The Science of Evaluation for Large Language Models

Lei Li

Language Technologies Institute

Carnegie Mellon University

June 19, 2025

Large Language Models drive the Productivity

Translate

Summarize

Editing

Write email

ChatGPT

Gemini

LLaMA

Chat

Answer questions

Suggest names

Write code

Recommend restaurants

How good is LLM generation?

Prompt: Translate " 新冠疫情危机爆发 ".

LLM output: The outbreak of the new crown crisis

Reference: The outbreak of the COVID-19 crisis

Evaluation

- Reference-based — Metrics: comparing output against references, used for testing.
- Source-based — Reward / Quality estimation (QE) model.

Rule-based and Learned Metrics

Rule-based

- BLEU
- chrF
- TER
- ROUGE

Only surface form difference

Supervised Metric

- BLEURT
- COMET
- MetricX

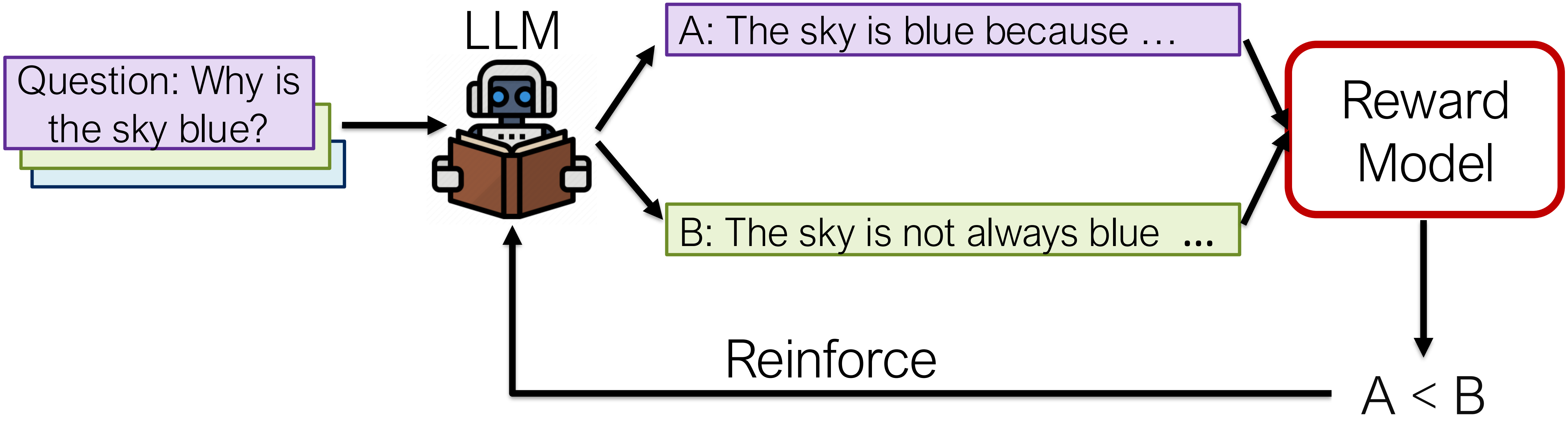
Human rating is scarce

Unsupervised Metric

- SEScore
- BERTScore
- PRISM
- BARTScore

LLM as evaluator?

Learning from Reward / Quality-Estimation Model(QE)



Ouyang et al. Training language models to follow instructions with human feedback. 2022

Challenges in Evaluating LLM

- BLEU/ROUGE will have significantly decreased correlations with human judgments.
- Comprehensive tasks instead of just one task (e.g. MT)
- Open-end generation tasks
- What if no ground truth is given?
    - Source-based evaluation is difficult

Outline

- Can we trust LLM evaluators?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - InstructScore: Interpretable text generation score
- Assessing Knowledge in LLMs (KaRR)

LLM as an Evaluator? (source-based)

Prompt: Translate " 新冠疫情危机爆发 ".

LLM output: The outbreak of the new crown crisis

ask LLM: how good is the above translation?
(MQM scheme: major error=-5, minor error=-1)
LLM output: -5

LLM Evaluator can Help Refine

Input: Translate " 新冠疫情危机爆发 ".

LLM output1: The outbreak of the new crown crisis

Input: Please evaluate the translation quality

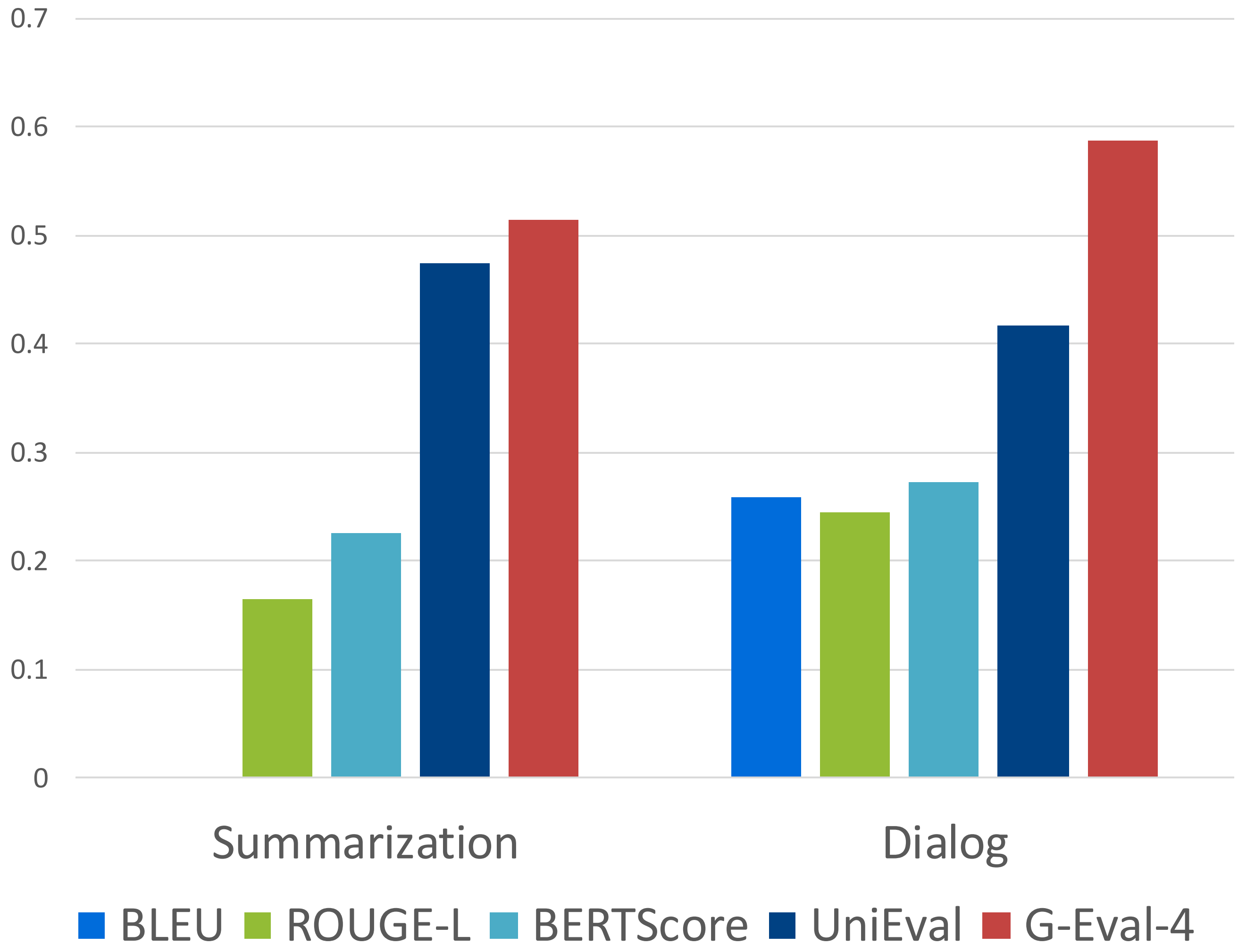
LLM evaluation: The score is -5. there is a major error.

Input: Please revise according to the evaluation.

LLM output2: The outbreak of the corona virus crisis

Aman Madaan, Niket Tandon ..., and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.
Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug.

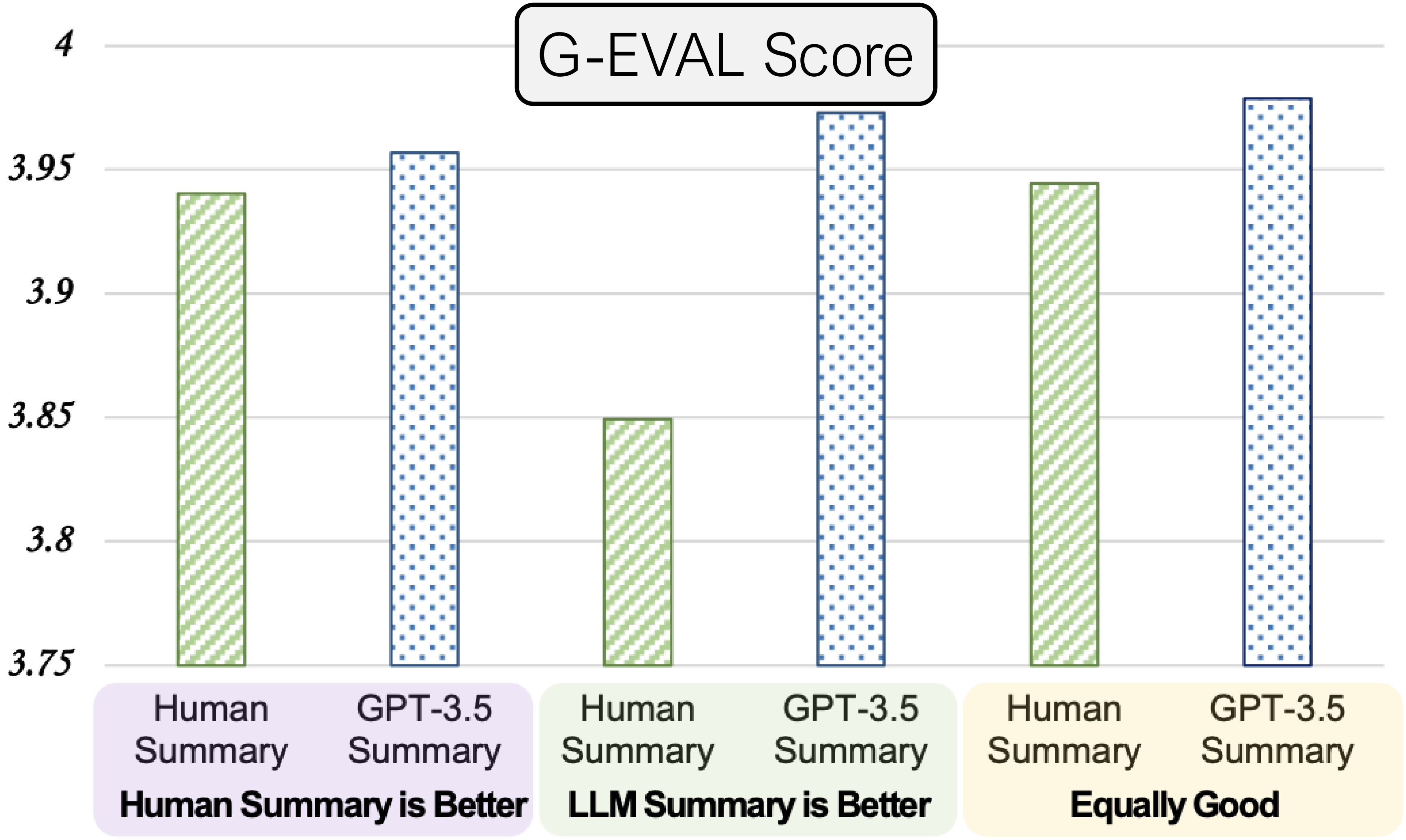
LLM (GPT4) evaluator highly correlates with human evaluation



