The Science of Evaluation and Alignment for Large Language Models

Lei Li
Language Technologies Institute
Carnegie Mellon University
October 30, 2024

Large Language Models drive the Productivity

Translate

Summarize

Editing

Write email

ChatGPT

Gemini

LLaMA

Chat

Answer questions

Suggest names

Write code

Recommend restaurants

Language Models: The Power of Predicting Next Word

$$Prob.(next\_word|prefix)$$

Santa Barbara has very nice ____	beach	0.5
	weather	0.4
	snow	0.01
Pittsburgh is a city of ____	bridges	0.6
	corn	0.02

Language Model: $P(x_{1..T}) = \prod_{t=1}^{T} \underline{P(x_{t+1}|x_{1..t})}$

Predict using Neural Nets

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. Alignment training

Rule-based and Learned Metrics

Rule-based

- BLEU
- chrF
- TER
- ROUGE

Only surface form difference

Supervised Metric

- BLEURT
- COMET

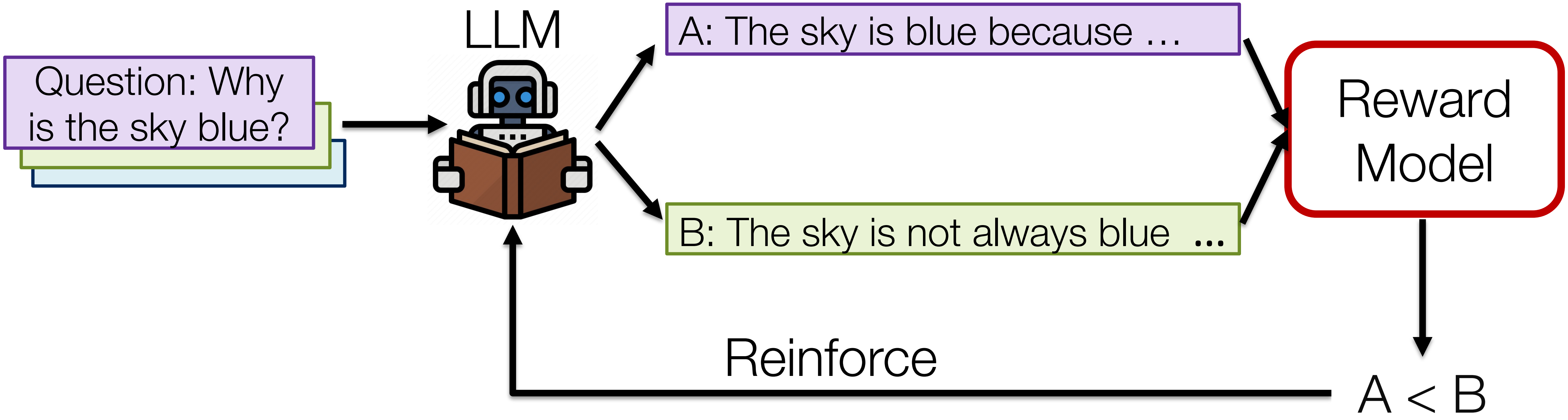
Human rating is scarce

Unsupervised Metric

- SEScore
- BERTScore
- PRISM
- BARTScore

LLM as evaluator?

Learning from Reward / Quality-Estimation Metric(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 evaluator?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - Interpretable text generation evaluation (InstructScore)
  - Assessing knowledge in LLMs (KaRR)
- Post-training Alignment
  - Online Preference Optimization (BPO)
  - Iterative refinement with fine-grained feedback (LLMRefine)

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?
(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

