LLM as a judge and Synthetic Dataset

Norapat Buppodom: TA Pangpuryiye;
AI Researcher, Menlo Research

SS4 LLM Hackathon

No Train Data!

Eval Data 6 examples!

Human Evaluation

LLM as a judge and synthetic dataset

Why use LLM as a judge

Automatic Evaluation

Classification

Named-entity recognition

Multiple Choice Exam

Translation

Summarization

Information Extraction (ex. Hack 2 Crime Charge)

Manual Evaluation

Text Generation

Chatbot

Agentic

RAG

Summarization/Deep Research

Role-playing

Why use LLM as a judge

```
[System]
Please act as an impartial judge and evaluate the quality of the response provided by an
AI assistant to the user question displayed below. Your evaluation should consider factors
such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of
the response. Begin your evaluation by providing a short explanation. Be as objective as
possible. After providing your explanation, please rate the response on a scale of 1 to 10
by strictly following this format: "[[rating]]", for example: "Rating: [[5]]".

[Question]
{question}

[The Start of Assistant's Answer]
{answer}
[The End of Assistant's Answer]
```

It is easy!

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena:
NeurIPS 2023 Dataset and Benchmark Track

Why use LLM as a judge

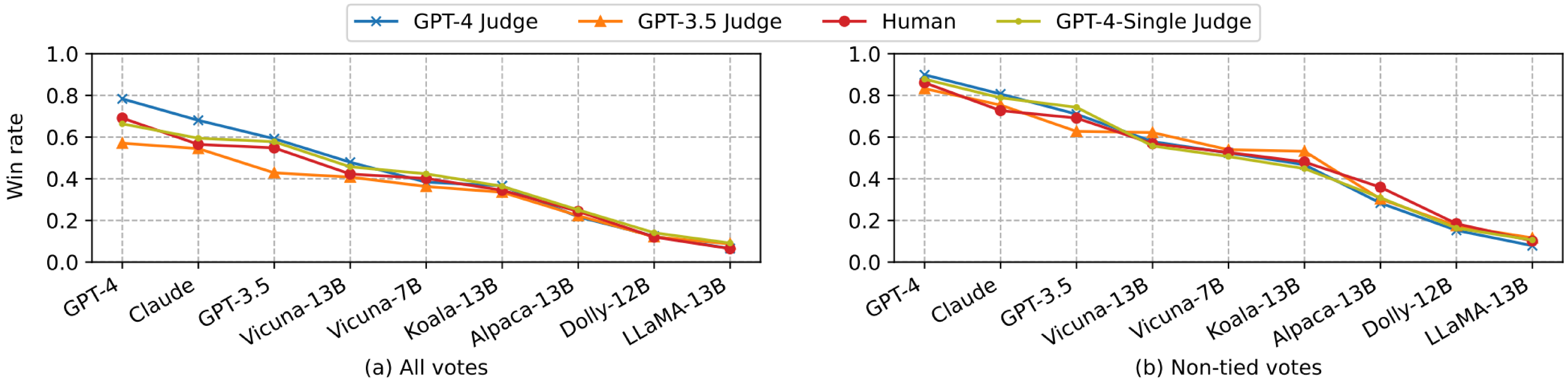


Figure 4: Average win rate of nine models under different judges on Chatbot Arena.

It is highly correlated with human eval!

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena:

NeurIPS 2023 Dataset and Benchmark Track

How to do LLM as a judge

1) Pointwise Scoring	Output	“score 4/5"	**“Most Scalability”**
2) Pairwise Scoring	Output A Output B	“A is better”	**“Most aligned with humans“** Can be used with DPO training
3) Reference based Scoring	Output Reference	“score 4/5"	**“Good with Reasoning Task”**

How to do LLM as a judge

Example Prompt

Human-like Summarization Evaluation with ChatGPT

1) Pointwise Scoring

Evaluate the quality of summaries written for a news article. Rate each summary on four dimensions: {Dimension_1}, {Dimension_2}, {Dimension_3}, and {Dimension_4}. You should rate on a scale from 1 (worst) to 5 (best).