Liu et al. G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment. 2023.
Chen et al. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. 2023.

But, are LLM evaluators fair?

GPT4 evaluator gives higher scores to its generation!



Liu et al. G-EVAL: NLG Evaluation using GPT-4 with Better Human Alignment. 2023.

A Translation Example

Yoruba text: Ní bayii a ni àwon eku oloshu merin ti ko ni dayabetesi telele to ti ni ayabetesi,” o she afikun.

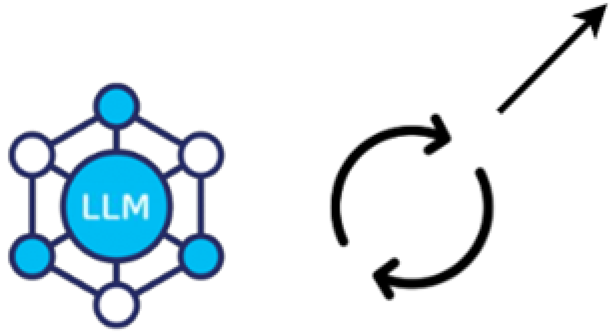
GPT-4’s translation: At this point, we have four rats without diabetes that have developed diabetes," he added.

Using LLM self-evaluate and refine

Human Post Edits: At this point, we have 4-month-old ~~rats~~ mice ~~without diabetes that have developed diabetes~~ that are non-diabetic that used to be diabetic ," he added.

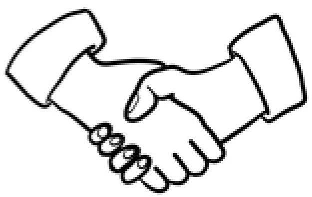
Major error (-5)   Minor error (-1)

GPT-4's evaluation: At this point, we have four rats without diabetes that have developed diabetes," he added.



Self-refine

Human Score: -11



GPT4 Score: -11

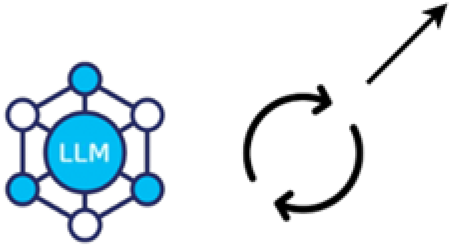
LLM self-refine leads to inflated self-score!

Human Post Edits: Currently, we have 4-month-old ~~healthy rats~~ mice ~~that have developed diabetes~~ that are non-diabetic that used to be diabetic ," he clarified.

Major error (-5)

Minor error (-1)

GPT-4's evaluation: "Currently, we have four healthy rats that have developed diabetes," he clarified.



Self-refine

Human Score: -11

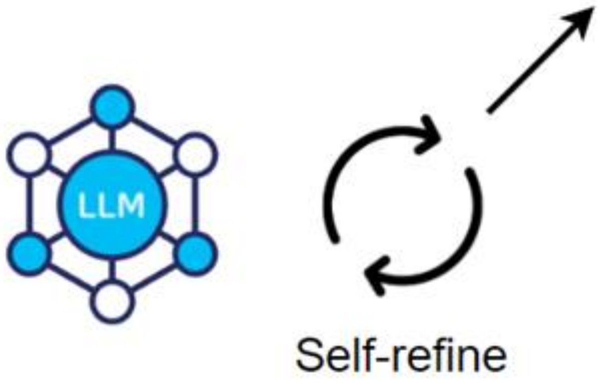
GPT4 Score: -10

LLM self-refine leads to inflated self-score!

Human Post Edits: Presently, we have 4-month-old ~~non-diabetic rats~~ mice ~~that have developed diabetes~~ that are non-diabetic that used to be diabetic ," he elaborated.

Major error (-5) Minor error (-1)

GPT-4's evaluation: Presently, we have four non-diabetic rats that have developed diabetes," he elaborated.



Human Score: -11

GPT4 Score: 0

While GPT-4 thinks it performed self-refine, humans observe all errors persist

LLM 1st generation: At this point, we have four rats without diabetes that have developed diabetes," he added.

LLM 2nd generation: "Currently, we have four healthy rats that have developed diabetes," he clarified.

LLM 3rd generation : Presently, we have four non-diabetic rats that have developed diabetes," he elaborated.

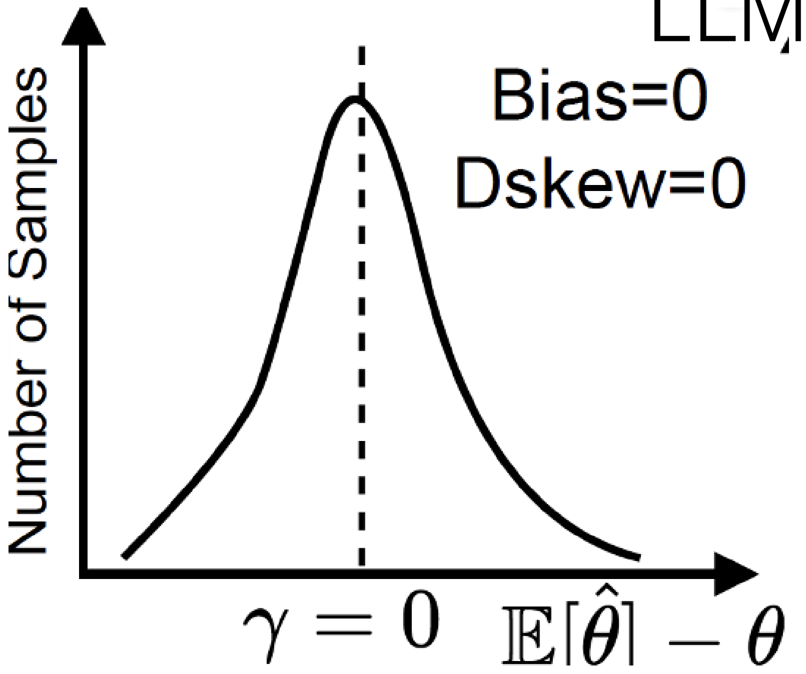
Defining bias in LLM Evaluators

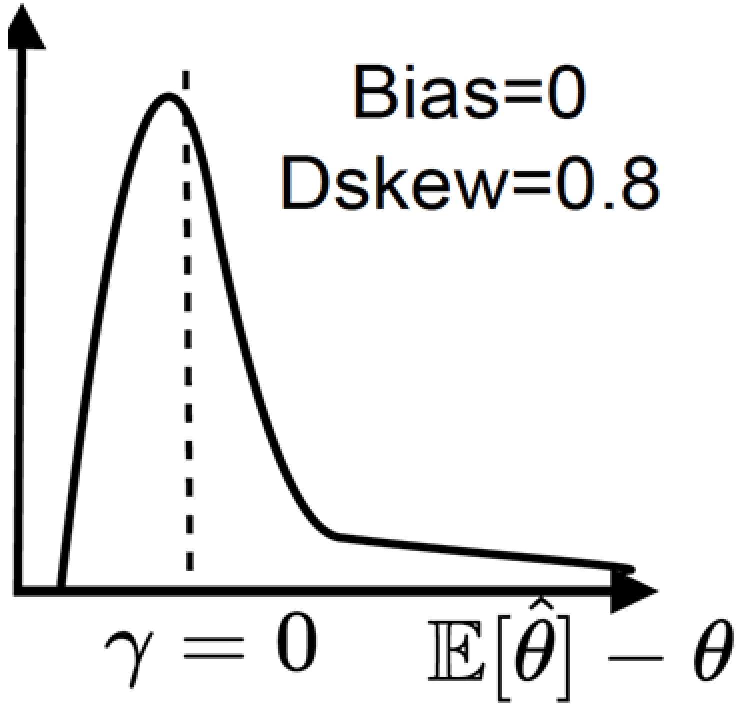
Bias definition 1: Statistical Bias Estimation

$$\text{Bias}(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{E}[\hat{\theta}] - \theta_i)$$

groundtruth score (human)

