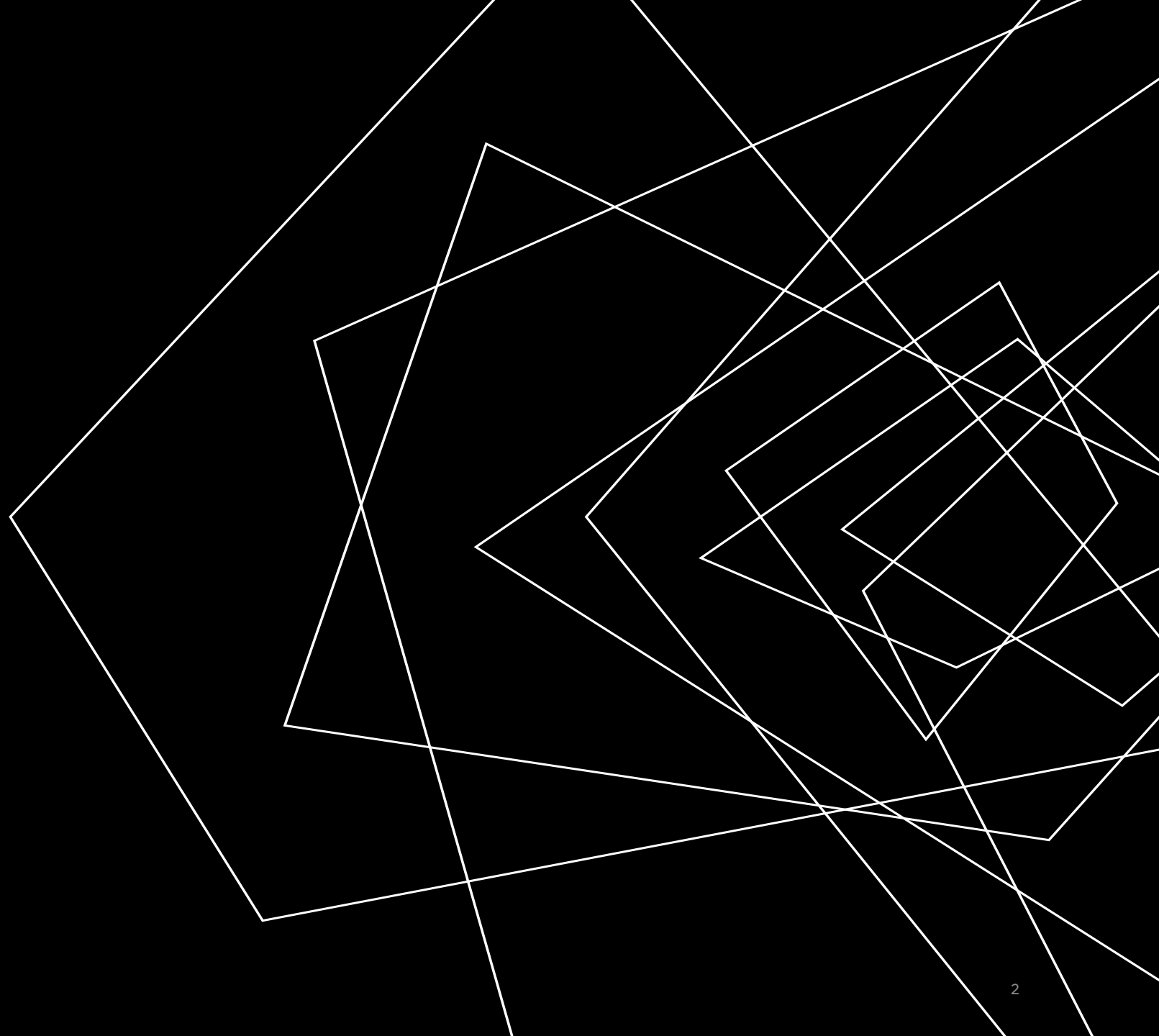
SEMINAR PRESENTATION ON

RELIABLE REASONING BEYOND NATURAL LANGUAGE

By Jagadeesh Reddy Vanga

AGENDA

- Introduction to paper abstract
- Intro to LLMs
- Prompting and CoT
- Novel Approach of Authors
- NLR Dataset Introduction

