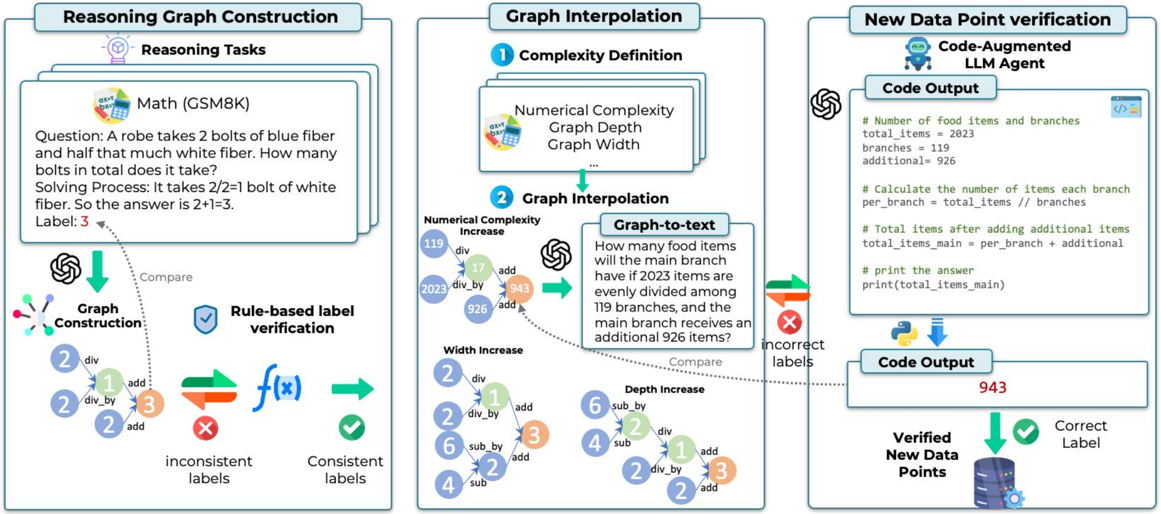
New Sample Generation

- Graph-to-text decoding using LLMs through in-context learning
  - Maintain consistent language style with original data (Easy: LLMs are good at style mimicking)
  - Encode reasoning graph structure in generated text (non-trivial, the generated new test sample's reasoning graph may be changed)

How to solve this?



New Sample Generation



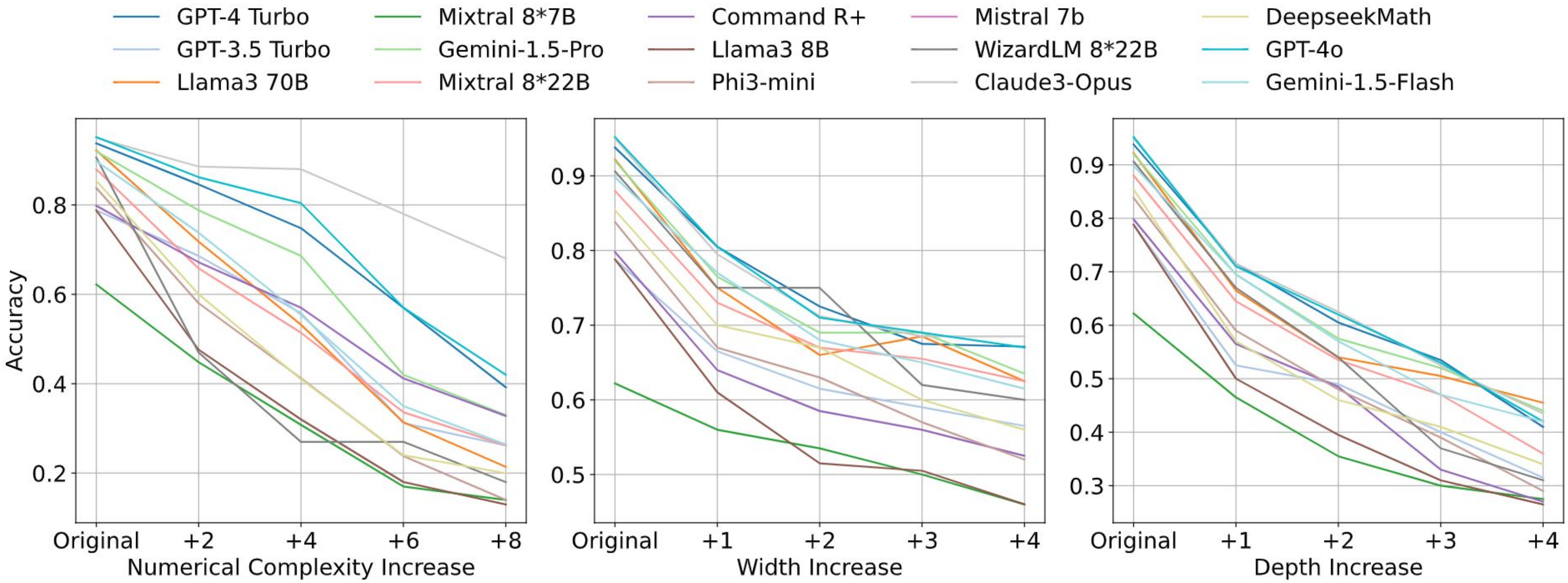
- Graph-to-text decoding using LLMs through in-context learning
  - Code-augmented LLM agent for verification
    - Motivation: SOTA LLMs are good at coding generation and execute code with external interpreter can avoid hallucination
  - Compare computed answers with graph-derived labels
  - Iterative refinement process for incorrect generations