LLM scores





Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

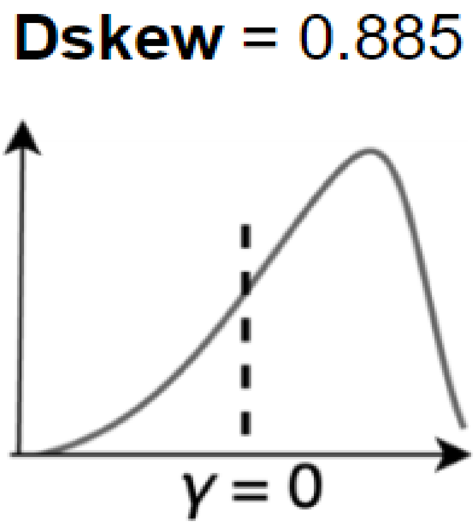
Defining bias in LLM

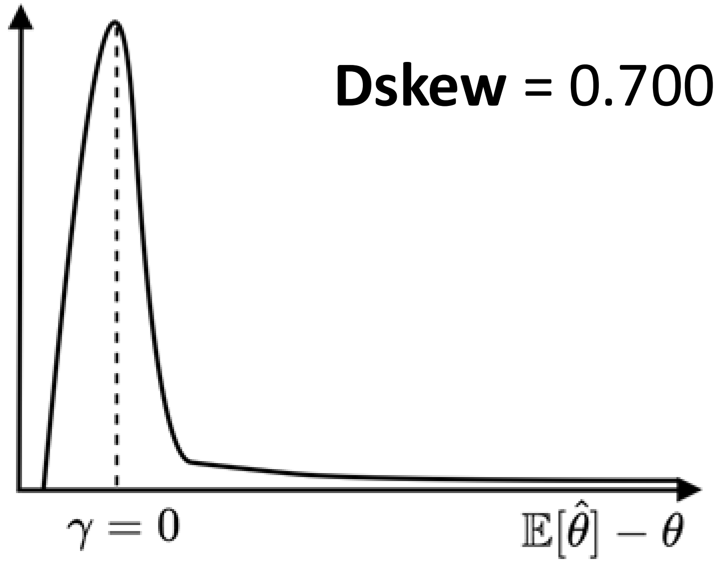
Bias definition 2: Distance Skewness estimation

$$\mathrm{dSkew}(X) = 1 - \frac{\sum_{i,j}\|x_i - x_j\|}{\sum_{i,j}\|x_i + x_j - 2\gamma\|}$$

LLM scores

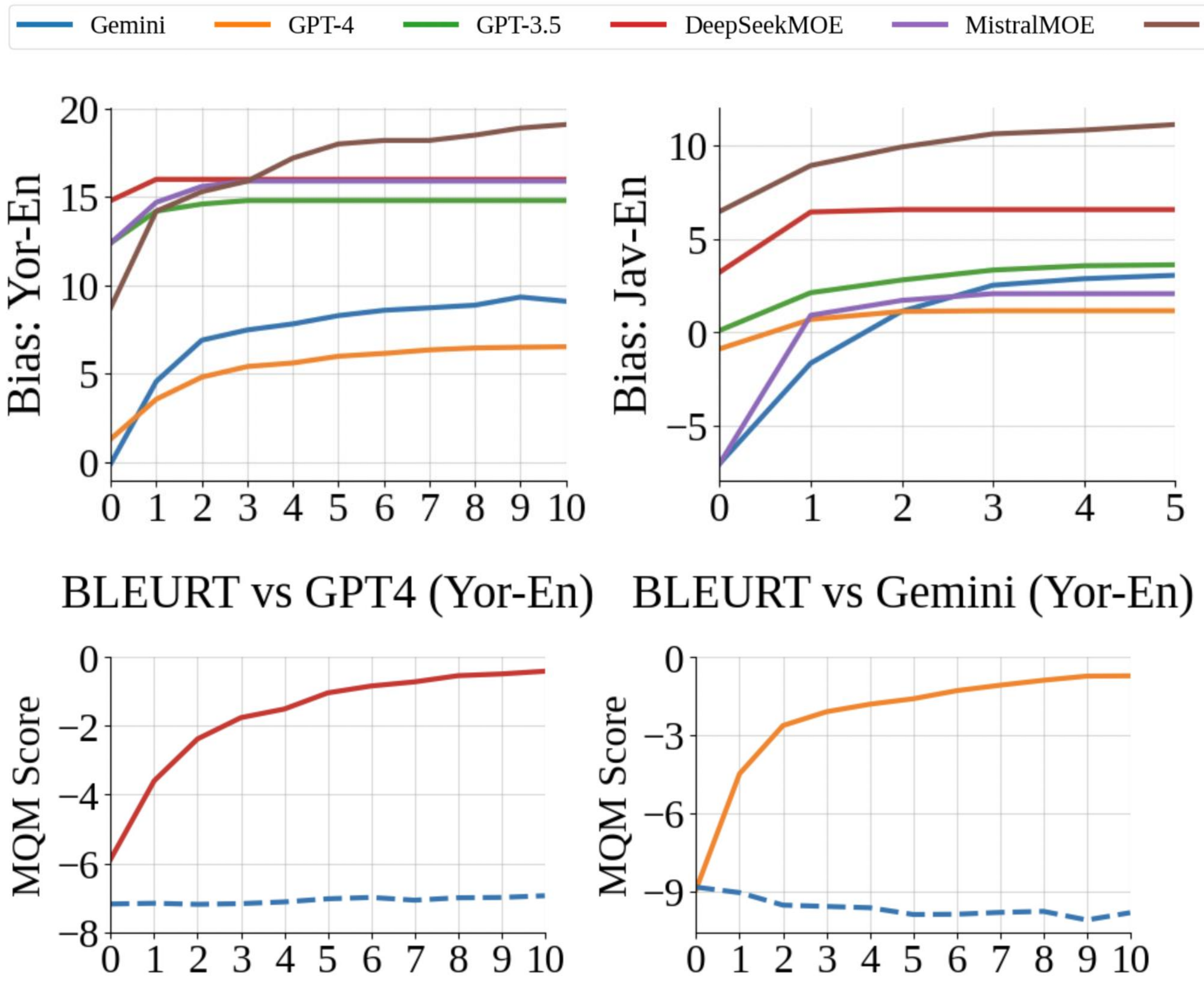
location





Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

Self-Bias Amplifies in LLM Translation

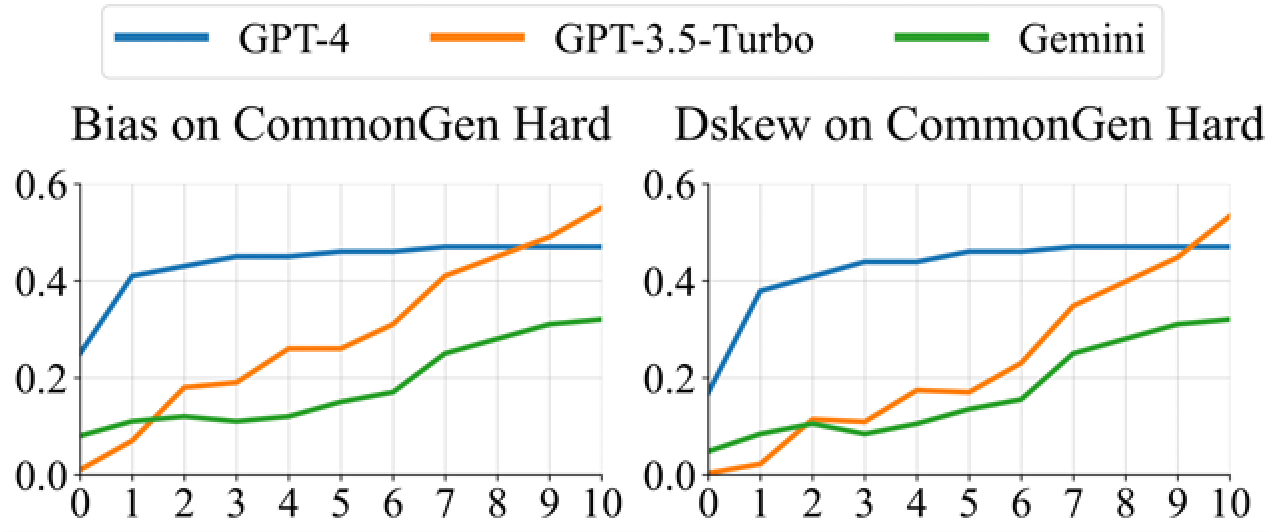


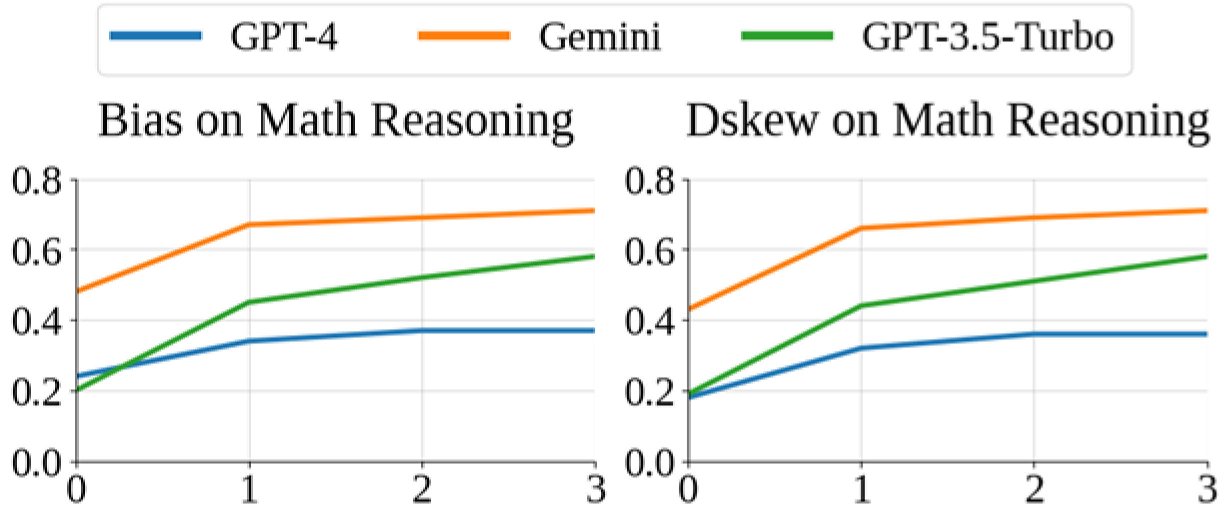
What is the root cause of self-bias amplification?