Article: {Article}
Summary: {Summary}

Figure 1: The template for Likert scale scoring.

2) Pairwise Scoring

Given a new article, which summary is better? Answer "Summary 0" or "Summary 1". You do not need to explain the reason.

Article: {Article}
Summary 0: {Summary_0}
Summary 1: {Summary_1}

2 Options

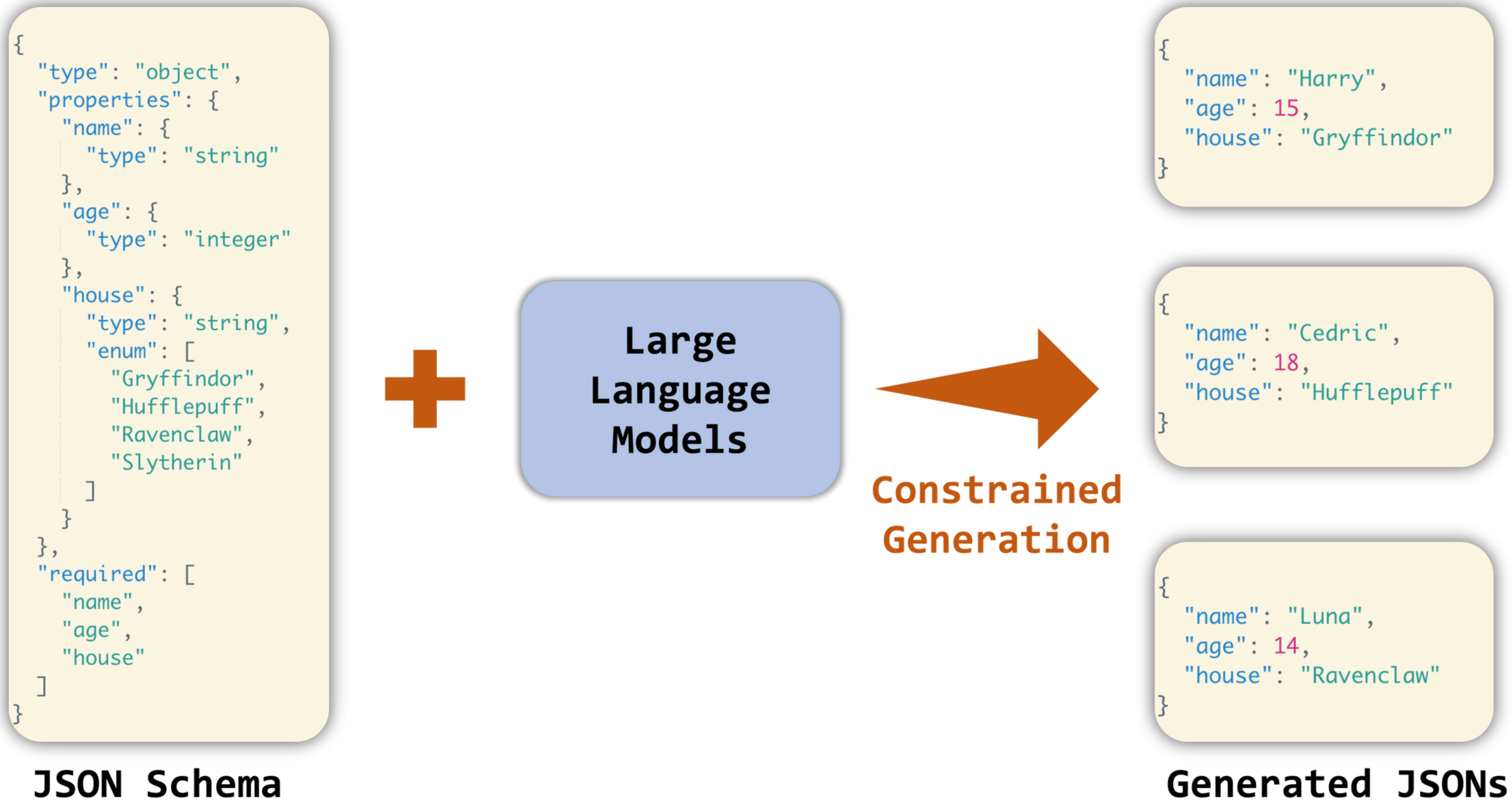
- Win, Loss
- Win, Loss, Tie

1) and 2) can be used with reference based Scoring:

Just add reference into prompt

How to do LLM as a judge

Constraint Decoding



https://lmsys.org/blog/2024-02-05-compressed-fsm/

Bias in LLM as a judge

Position Bias

Judge	Prompt	Consistency	Biased toward first	Biased toward second
Claude-v1	default	23.8%	**75.0%**	0.0%
	rename	56.2%	11.2%	**28.7%**
GPT-3.5	default	46.2%	**50.0%**	1.2%
	rename	51.2%	38.8%	6.2%
GPT-4	default	**65.0%**	30.0%	5.0%
	rename	**66.2%**	28.7%	5.0%

please compare
<A Response>
<B Response>

LLM Judge

A Wins

please compare
<B Response>
<A Response>

LLM Judge

B Wins

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena
NeurIPS 2023 Dataset and Benchmark Track

Bias in LLM as a judge

How to fix position bias?

1. **Random Position of A/B**
2. **Compare twice, both position**

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena:
NeurIPS 2023 Dataset and Benchmark Track

Bias in LLM as a judge

Response Length Bias

Q: Super AI SS5 จัดที่ไหน

A: Super AI SS5 **จัดที่ The pine อาหารอร่อย**

B: **จัดที่ Thepine อาหารอร่อย**

“A Wins”

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena:
NeurIPS 2023 Dataset and Benchmark Track