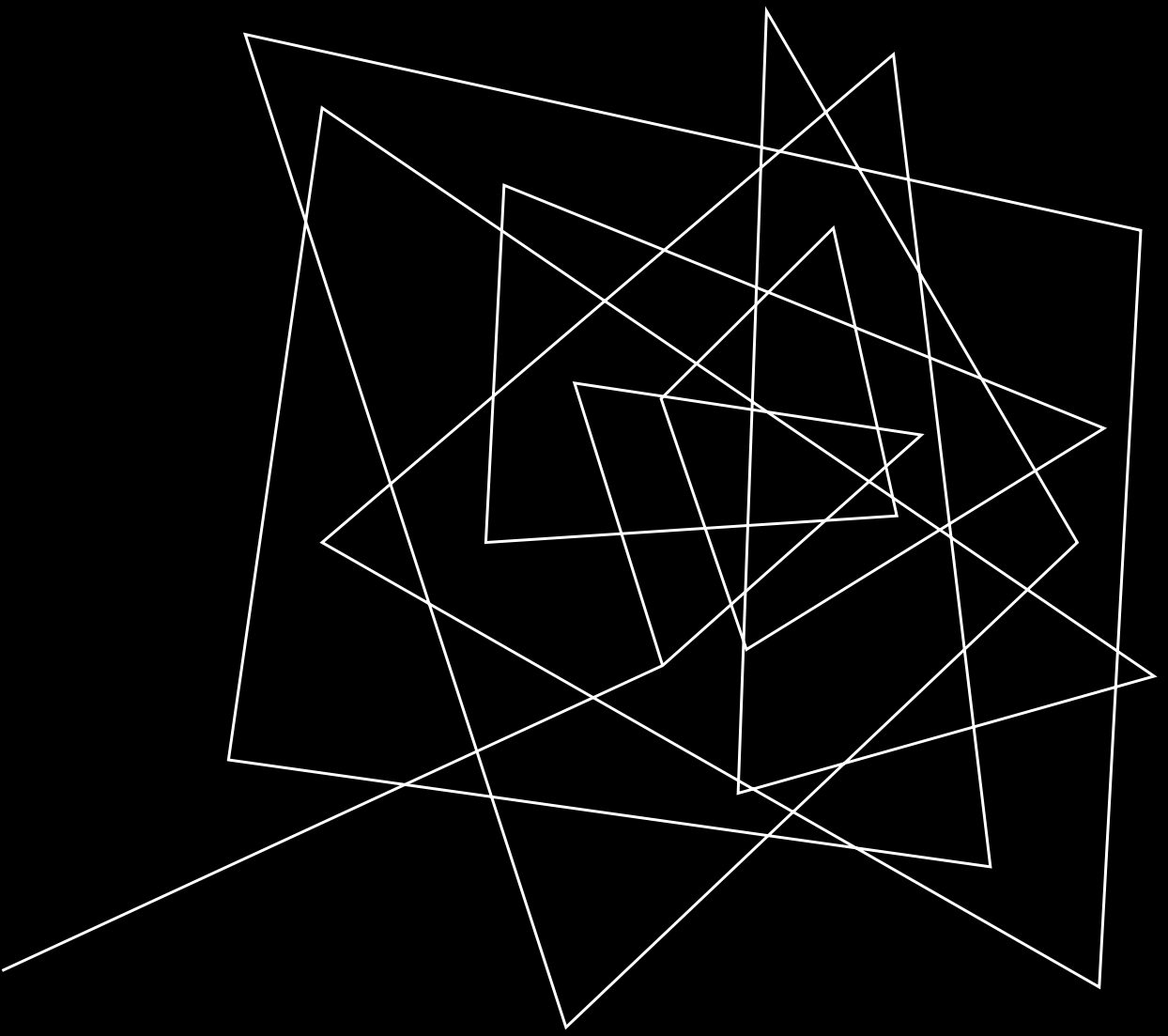


RELIABLE REASONING BEYOND NATURAL LANGUAGE

By Nasim Borazjanizadeh, Steven T. Piantadosi
UC Berkeley

ABSTRACT

- **Problem with LLMs**: LLMs show limitations in reasoning—particularly in flexible, reliable deductive reasoning.
- **Proposed Solution**: Introduced a **NeuroSymbolic** approach using **Prolog** to encode problem statements into logical code.
- **Approach Benefits**: Enhanced performance on reasoning benchmarks like GSM8k and Navigate dataset.
- **New Dataset**: Developed the Non-Linear Reasoning (NLR) dataset, focusing on non-linear reasoning with basic arithmetic.
- **Key Outcome**: Integration with Prolog outperforms traditional LLMs, even advanced models like GPT-4, in solving NLR problems.