- GPT-4 and Gemini overestimate improvements in self-refined outputs, compared to actual performance measured by BLEURT

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

Self-Bias Amplifies in LLM Data-to-Text and Math

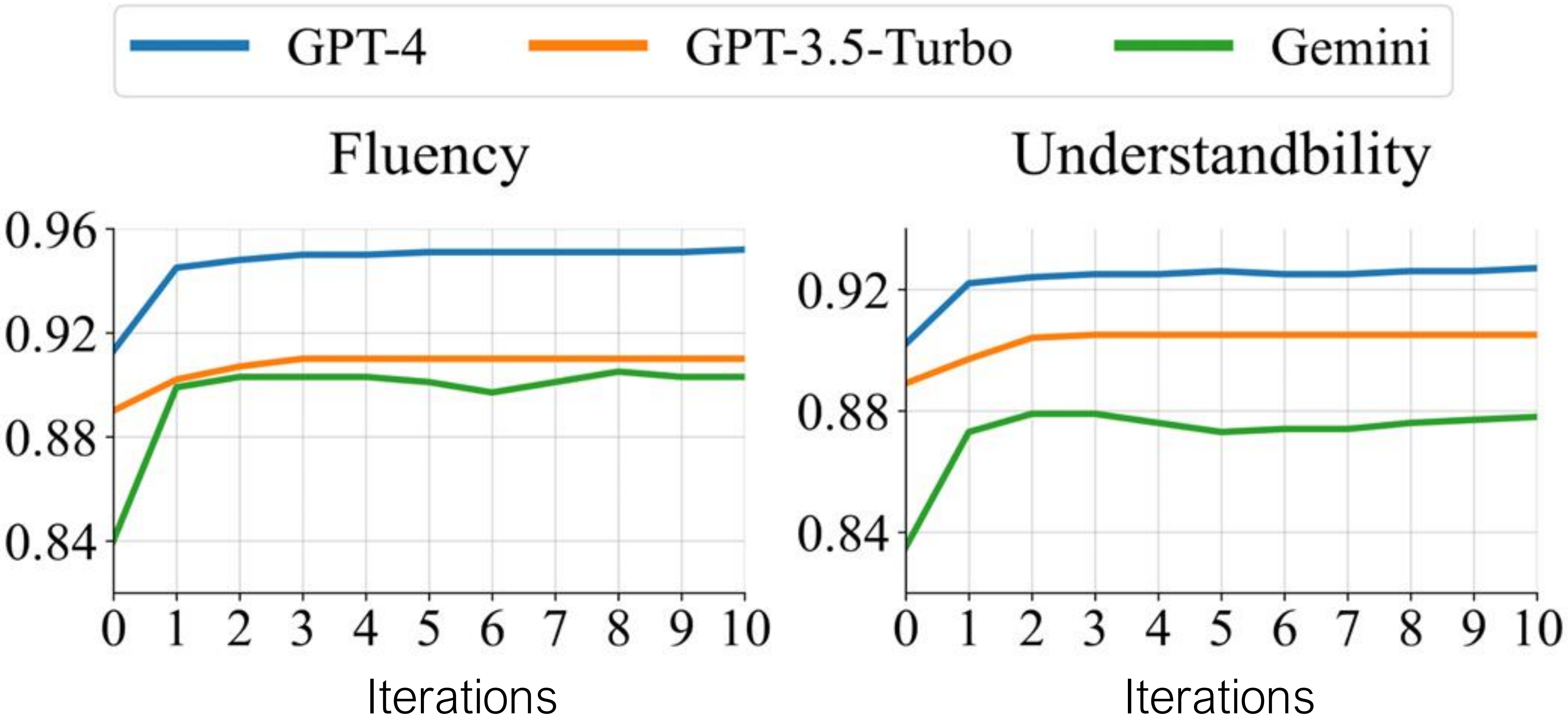




Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

What is improving at Self-refine if not quality?

Self-refine improves understanding and fluency of the text



Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

Key insights

- LLM evaluators have strong self-bias
- Self-bias is amplified during LLM self-refine/self-rewarding process
- Self-refine can improve fluency of text but not necessarily quality
- LLMs favor texts that follow their ‘style’

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

Outline

- Can we trust LLM evaluators?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - InstructScore: Interpretable text generation score
- Assessing Knowledge in LLMs (KaRR)

When you made a mistake…

Teacher 1: You have a bad translation. You get score of 20/100

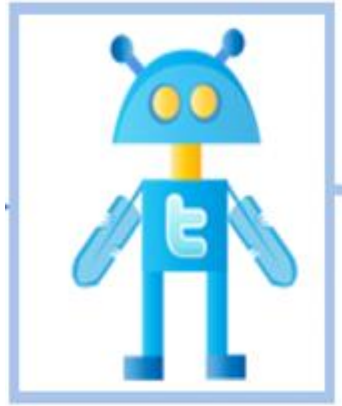
outbreak of

Teacher 2:
'New crown' is a major mistranslation error. The correct translation is 'COVID-19'.
Score: 20/100

Evaluating Text Generation Quality – Existing metrics

Reference: The outbreak of the COVID-19 crisis

Gen Candidate: The outbreak of the new crown crisis



BLEU: 0.661

BertScore: 0.925

COMET: 0.711

BLEURT: 0.519

SEScore2: -5.43

Ideal Metric: Fine-grained Explanation

Reference: The outbreak of the COVID-19 crisis

Candidate: The outbreak of the new crown crisis

Error location: new crown

Error type: Terminology is used inconsistently

Major/Minor: Major

Explanation: The term " new crown" is not the correct term for “Covid-19".

Direct Prompting ChatGPT

Raw text: "The art … between providing enough detail to … too much information."

Error type 1: Translation includes information not present in the correct translation

Major/minor: Major

Incorrect generation:

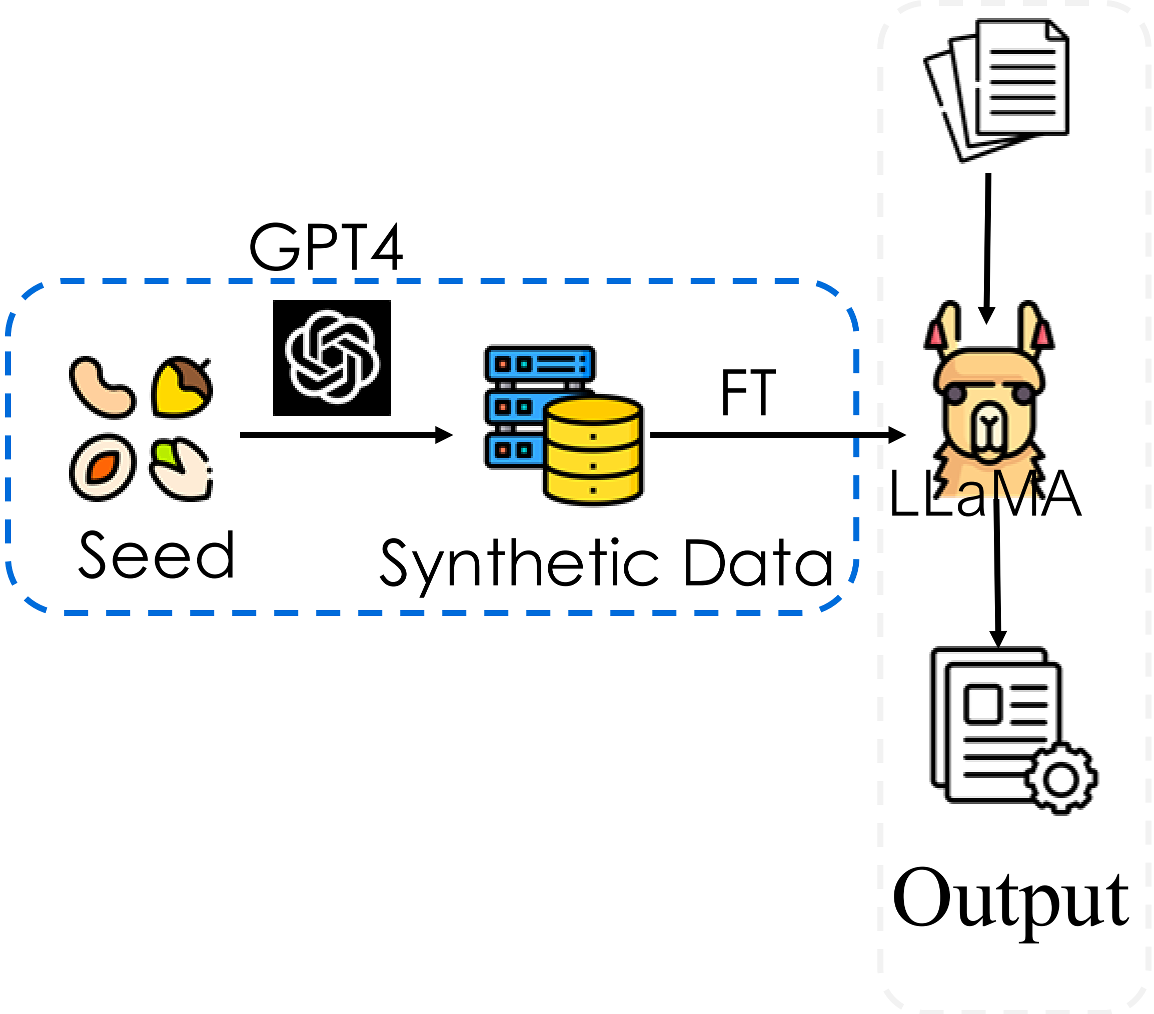
[GPT4 fill in]

Error location 1: [GPT4 fill in]

Explanation for error 1:

[GPT4 fill in]