Bias in LLM as a judge

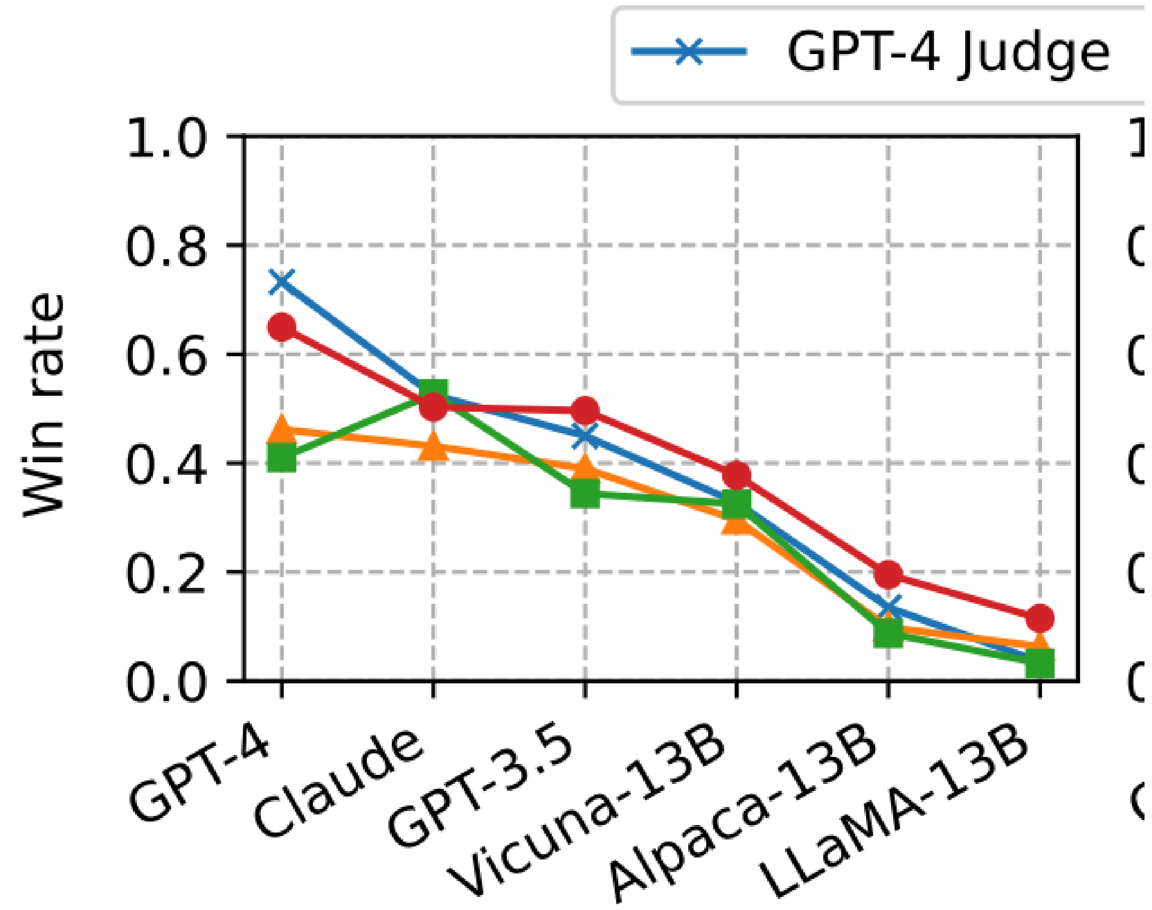
How to fix response length bias

1. **Calibration by length**
2. **Calibration by score/length**

Note: Human also have response length bias

Bias in LLM as a judge

Self-consistency bias



(a) All votes, first turn

Pathumma
Typhoon
→ **Typhoon** → **"Typhoon Wins"**

Pathumma
Typhoon
→ **Pathumma** → **"Pathumma Wins"**

<u>Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena</u>: NeurIPS 2023 Dataset and Benchmark Track

Eval LLM as a judge

Agreement with human

GPT-4 Judge

Human Judge (Random)

Example (Random)