LLMs AND TRANSFORMERS

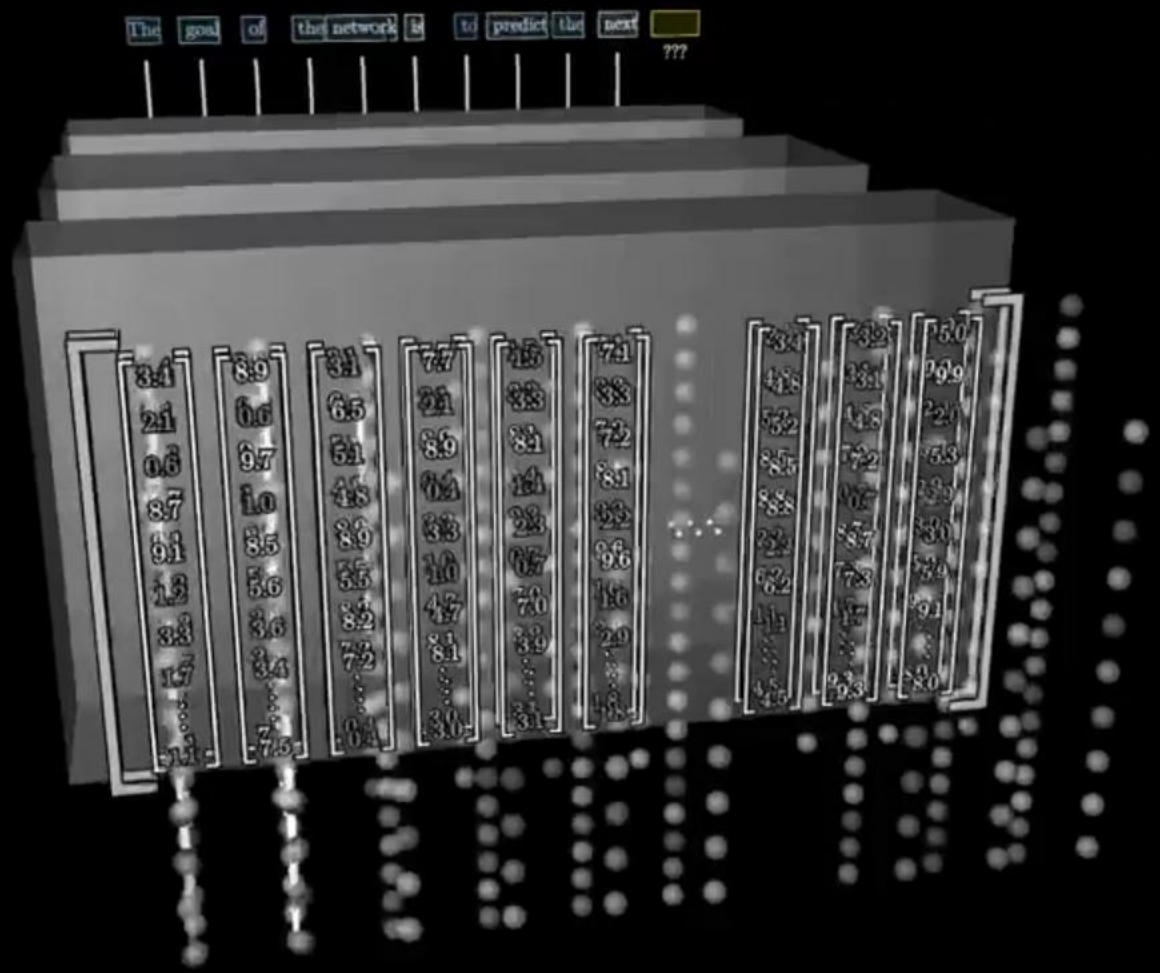
Google Bert

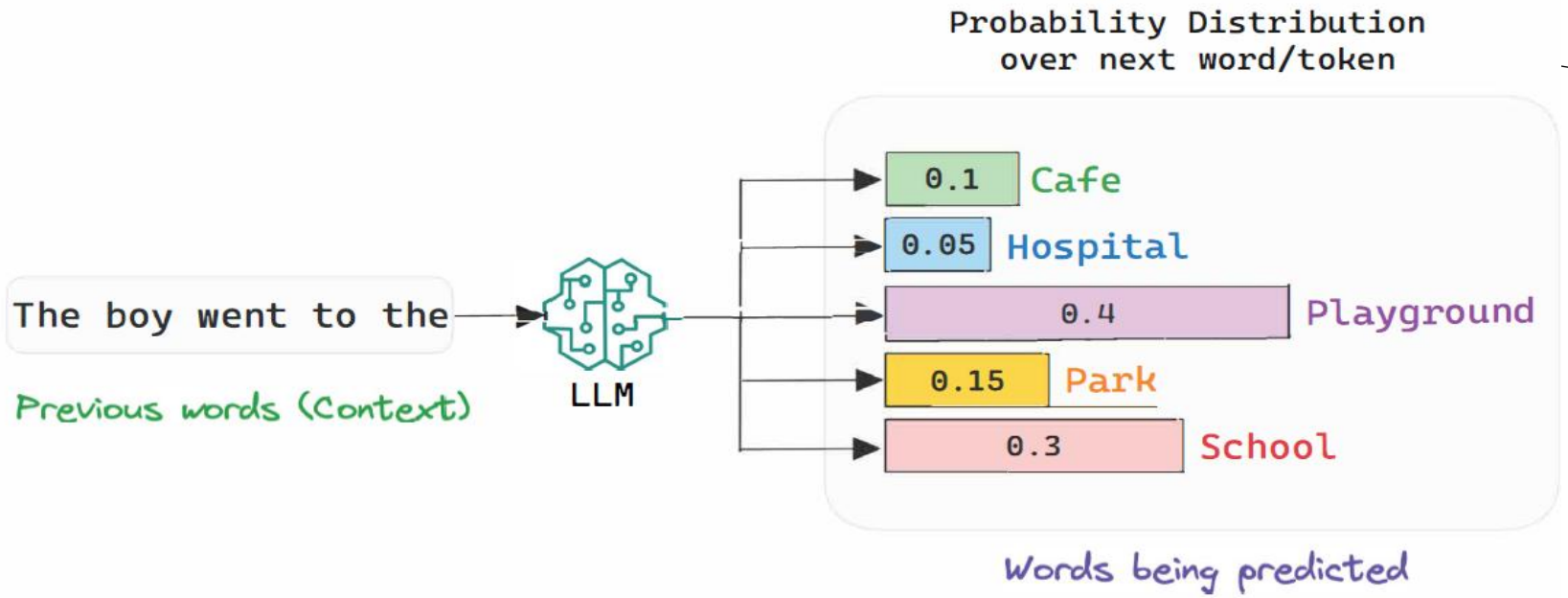
LARGE LANGUAGE MODEL

- **What are LLMs**?: Large Language Models trained on massive datasets for natural language understanding and generation.
- **Built on Transformer Architecture**: Utilizes self-attention mechanisms to process data efficiently.
- **Capabilities of LLMs**: From text generation and translation to summarization and coding assistance.
- **Wide Applications**: Used in chatbots, virtual assistants, content creation, and more.