Reasoning Tasks

Domain	Dataset	Node Definition	Edge Definition	Complexity
Math Reasoning	GSM8K [19]	Numbers	$\{+, -, \times, \div, \ldots\}$	# of digits in calculation Width; Depth of calculations
Social Reasoning	BBQ [75]	Persons, Attributes	Relations: 'has'	Attributes' polarity # of attributes involved
Spatial Reasoning	BBH Navigate [91]	Unit action	Sequential order	# of actions
Symbolic Reasoning	BBH Dyck Language [91]	$\{\}, \langle\rangle, [], ()$	Sequential order	# of brackets in the input # of brackets in the label

- The reasoning graph definition in DARG are general and can be applied and extended to other tasks

Math Reasoning (GSM8K)

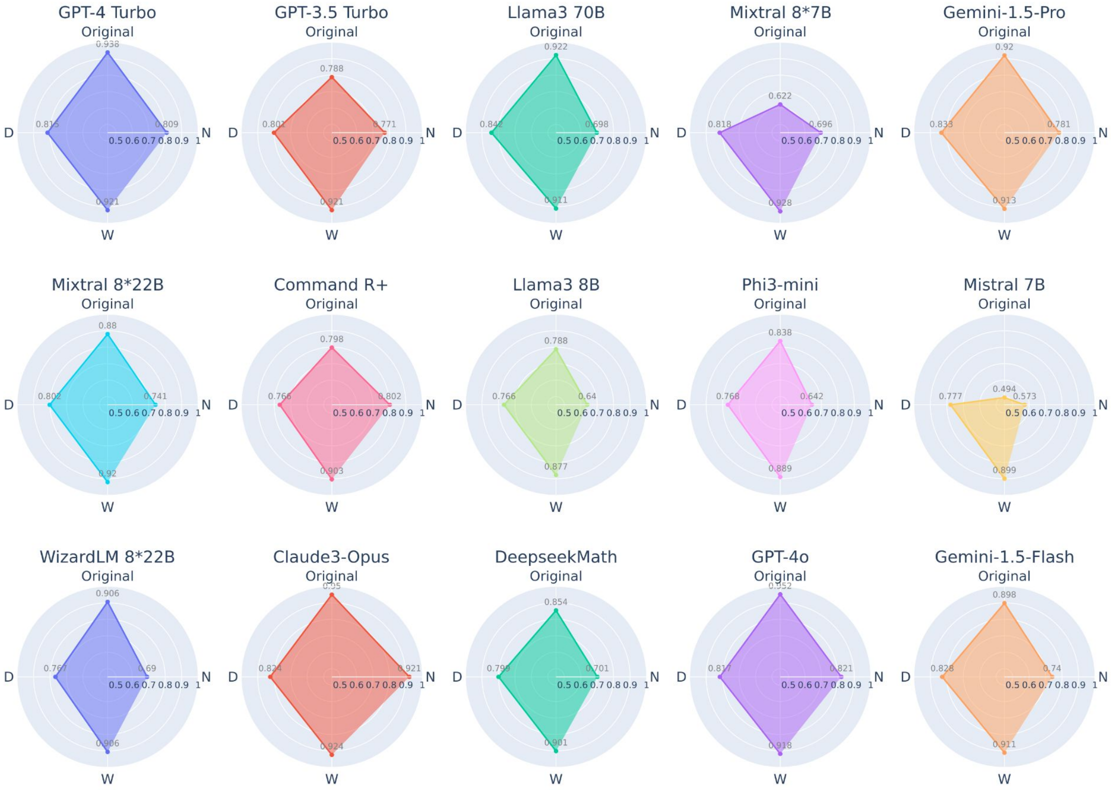


Math Reasoning (GSM8K)