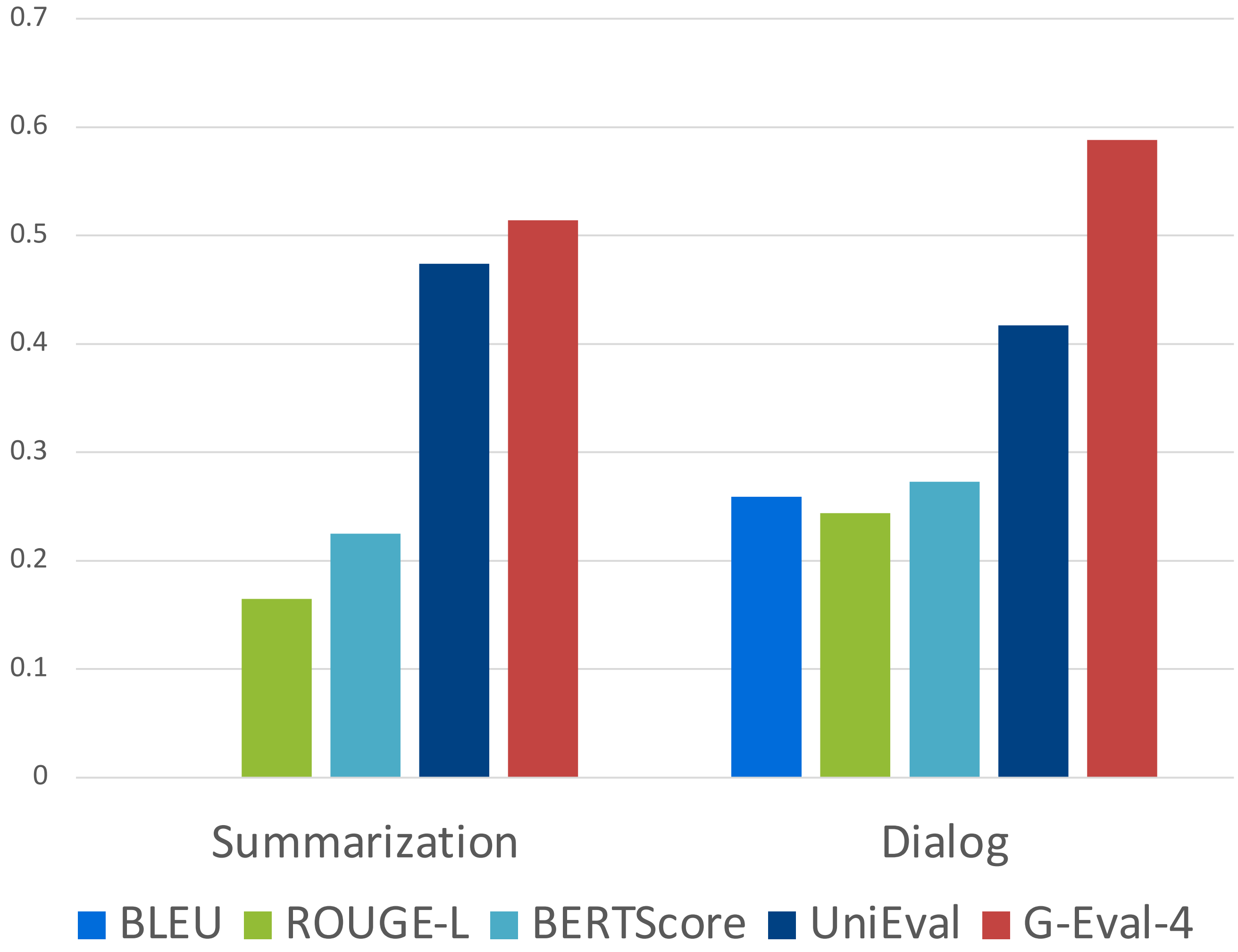
LLM evaluation: The score is -5. there is a major error.

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.