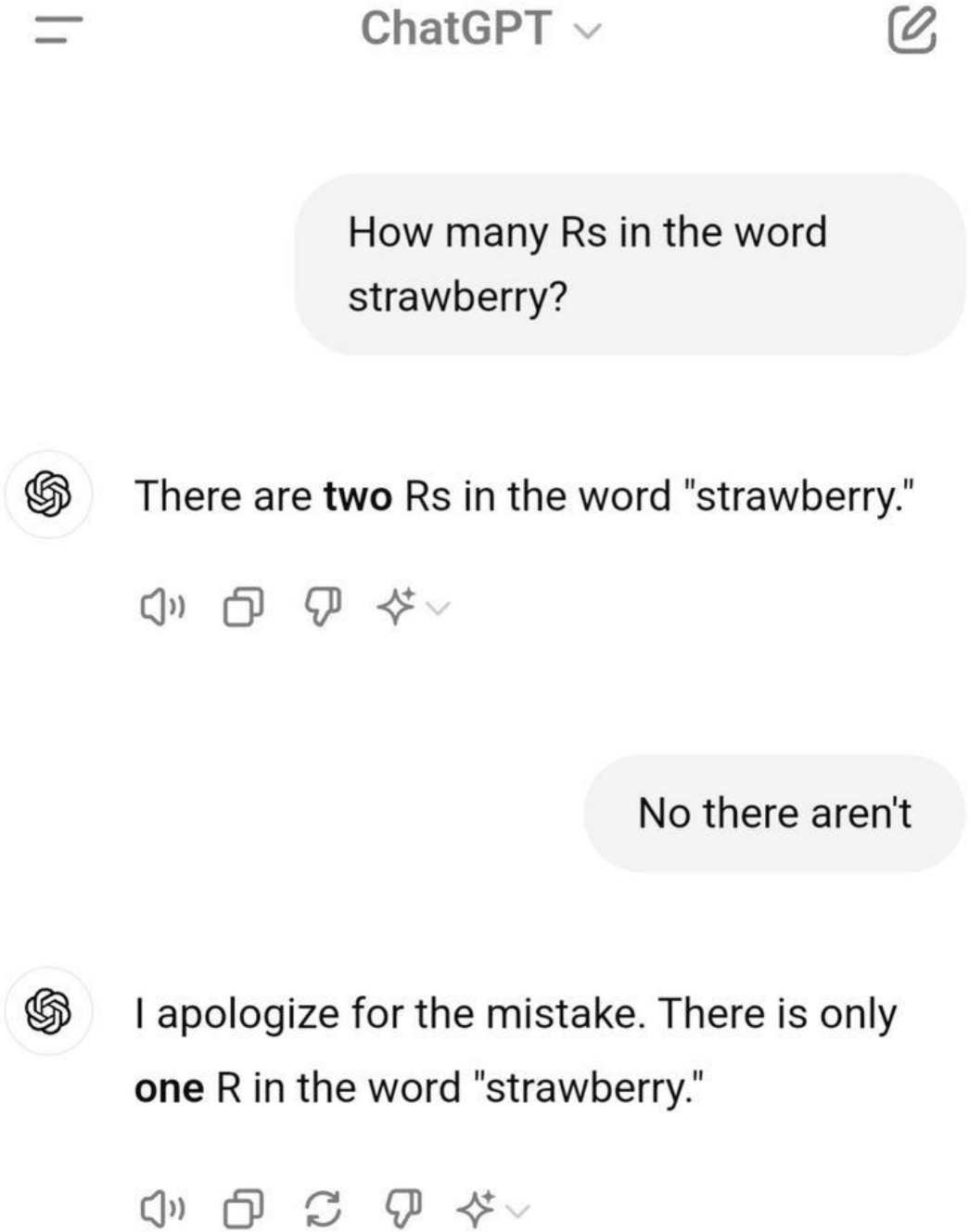
Transformer

Credit: 3Blue1Brown





- **Dialogue Process**: The system prompt is followed by the user's input (question or prompt), which is treated as the first part of the dialogue. GPT-3 then begins to predict responses that a helpful AI assistant would likely give.
- **Fine-Tuning for Quality Responses**: While this setup forms the basic framework, additional steps are required to improve the quality and accuracy of the responses, making the AI more effective as a conversational assistant.



THE PROBLEM

- Limited Reasoning Abilities Due to Sequential Architecture
- Difficulty with Implicit Reasoning
- Mismatch Between Linear Language and Non-linear Reasoning
- Short Working Memory

Any thing we could do to improve these problems?

Like ... **Prompt Engineering**

WHAT IS PROMPT ENGINEERING?

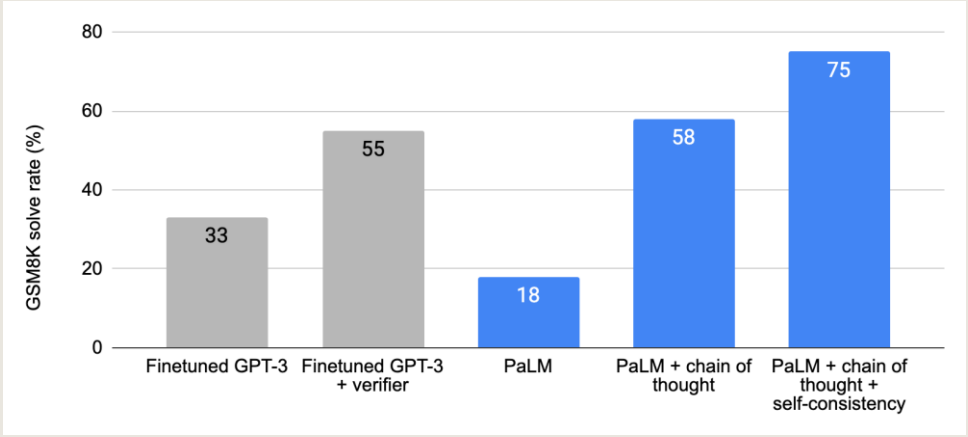
Crafting structured prompts for better AI task interpretation.

- **Purpose of a Prompt**: Instructs the LLM on tasks, includes instructions, context, input, and output indicators.
- **Applications**: Enables LLMs to perform a variety of tasks from Q&A to creative text generation.
- **Key Techniques**: Zero-shot, few-shot, active, and chain-of-thought (CoT) prompting.

CHAIN-OF-THOUGHT PROMPTING

A technique that guides LLMs to break problems into intermediate steps.

- **Improves Reasoning**: Helps LLMs solve complex problems by focusing on one step at a time.
- **Benefits**: Provides interpretable reasoning and is effective for tasks like math problems and commonsense reasoning.
- **Scalable Technique**: Improves performance without extra training—e.g., PaLM model improved from 17.9% to 58.1% on GSM8K benchmark.



PRE-EXISTING RESEARCH ATTEMPTS TO IMPROVE REASONING IN LLMS

- API Calls to External Tools: Used to reduce errors but doesn't fully solve LLM reasoning limitations. [1, 2]
- LINC (Neuro-Symbolic Reasoning): Converts language into formal logic but struggles with implied information. [3]
- Consistency Work by Nye et al.: Cross-checks generated text with a manually crafted world model, limiting its general applicability. [4]
- Program of Thought (PoT): Disentangles computation from reasoning but often lacks deeper reasoning depth.[5]

PROPOSED METHOD: CHAIN-OF-THOUGHT + PROLOG

Solution: Combine LLMs with Prolog, a logic programming language, to leverage both natural language understanding and deductive reasoning.