- New Metric:Complexity-Induced Accuracy Retention Rate (CIARR)
  - A higher value indicates greater robustness to complexity increases in that dimension.

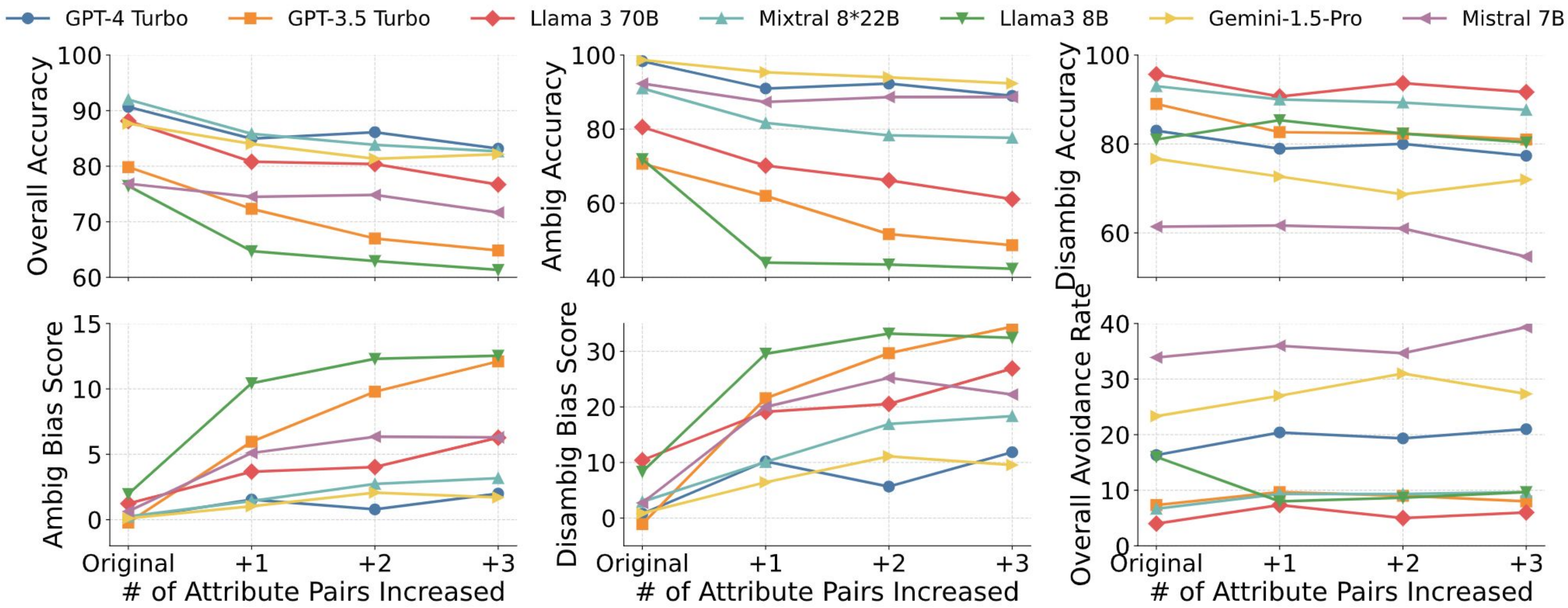
$$\text{CIARR}_D = \frac{1}{n-1}\sum_{i=1}^{n-1}\left(\frac{A_{i+1}}{A_i}\right) \times 100\%$$

Math Reasoning (GSM8K)