But, failed explanation in GPT4

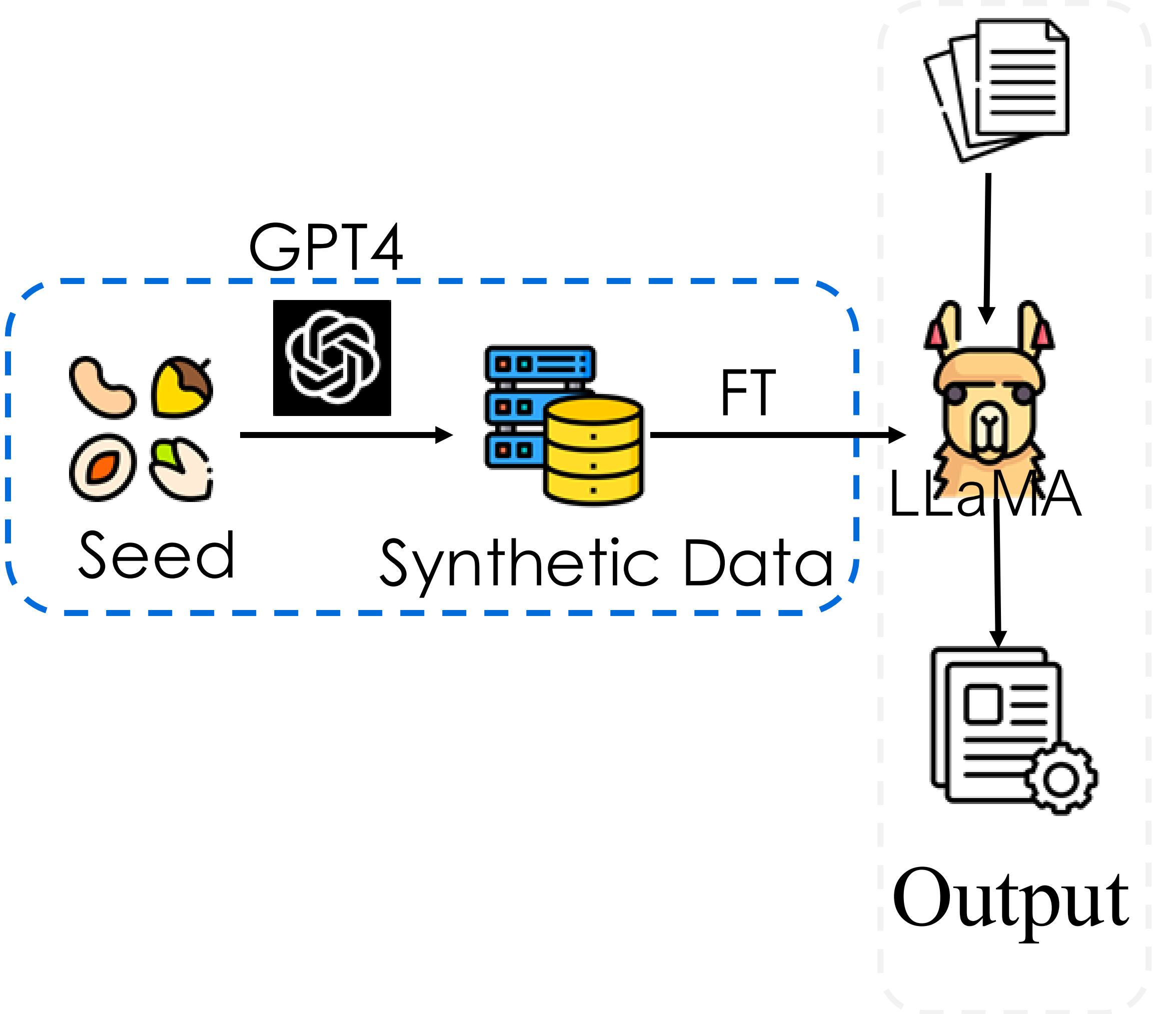


Error type 3: Missing information

Explanation for error 3: The incorrect translation adds the word "annual" to the phrase ...

Error type is inconsistent with explanation

But, failed explanation in GPT4

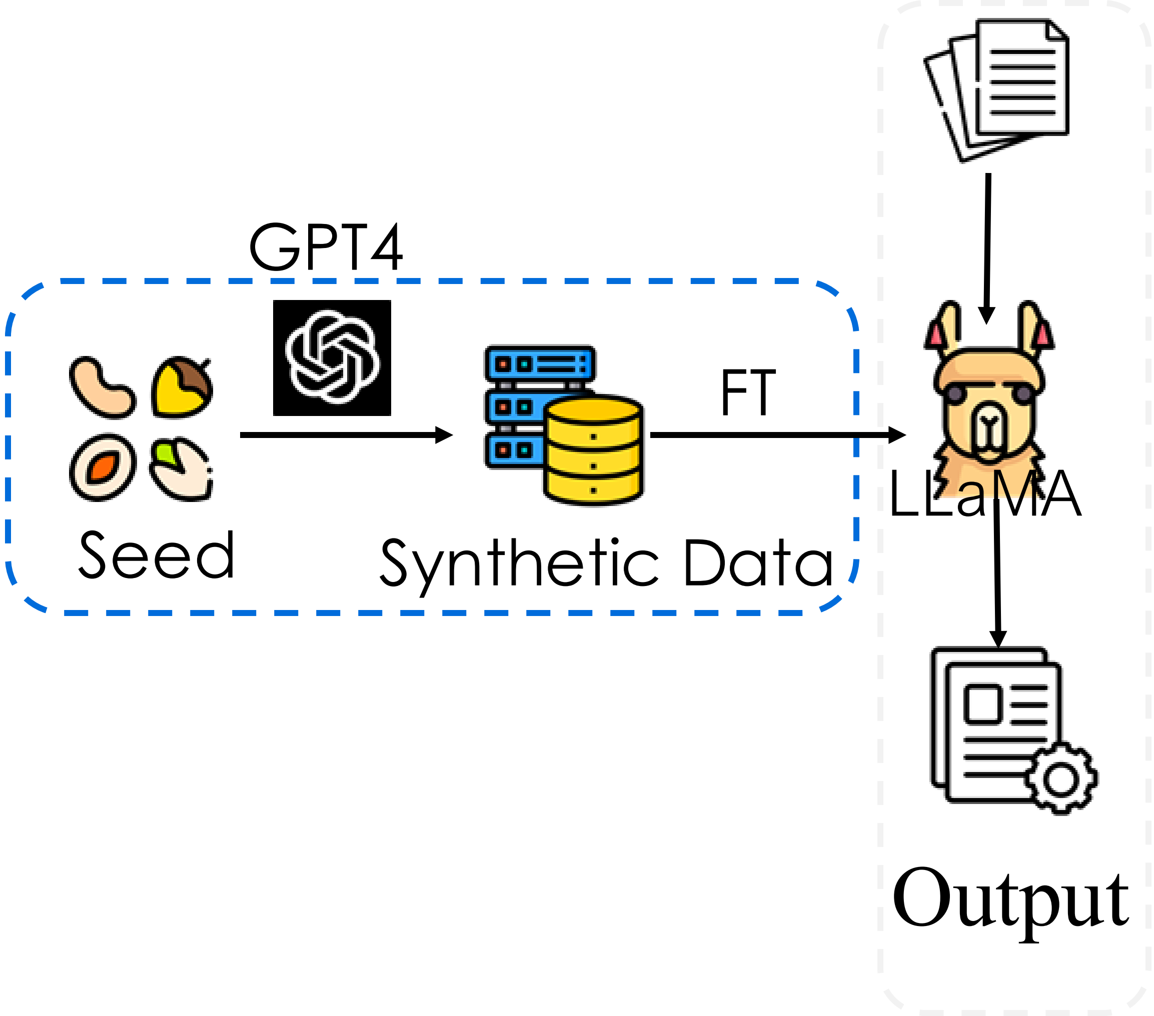


Evaluated text: The outbreak of the new crown crisis

Error location: 'virus'

Hallucination

But, failed explanation in GPT4



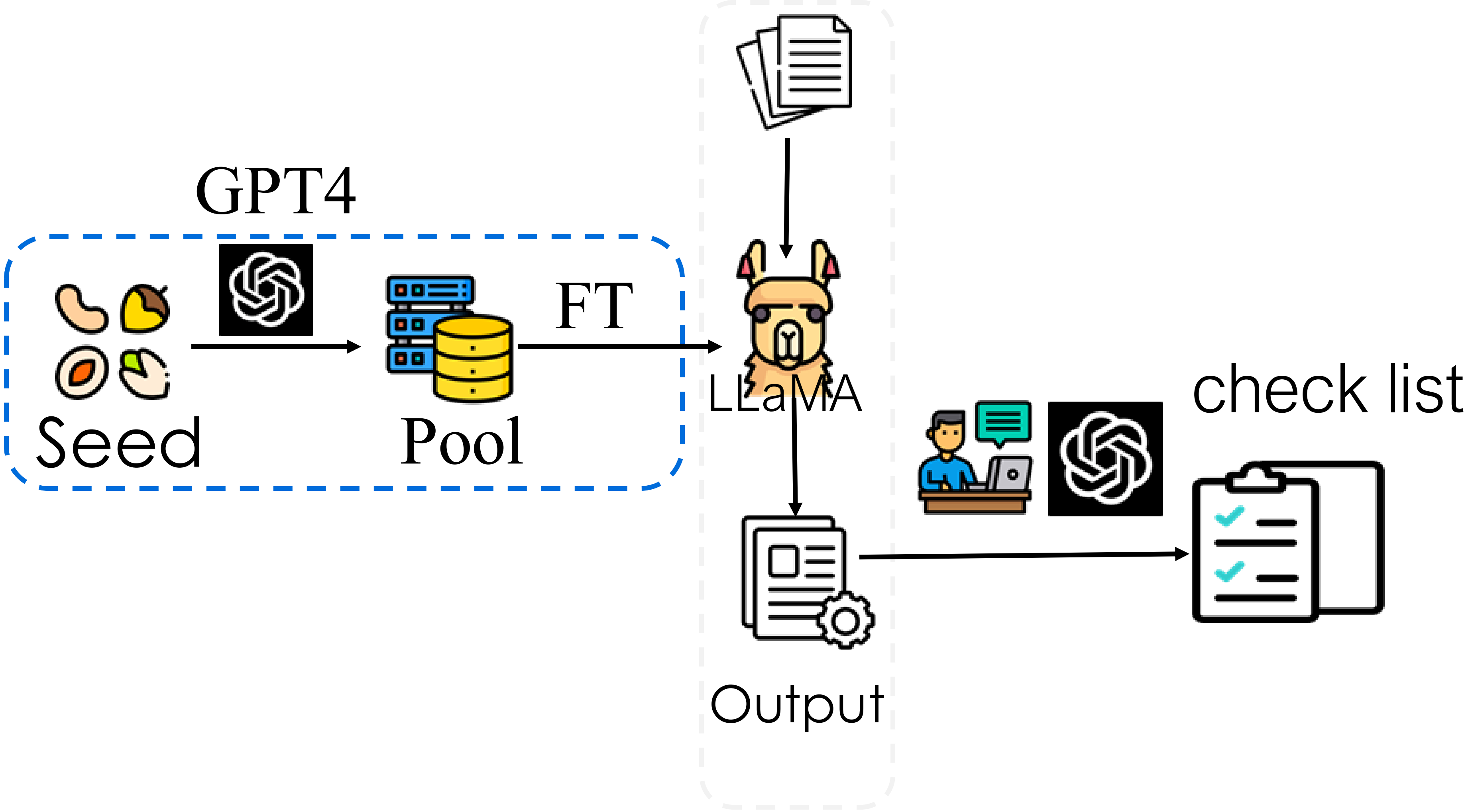
Explanation for error 1: The incorrect translation uses the word "annual" instead of "annual"