Method:

1. Prompt LLMs to translate problems into Prolog code.
2. Use Chain of Thought (CoT) reasoning in text and code to guide translation and capture implicit information.
3. Employ a Multiple-Try inference algorithm for robust code generation.

Benefits:

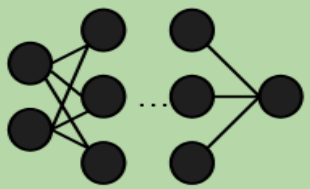
1. Significantly improved accuracy on challenging logical reasoning tasks.
2. Robustness against errors and ability to handle complex problems.

Problem: There are 4 brothers, John, Blake, Sam, and Frank. In three years, Sam will be twice as old as John, and Blake's age will be equal to the age difference between Sam and Frank. In eight years, John's age would be twice Blake's age, and John would be 2 years younger than Frank's current age. Also, the sum of Sam's and Blake's ages is 5 years less than the sum of John's and Frank's ages. What is the sum of the 4 brother's ages right now?

Input

Prolog prompt with chain of thought reasoning in text and code

LLM



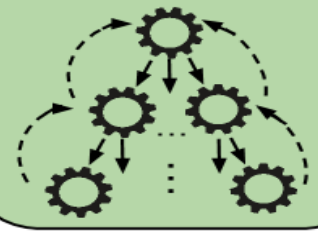
```
problem(Sum_4_brothers_ages):-
{% In three years, Sam will be twice as old as John
Sam_current_age + 3 = 2 * (John_current_age + 3),
% In three years, Blake's age will be equal to the
age difference between Sam and Frank
Blake_current_age + 3 = abs(Sam_current_age + 3
- (Frank_current_age + 3)),
...
```

LLM Generation 1

```
problem(Sum_of_brothers_ages):-
{% In three years, Sam will be twice as old as John
S + 3 = 2 * (J + 3),
% Blake's age will be equal to the age difference
between Sam and Frank
B = S - F,
% In eight years, John's age would be twice Blake's age
J + 8 = 2 * (B + 8),
% John would be 2 years younger than Frank's current
age
J = F - 2, ...
```

LLM Generation 2

Prolog Interpreter



Code executed

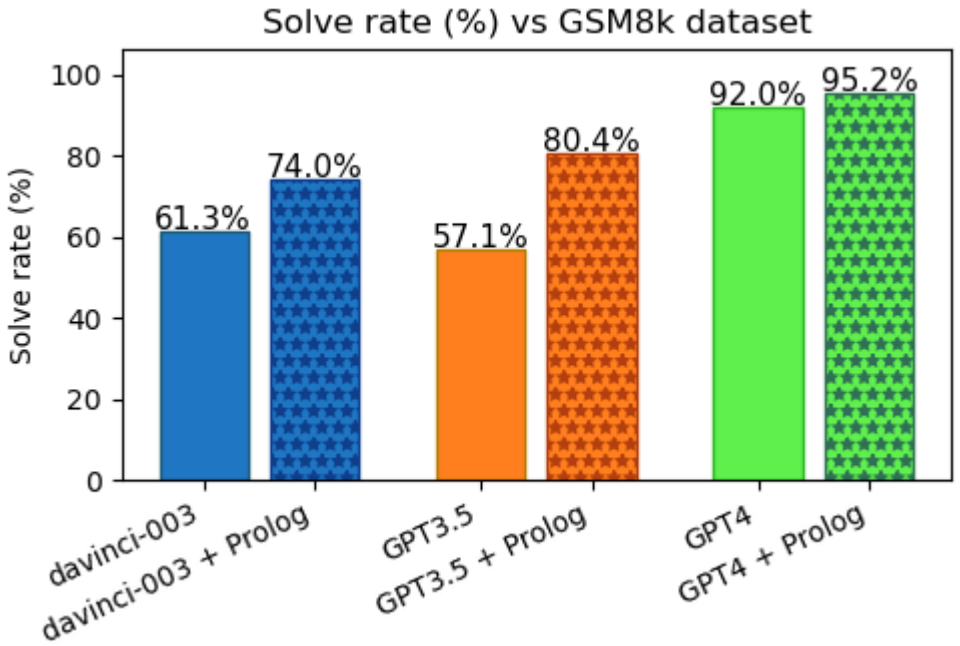
63

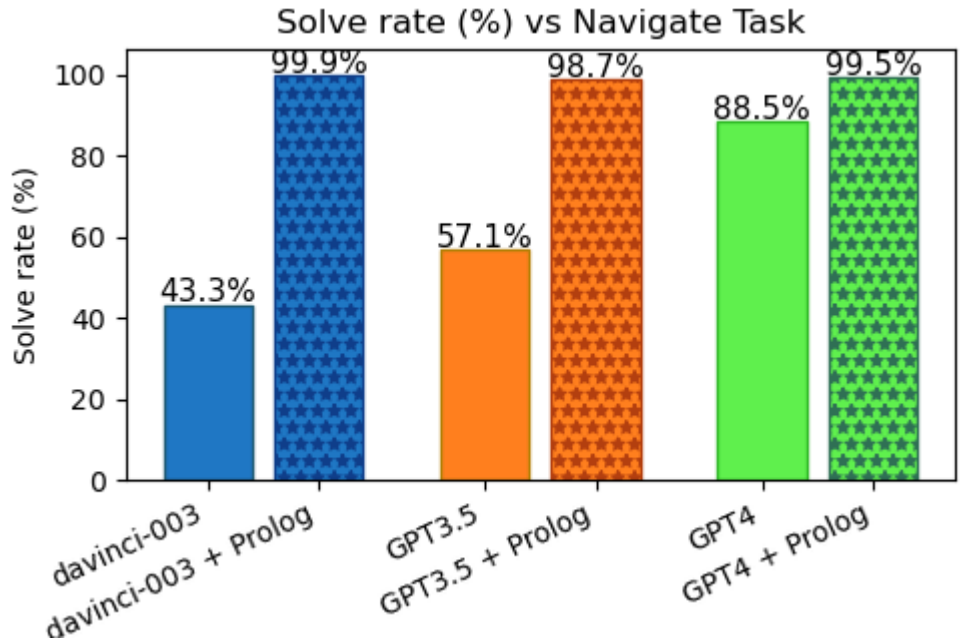
Output

Code was not executed

Multiple-Try Inference Algorithm

HOW DID IT WORK?