Agreement wit human = probability that two judge has the same result

Example of eval frameworks

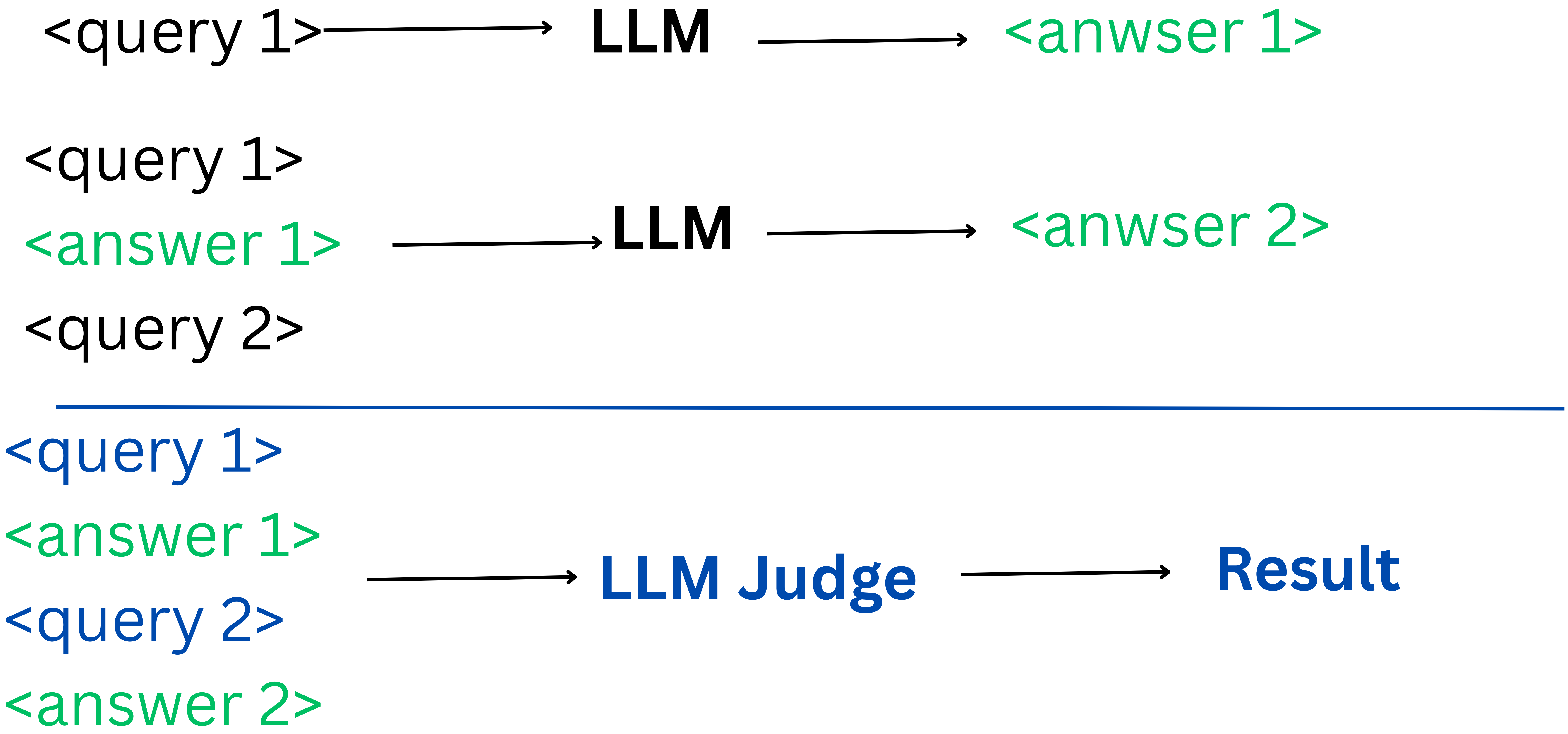
MT-Bench: general LLM as a judge eval

General Domain - **Pairwise eval**
Math Domain - **Reference eval**

Questions: **80 questions, 2 turns**
Topic: writing, roleplay, extraction, reasoning, math, coding, 3 knowledge I (STEM), and knowledge II (humanities/social science).

Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena:
NeurIPS 2023 Dataset and Benchmark Track

MT-Bench: general LLM as a judge eval



Eval Summarization: reference free

Human-like Summarization Evaluation with ChatGPT

Original Text

Summarization

- **consistency:** fact check with original text
- **relevancy:** compare content in summary with source

- **fluency:** check each sentence
- **coherence:** check story telling

scale 0-5

Eval Summarization: reference free

Human-like Summarization Evaluation with ChatGPT

consistency: fact check with original text

Original Text

Summarization

Is the sentence supported by the article? Answer "Yes" or "No".