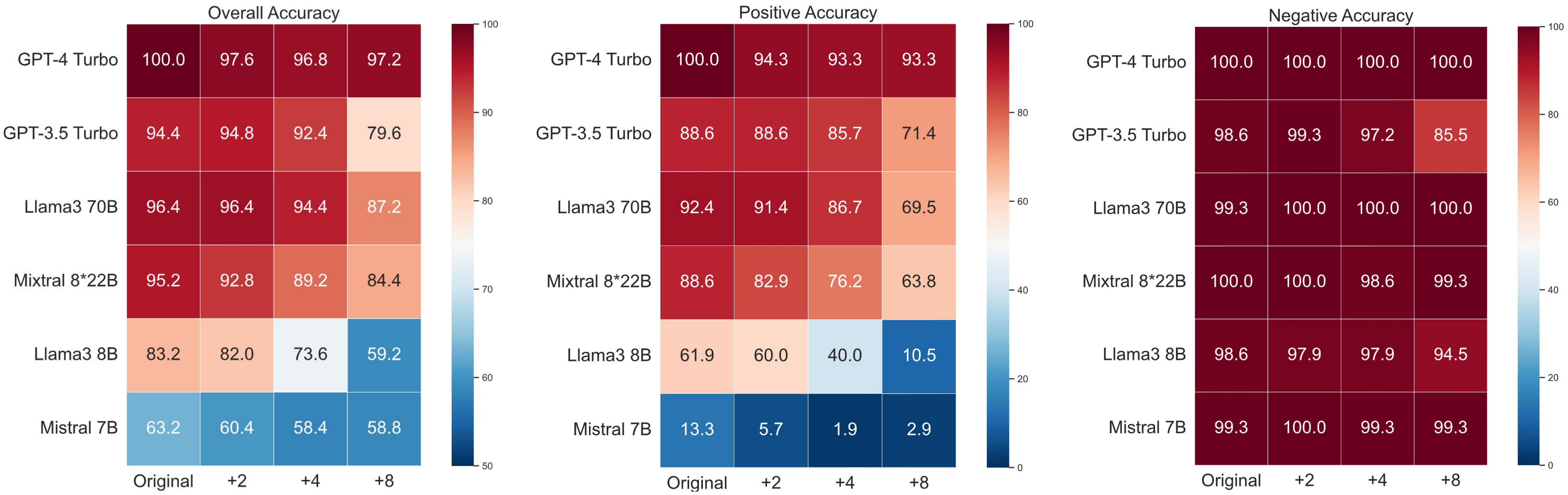
- Larger models and MoE models generally have greater robustness towards complexity increase

Social Reasoning (BBQ)



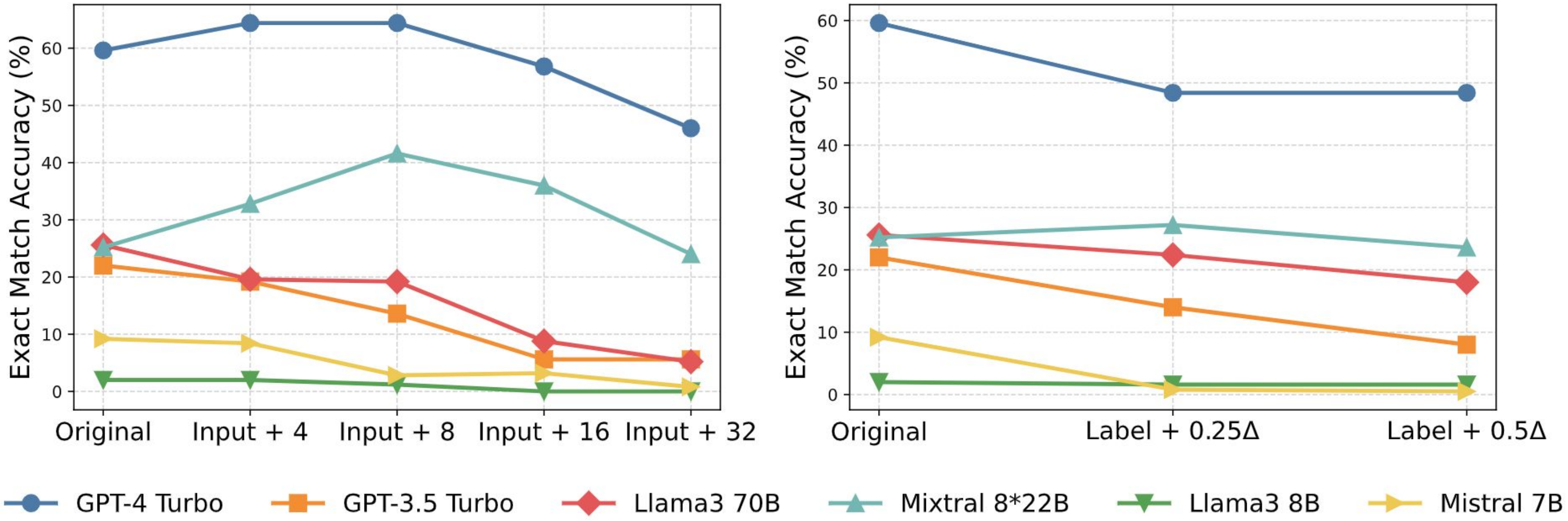
- Key observation: Increased bias with complexity
- Note on over-sensitivity of some models (e.g., GPT-4 Turbo, Gemini-1.5-Pro)