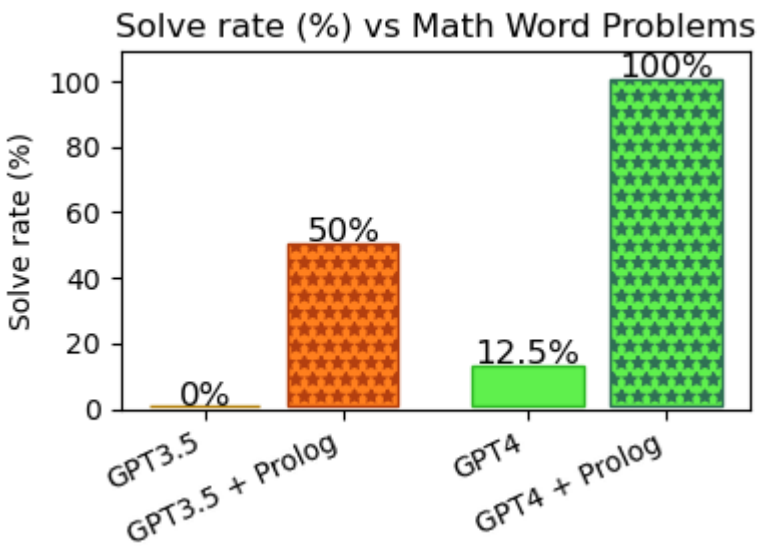
NON-LINEAR REASONING DATASET

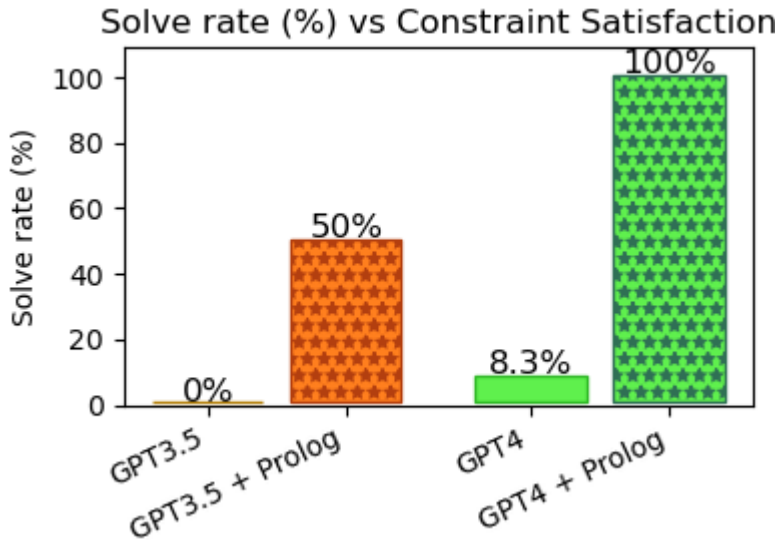
- A benchmark with 55 unique problems for evaluating LLM reasoning.
- **Out-of-Distribution Generalizability**: Problems designed to test true reasoning abilities beyond memorization.
- **Focus on Non-Linear Reasoning**: Emphasizes complex interrelationships and iterative problem-solving.
- **Diversity of Problem Categories**: Includes math word problems, constraint satisfaction, and algorithmic instructions.
- **Minimal Arithmetic Complexity**: Focus on logical reasoning rather than complex calculations.
- **Varying Entanglement Dataset**: A subset containing varying number of interconnected variables

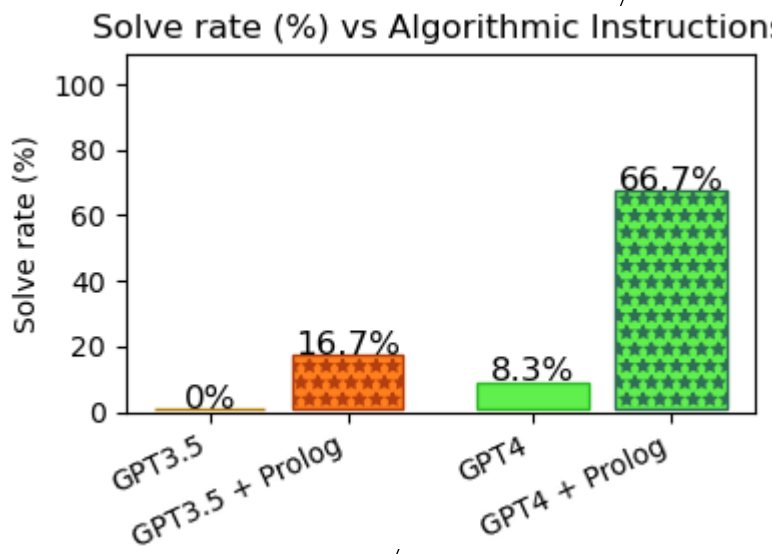
Dataset, Varying Entanglement Dataset

NLR Problem Statement	Characteristics
Math Word Problem: When I was half my current age, my father was 30. When I was 1/3 my current age, my mother was 25. And when I was 1/6 of my current age, my sister was 7. If the sum of my age, my sister's age, my father's age, and my mother's age is 116, then how old am I now?	4 entangled variables in the problem
Constraint Satisfaction: In a line to enter a cinema, 4 people are standing between Bob and Alex. Chad's index in the line is 1 after Bob's, he's standing right behind Bob considering the order of people left to right. Frank is right behind Alex. Sam is right in front of Bob. There are 2 people between Sam and Frank. If Bob is in the 7th person in the line, counting left to right, what is the number of Alex?	2 constraints encoding multiple possibilities
Algorithmic Instructions: There's a cinema with 12 seats organized in 3 rows and 4 columns. Due to covid there's a policy that a seat can be filled only if none of the seats right next to it in the same column or the same row are not filled. If we place a person in the seat in the second column of the first row and then start to fill the seats left to right, row by row, starting row with 1, how many people can be seated in the cinema in total?	5 entangled variables in each state