Article: {Article}
Sentence: {Sentence}

- fact sentence 1 ✅
- fact sentence 2 ✅
- fact sentence 3 ❌

score = 2/3

Eval Open-ended Q/A with context

Ragas: Automated Evaluation of Retrieval Augmented Generation: EACL 2024 demo

Source

Question

LLM

Answer

- **factuality:** is **answer** mentioned in the **source**

- **answer relevancy:** is **answer** related to the **question**

- **fluency:** check each sentence
- **coherence:** check story telling

Eval RAG: reference free

Ragas: Automated Evaluation of Retrieval Augmented Generation: EACL 2024 demo

Can be automatic eval: ex. precision

- **context relevancy:** is **retrieved documents** related to the **question**

Retrieved Documents ← Search Engine ← Question

↓

LLM

↓

Answer

- **factuality:** is **answer** mentioned in the **source**

- **answer relevancy:** is **answer** related to the **question**

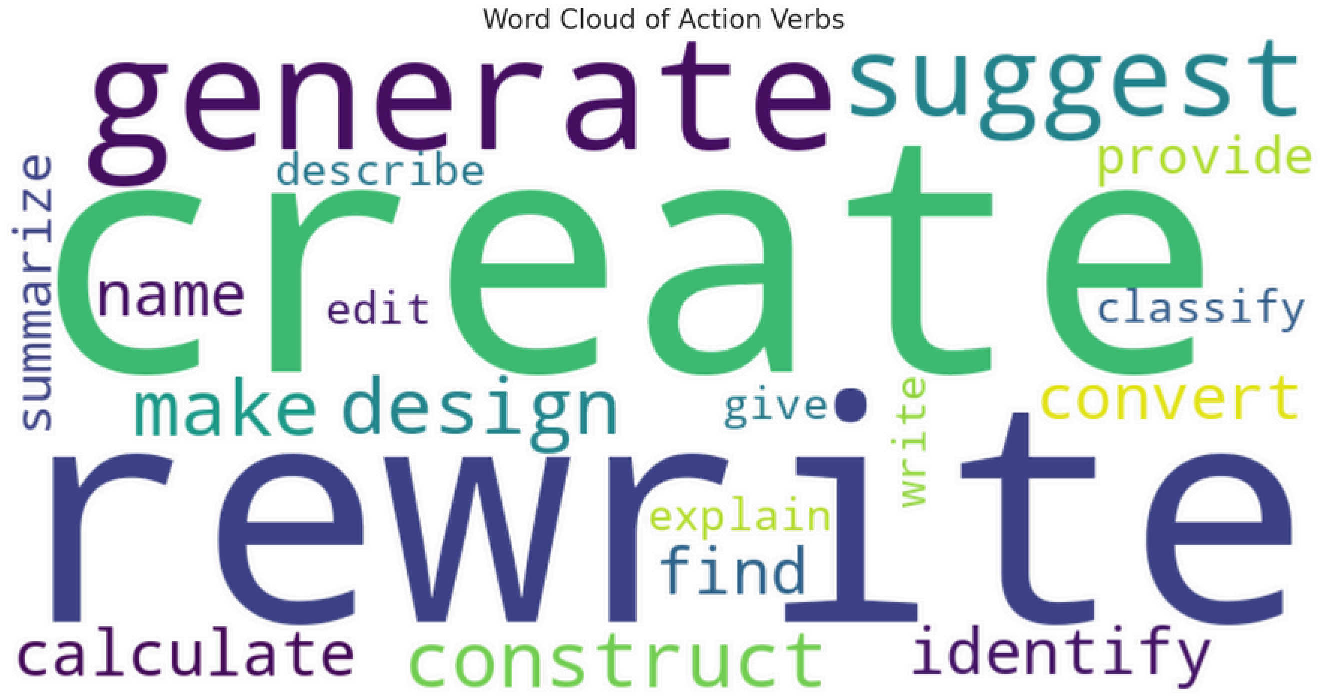
- **fluency:** check each sentence
- **coherence:** check story telling