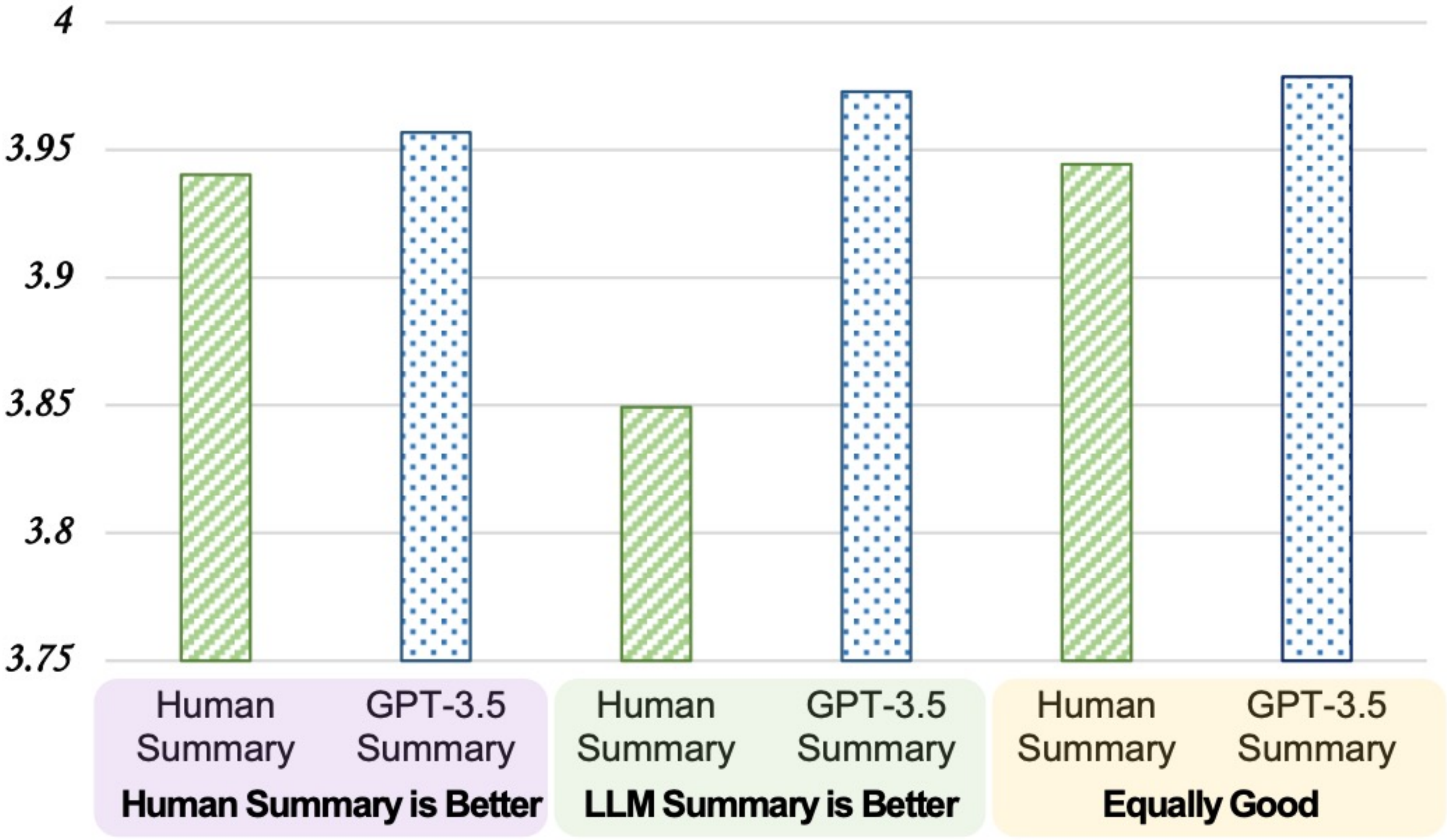
LLM (GPT4) evaluator highly correlates with human evaluation



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!



Translation Example

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

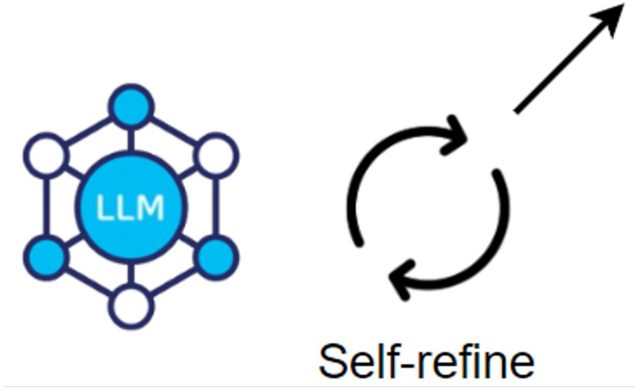
GPT-4's translation: At this point, we have four rats without diabetes that have developed diabetes," he added.

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.



Human Score: -11

GPT4 Score: -11

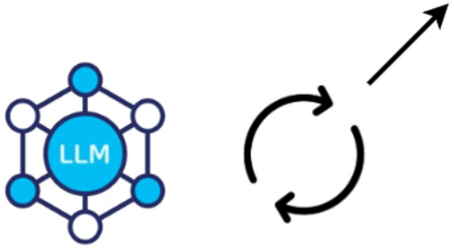
LLM self-refine leads to inflated self-score!