Explanation is illogical

Failures of GPT4 generated explanation

Fields	**Failure Mode**	**Description (M is local failure mode, G is global failure mode)**
Error Type	Inconsistency to explanation	M1: Error type is inconsistent with explanation
Error Location	Inconsistency to explanation	M2: Error locations are not consistent with the explanation
	Hallucination	M3: Error locations are not referred in the output text
Major/Minor	Major/Minor disagreement	M5: Major and minor labels are not correct
Explanation	Hallucination	M4: Error locations are not referred in the output text
	Explanation failure	M6: Explanation is illogical
All 4 Fields	False negative error	G1: Error described in the explanation is not an error
	Repetition	G2: One error is mentioned more than once among explanations
	Phrase misalignment	G3: Incorrect phrase and correct phrase are not aligned
	Mention multiple errors	G4: One error span mentions multiple errors

Introducing InstructScore



Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

Use GPT-4 as a checking Model

Reference: *…… revolutionary base area……*
Output: *……the old revolutionary district……*

Correct: revolutionary base area

Incorrect: old revolutionary district

Does output contain this error?

Is the error type consistent with explanation?

Are two phrase aligned?

InstructScore: Automatic Feedback

Reference
Candidate

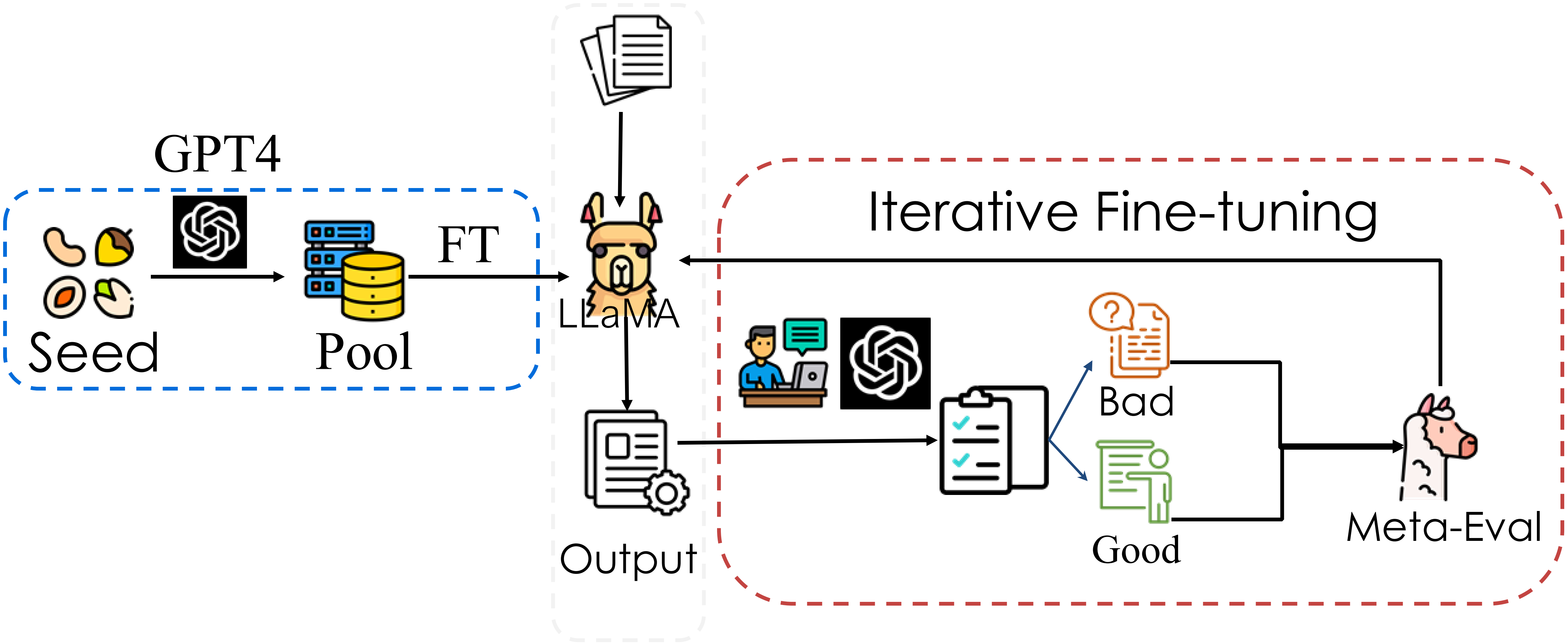
Error location1
Error Type1
Major/Minor
Explanation1

Error location2
Error Type2
Major/Minor
Explanation2

Error1	Error location	✓
	Error type	✓
	Major/minor	✗
	Explanation	✓
Error2	Error location	✓
	Error type	✓
	Major/minor	✓
	Explanation	✓