Synthetic Dataset

https://crfm.stanford.edu/2023/03/13/alpaca.htm

Early work on synthetic dataset: Alpaca

175 seed instructions-outputs



few shot **3** examples

↓

Strong LLM

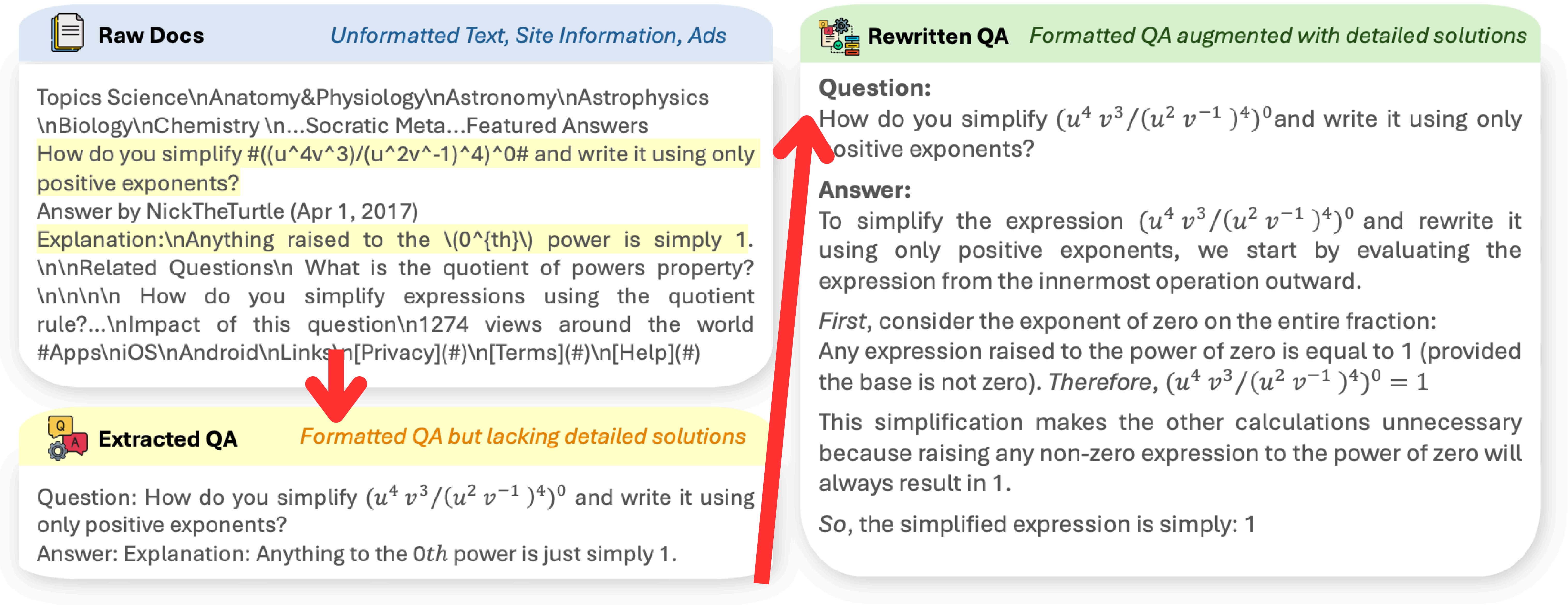
↓

output **20** examples once

52K generated instructions-outputs

WebInstruct

Use LLM to create instruct dataset from webpage



[MAmmoTH2: Scaling Instructions from the Web]

Machine Translation

English Dataset

each example

Strong Translator

output translated example

X language Dataset

Automatic QC

1. **Length filtering**
2. **COMET Score filtering**

COMET

by Unbabel

https://github.com/Unbabel/COMET

Challenges in synthetic data generation: Diversity

Figure 3: The top 20 most common root verbs (inner circle) and their top 4 direct noun objects (outer circle) in the generated instructions. Despite their diversity, the instructions shown here only account for 14% of all the generated instructions because many instructions (e.g., "Classify whether the user is satisfied with the service.") do not contain such a verb-noun structure.

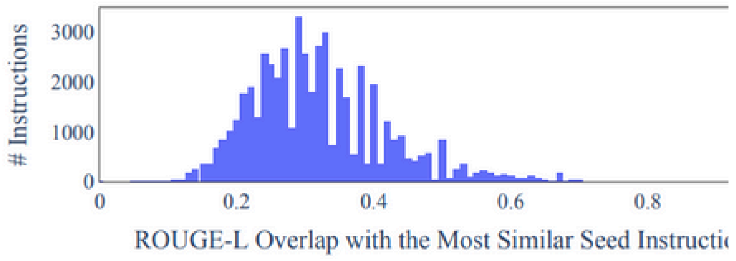


Figure 4: Distribution of the ROUGE-L sc between generated instructions and their n similar seed instructions.