Spatial Reasoning (BBH Navigate)



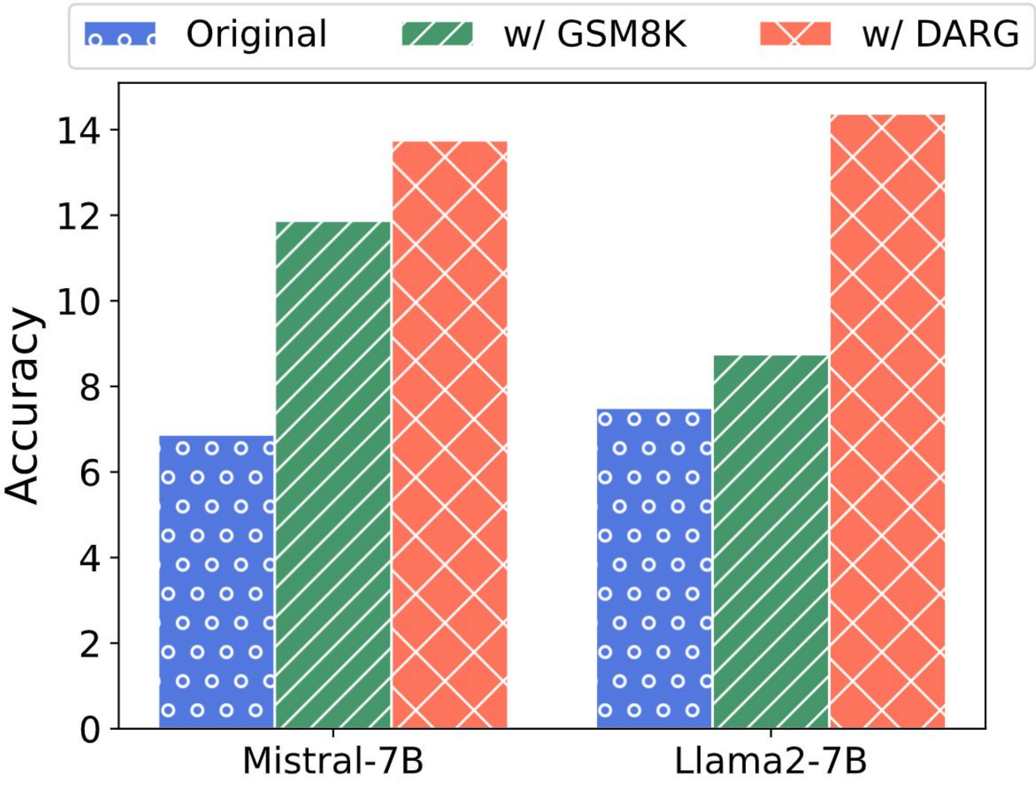
- Highlight: Dramatic decrease in positive accuracy, biases towards generating the negative label

Symbolic Reasoning (BBH Dick Language)



- Highlight: LLMs show performance decrease when the input the expected output length increase

Fine-tuning with DARG



- Comparison between fine-tuning with DARG generated data and the same amount of GSM8K's training data.
- Test on an unseen test set with diverse range of complexity

- Highlight: DARG shows potentials in generating effective training data for LLM improvement

Conclusion

- DARG: A novel framework for dynamic LLM evaluation
- Reveals performance decline and bias increase with complexity
- Demonstrates the need for adaptive evaluation methods
- Potential impact on LLM Improvement and benchmarking practices