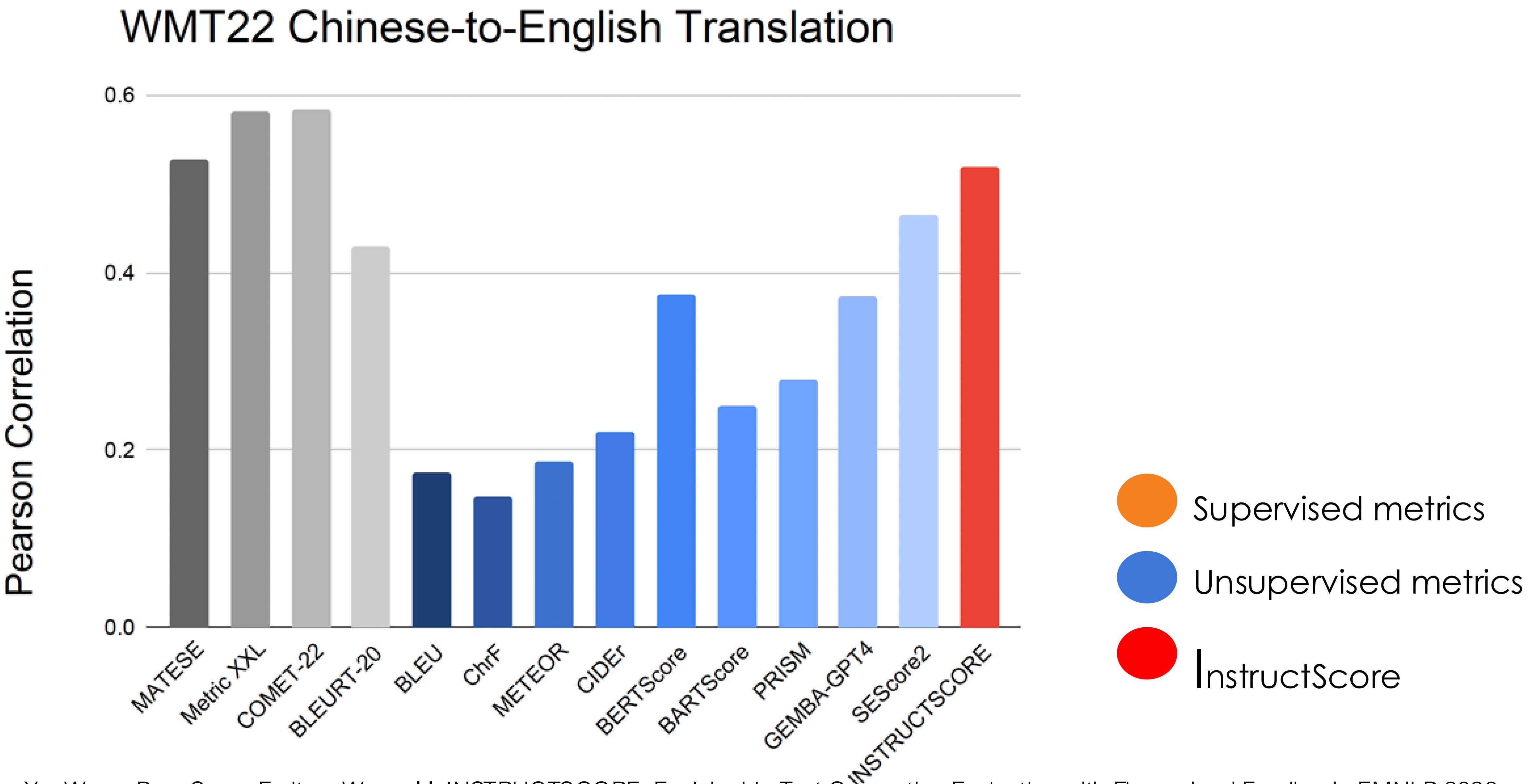
Alignment Score: 7/8

InstructScore: Training with Refinement



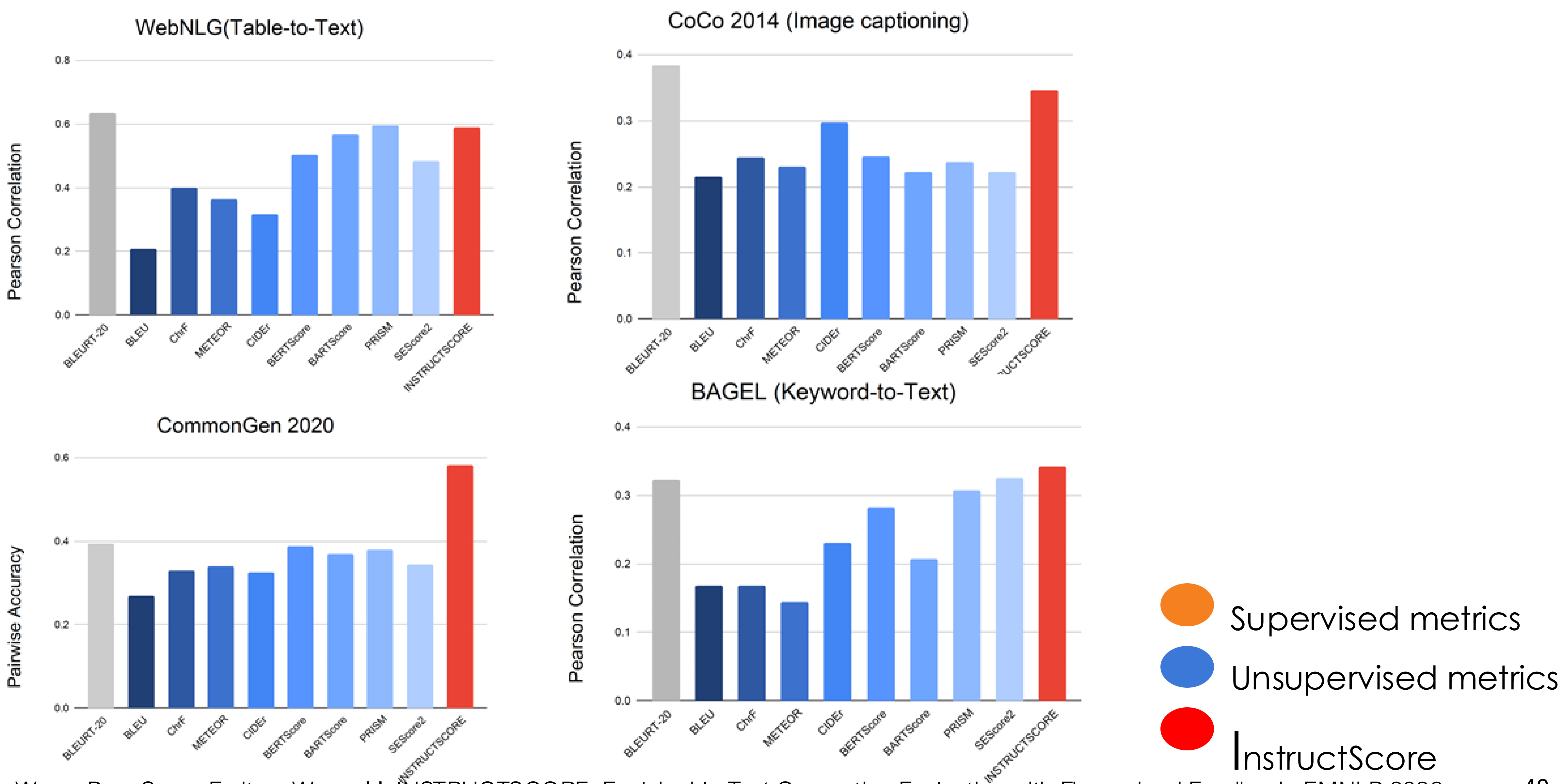
Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

InstructScore can judge machine translation!



Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

InstructScore can evaluate text generation!



Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

InstructScore vs GPT4

Candidate: Or how long before the restaurant is ready? Can you ask for me?
Reference: And how much longer does it take for the restaurant to prepare it? Could you help me ask them?

InstructScore vs GPT4

Error Type: Lexical Selection
Major/Minor: Major
Error Location: "Or" instead of "And"

Error Type: Lexical Selection/Omission
Major/Minor: Major
Error Location: "Can you ask for me?" instead of "Could you help me ask them?"

2X

Error Type: Omission
Major/Minor: Minor
Error Location: Missing "Could" in "Could you help me ask them?"

Error type 1: Incorrect translation is missing content from the correct translation
Major/minor: Major
Error location 1: how long before

Error type 2: Problems with grammar, other than orthography
Major/minor: Minor
Error location 2: help me ask

Highlights of InstructScore

- We develop a new model-based evaluation metric for Explainable text generation-based metric and leverage automatic feedback to align with human requirements!

1. Fine-grained Explainability
2. Highly Aligned with Human
3. Generalizability (No human ratings are required!)

Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

Outline

- Can we trust LLM evaluators?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - InstructScore: Interpretable text generation score
- Assessing Knowledge in LLMs (KaRR)

LLMs generates Unreliable Answers

- e.g. LLaMA-7B

When did Shakespeare die?

Llama-7B : 23rd April 1616.

LLMs generates Unreliable Answers

- e.g. LLaMA-7B

On what date did William Shakespeare's death occur?

Llama-7B : It was on 23 august 1616.

Knowing versus Guessing

1. Distinguish if text generation stems from genuine knowledge or just high co-occurrence with given text.

   William Shakespeare's job is a writer.

   John Smith's job is a writer.

Unreliable Factual Knowledge in LLMs

- LLMs often generate **unreliable answers given varying prompts.**

- Example1: Alpaca-7B

William Shakespeare's job is?
: A playwright. ✓

The job of Swan of Avon is?
: A boatman. ✗

- Example2: ChatGPT

William Shakespeare's job is?
: A playwright and teacher. ✓

Is William Shakespeare a teacher?
: None. ✗

Assessing LLM's Knowledge

- Given varying prompts regarding a factoid question, can a LLM **reliably** generate factually **correct** answers?

When did Shakespeare die?

On what date did William Shakespeare's death occur?

Generative Language Model

23rd April 1616. He is …

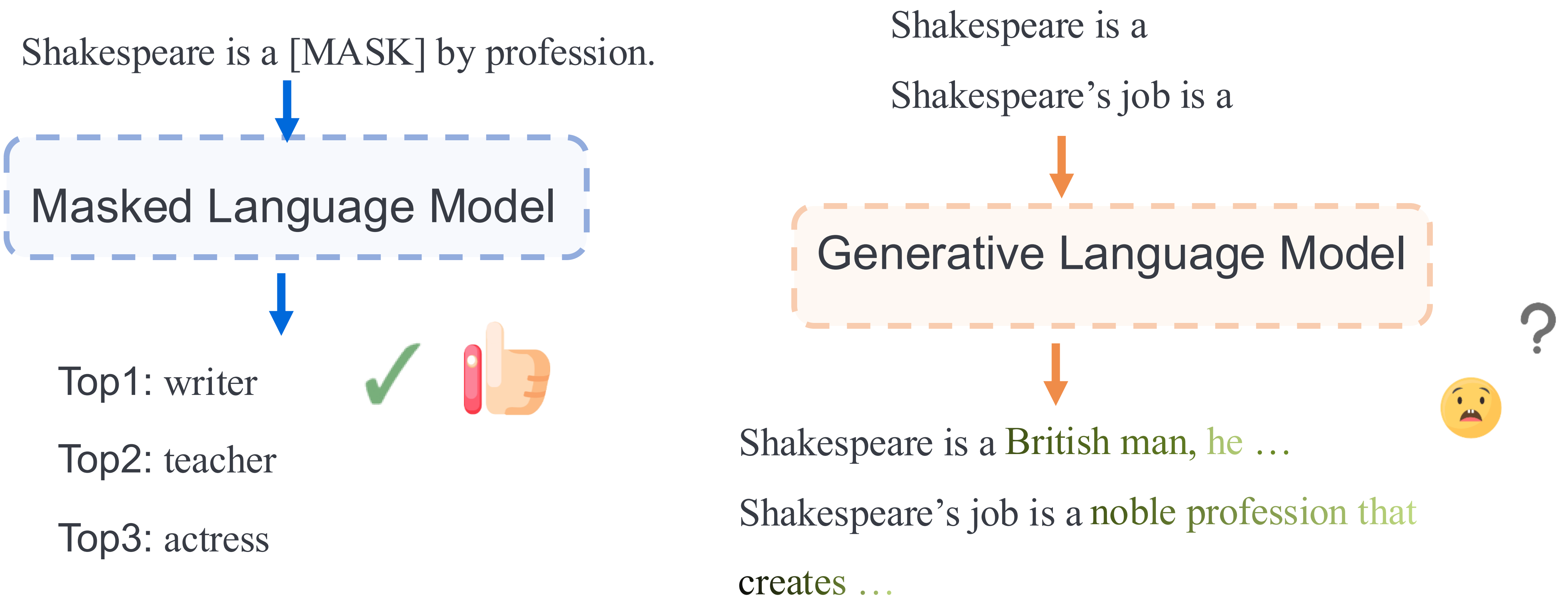
It was on 23 April 1616 and…

Reliable?

Dong et al. Statistical Knowledge Assessment for LLMs. Neurips 2023

Challenges in LLM Knowledge Assessment

- **Knowledge irrelevant generation:** The freely generated results of generative models might be irrelevant to factual knowledge.



Risk Ratio

- In statistics, **risk ratio** estimate the strength of the association between exposures (treatments or risk factors) and outcomes.