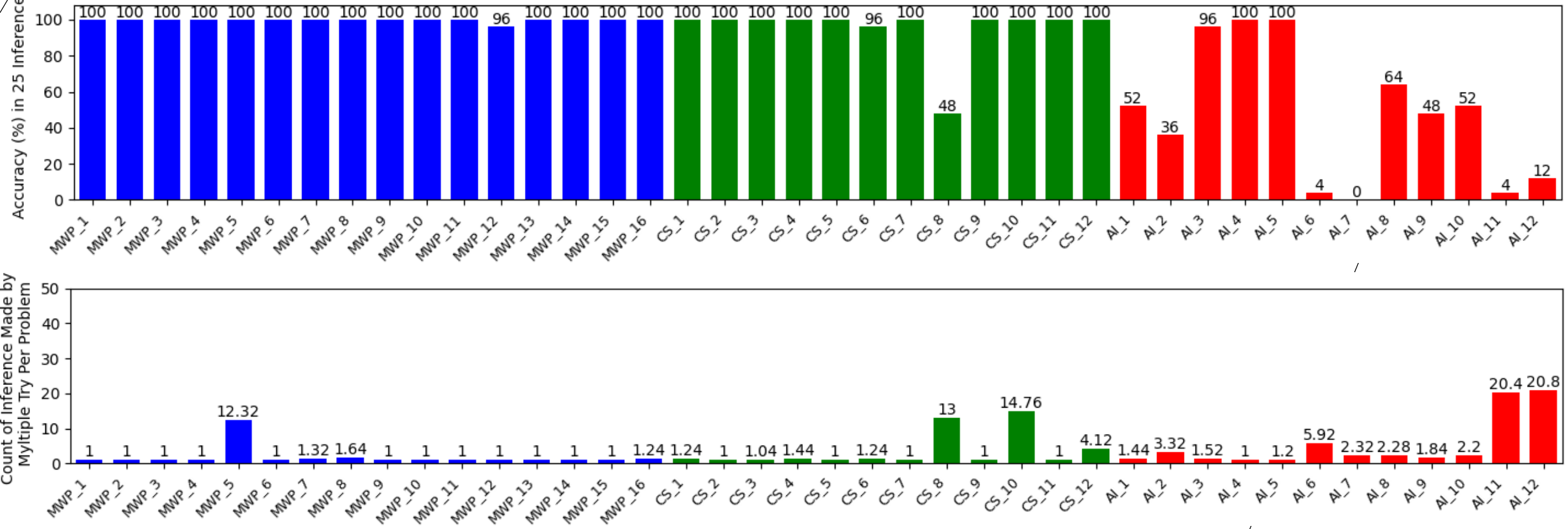
PERFORMANCE ON NLR DATASET





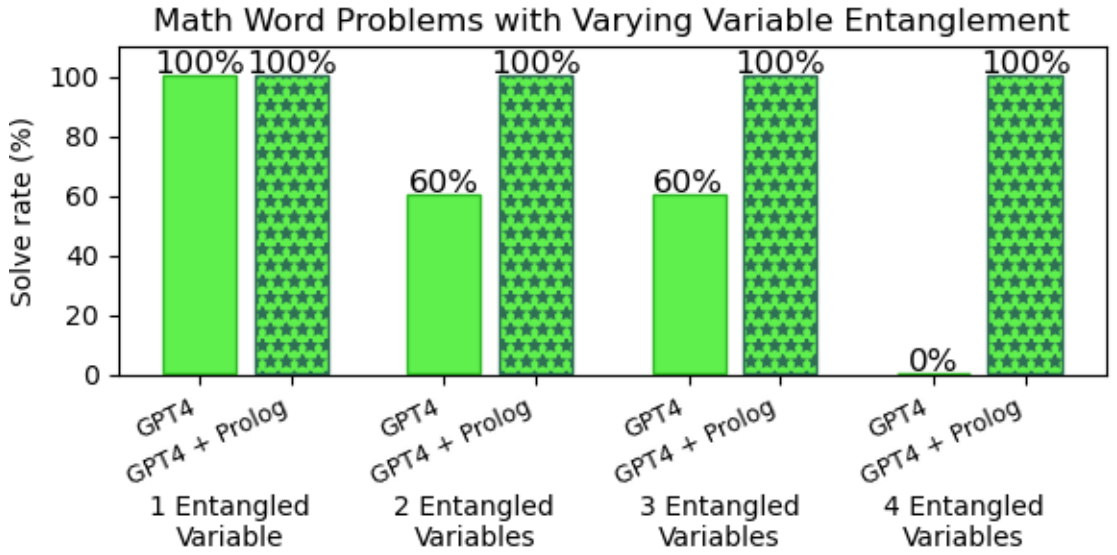


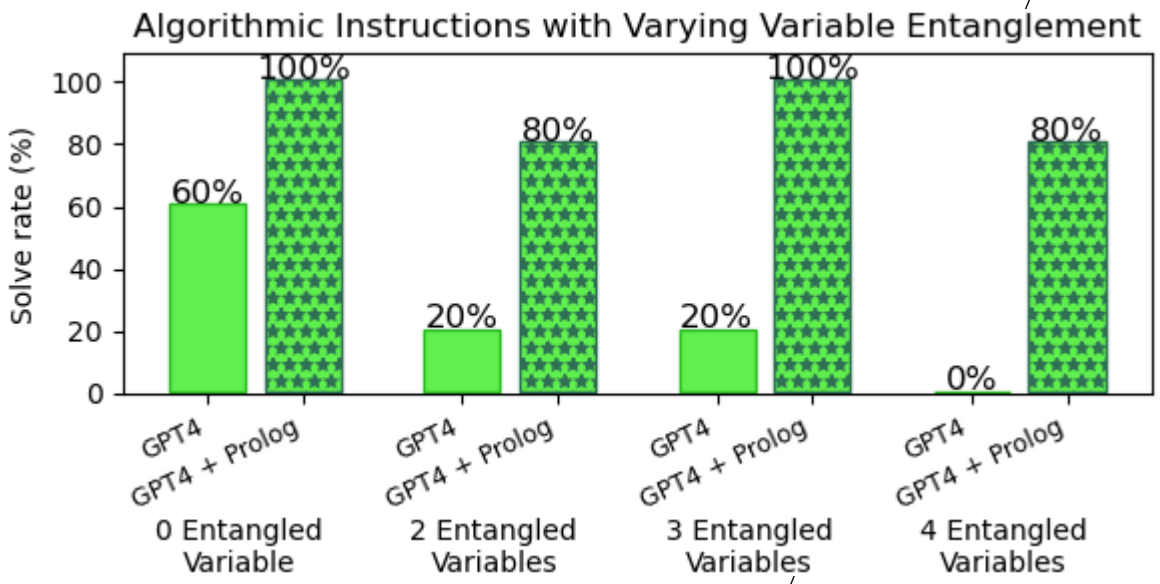
ROBUSTNESS & PERFORMANCE VARIABILITY



Math Word Problems are abbreviated as ’MWP’, Constraint Satisfaction as ’CS’, and Algorithmic Instruction as ’AI’.

PERFORMANCE ON VAR-ENTANGLED DATASET





LIMITATIONS AND BROADER IMPACT

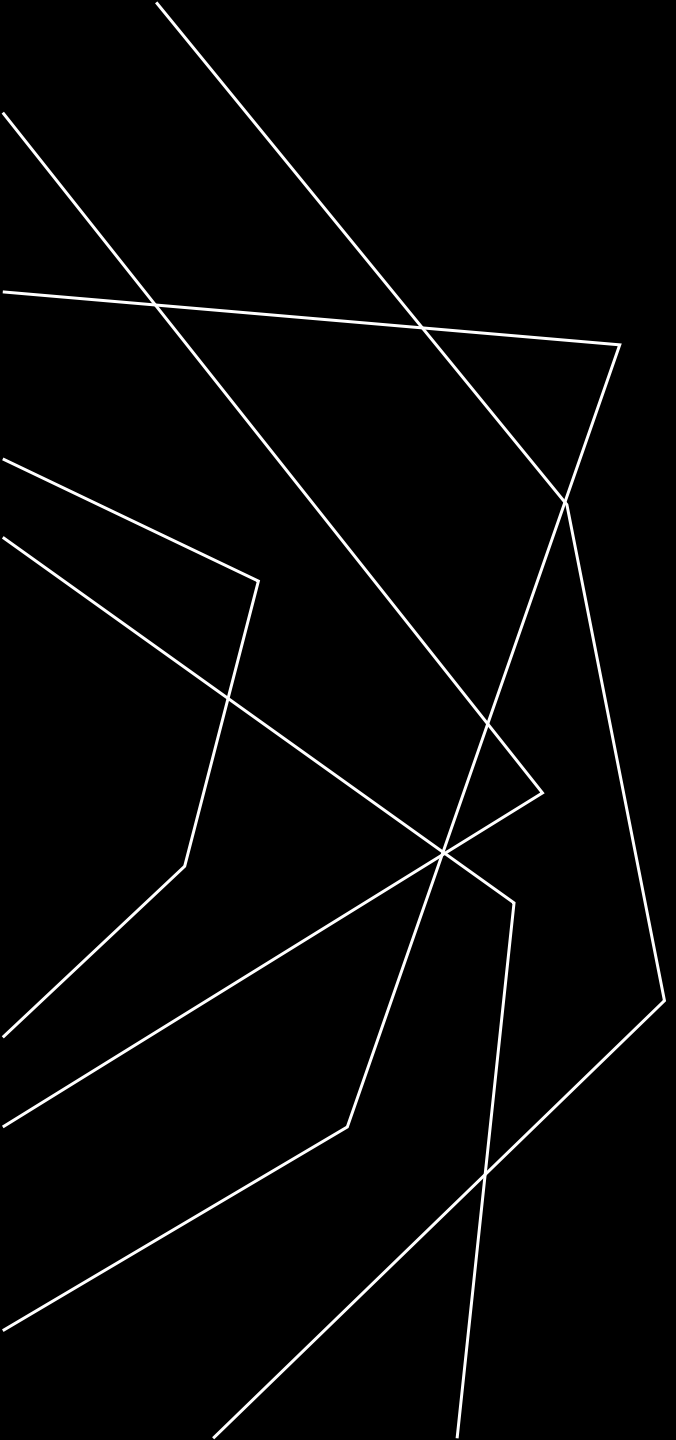
Positive Societal Impacts:

- Improved reasoning capabilities can automate complex tasks, leading to efficiency gains but also potential job displacement.

Limitations of the NLR Dataset:

- **Scalability**: Designing unique reasoning problems is time-consuming and complex.
- **Error Detection**: LLMs may generate coherent but incorrect solutions.
- **Prolog Limitations**: Prolog's limited support for complex data structures can hinder its use in problems involving higher-dimensional data.

QUESTIONS?



THANK YOU