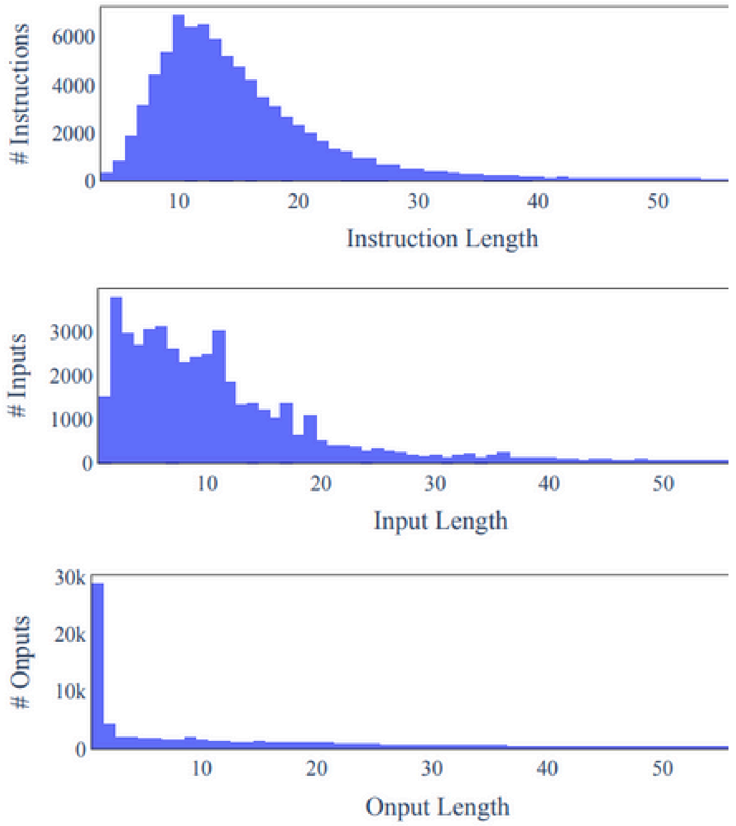


Figure 5: Length distribution of the gener instructions, non-empty inputs, and outpu

- **Root Verbs → Noun** plot
- **ROUGUE-L Similarity** plot
- **Length** plot

[SELF-INSTRUCT: Aligning Language Models with Self-Generated Instructions]()

Challenges in synthetic data generation: Factuality

Need evaluation of synthetic dataset

Human Eval

LLM as a judge

SS4 LLM Hackathon

Question → Query → Table

มีใครบ้างที่อยู่แถวเชียงใหม่ และบ้านอยู่ไม่เกินกว่า 2 เมตร จากแหล่งน้ำ

```sql
SELECT name
FROM table1
WHERE
    name = 'เชียงใหม่'
AND elevation > 2
```

id	name	place	elevation
0	A	เชียงใหม่	1
1	B	กรุงเทพ	5
2	C	กรุงเทพ	1.5

Input รับ
1) คำถาม
2) ตำอย่างตารางบางส่วน

LLM

Output
SQL Code

SS4 LLM Hackathon Solution

Define Scope → Curated Table Scema Dataset → LLM self instruct question and SQL query pair → Automatic Quality Assurance

Domains

- Retail
- Education
- Legal
- Finance

Tasks

- Simple
- Aggregation
- Window
- Set operation

```sql
-- Create a table
CREATE TABLE users (
    id INT,
    name TEXT
);

-- Insert a value
INSERT INTO users (id, name) VALUES (1,
'Alice');
```

Prompt:
"From table **<table schema>** create question and SQL code of task **<task>** that can answer the question, result in json {"sql": <str>, "q":<str>}

Run SQL on a simulated database and removed any execution error

Q/A