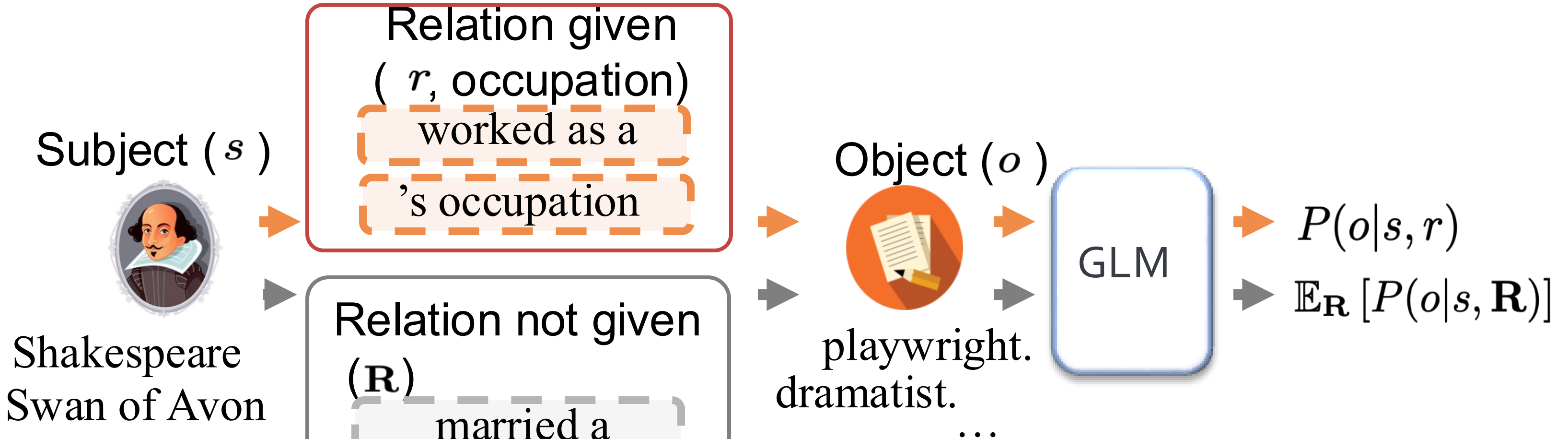
- Example: a disease noted by D, and no disease noted by $\neg D$, exposure noted by E, and no exposure noted by $\neg E$. The risk ratio can be written as:

- $\text{Risk } Ratio = \dfrac{P(D|E)}{P(D|\neg E)}$

	E （exposure）	$\neg E$ （no exposure）
D (disease)	P(D|E)	P(D|¬E)
¬D (no disease)	P(¬D|E)	P(¬D|¬E)

Knowledge Assessment Risk Ratio (KaRR)

- Assesses the joint impact of subject and relation symbols on the LLM's ability to generate the object symbol.



$$KaRR_r(s,r,o) = \frac{P(o|s,r)}{\mathbb{E}_{\mathbf{R}}[P(o|s,\mathbf{R})]}$$

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, Lei Li. Statistical Knowledge Assessment for LLMs. Neurips 2023

KaRR Dataset

- Broad coverage
  - 1million entities
  - 600 relations

Method	Subj. Alias	Obj. Alias	Rel. Alias	Rel. Cvg.
LAMA@1	✗	✗	✗	6.83%
LAMA@10	✗	✗	✗	6.83%
ParaRel	✗	✗	✓	6.33%
KaRR	✓	✓	✓	100%

"P36": {

"capital city": "[X] is the capital city of [Y].",

"administrative capital": "[X] is the administrative capital of [Y].",…

},

"P19": {

"birthplace": "[X]'s birthplace is [Y].",

"born in": "[X] was born in [Y].",

"POB": "The POB of [X] is [Y].",

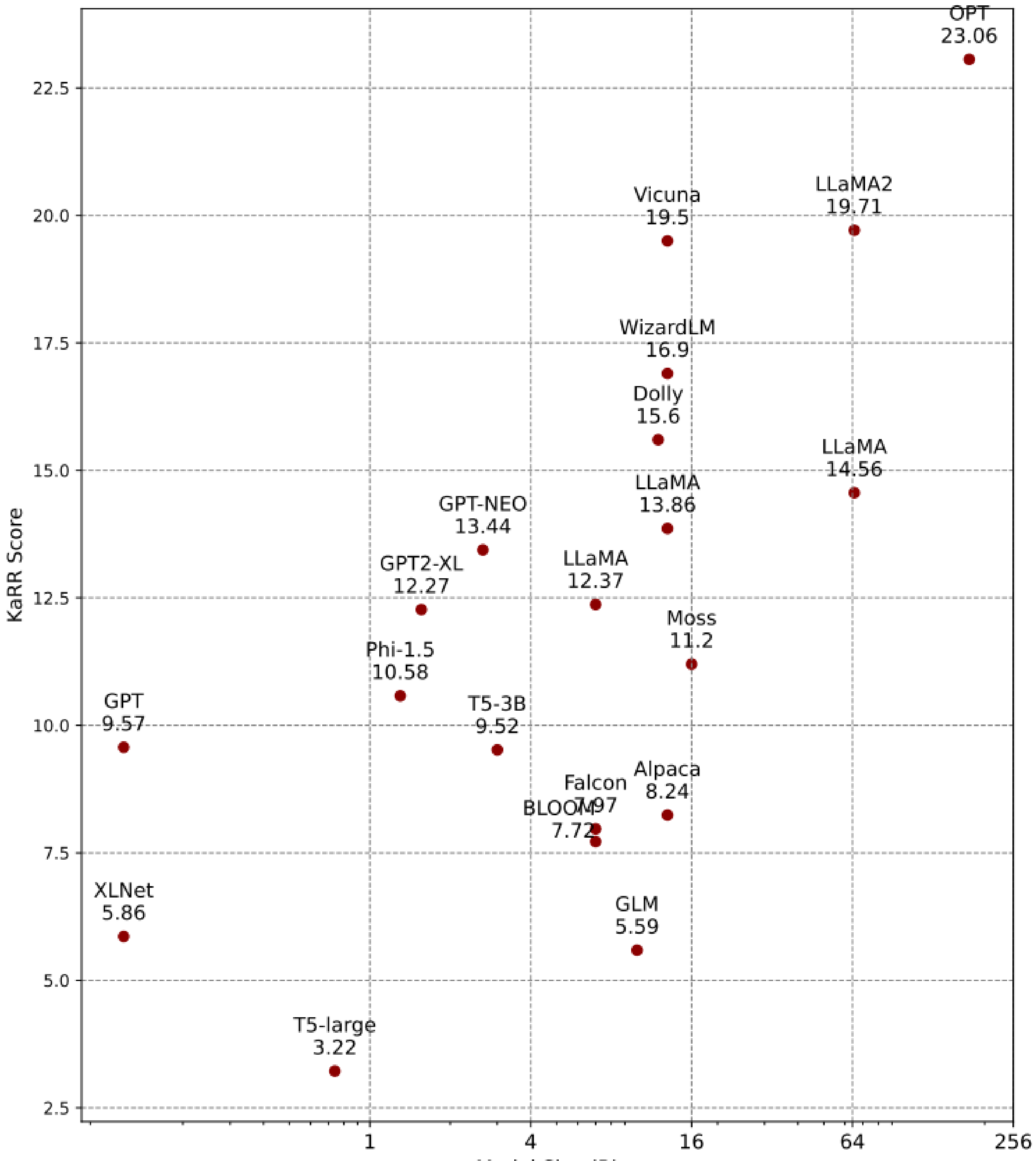
"birth place": "The birth place of [X] is [Y].",

"location of birth": "The location of birth of [X] is [Y].", …

Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, Lei Li. Statistical Knowledge Assessment for LLMs. Neurips 2023

KaRR Scores for 20 LLMs

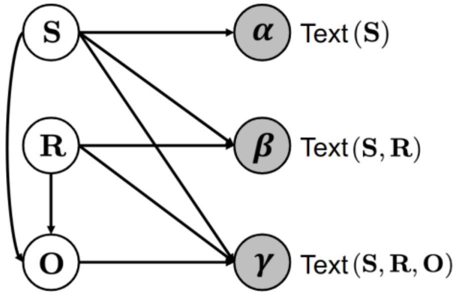
- Small and medium-sized LLMs struggle with generating correct facts consistently.
- Finetuning LLMs with data from more knowledgeable models can enhance knowledge.

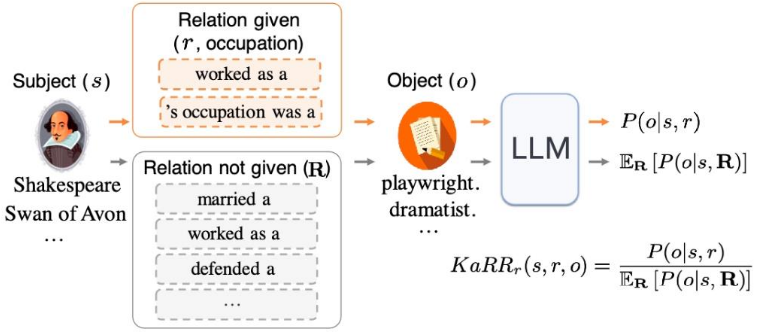


Summary of LLM Knowledge Assessment

- Graphical model for knowledge Assessment
- New metric -- KaRR Score
- High human correlation
- Less evaluation bias

Code and data:
dqxiu/KAssess (github.com)





Summary

- Can we trust LLM evaluators?
  - LLM Evaluators exhibit strong bias towards itself
  - Self-bias is amplified in LLM self-refine
- Evaluating LLM Generation Quality
  - InstructScore: Interpretable text generation score
- Assessing knowledge in LLMs (KaRR)
  - KaRR measures how reliable are LLM in generating fact-related answers

Future thoughts

- Evaluating
  - complex knowledge
  - LLM RAG
  - LLM Agent
- Evaluation for open-end generation
  - PerSE at EMNLP 2024

Reference

- Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024.
- Xu, Wang, Pan, Song, Freitag, Wang, Li. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.
- Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023.
- Xu, Qian, M. Wang, Li, W. Y. Wang. SESCORE2: Learning Text Generation Evaluation via Synthesizing Realistic Mistakes. ACL 2023.
- Xu, Tuan, Lu, Saxon, Li, Wang. Not All Errors are Equal: Learning Text Generation Metrics using Stratified Error Synthesis (SEScore). EMNLP 2022.