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

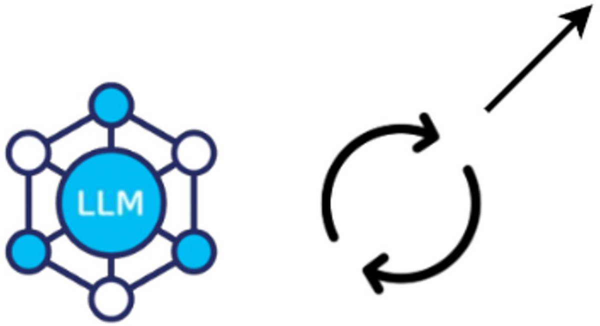
GPT4 Score: -10

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.



Self-refine

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.

LLM self-bias goes beyond translation!

Concepts: ['fruit', 'motorcycle', 'perform', 'jacket', 'vehicle', 'place', 'mat', 'walk', 'world', 'area', 'kiss', 'mother', 'pass', 'report', 'club', 'axis', 'tricep', 'patient', 'listen', 'owner', 'uniform', 'floor', 'hamburger', 'use', 'wine', 'cross', 'bull', 'sell', 'lawn', 'friend']

GPT-4's generation: In a world where a fruit can perform like a motorcycle ……

We test word coverage ratio

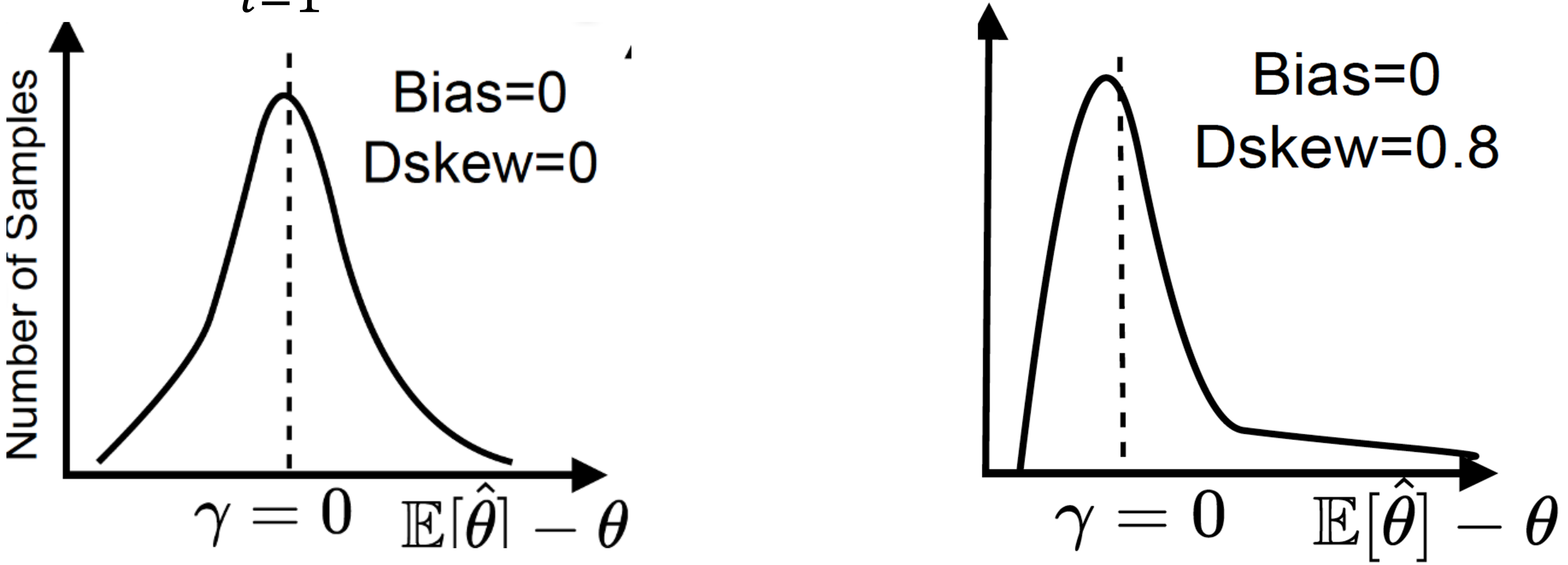
Human Score: 40%

GPT4 Score: 80%

Defining bias in LLM Evaluators

Statistical Bias Estimation

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

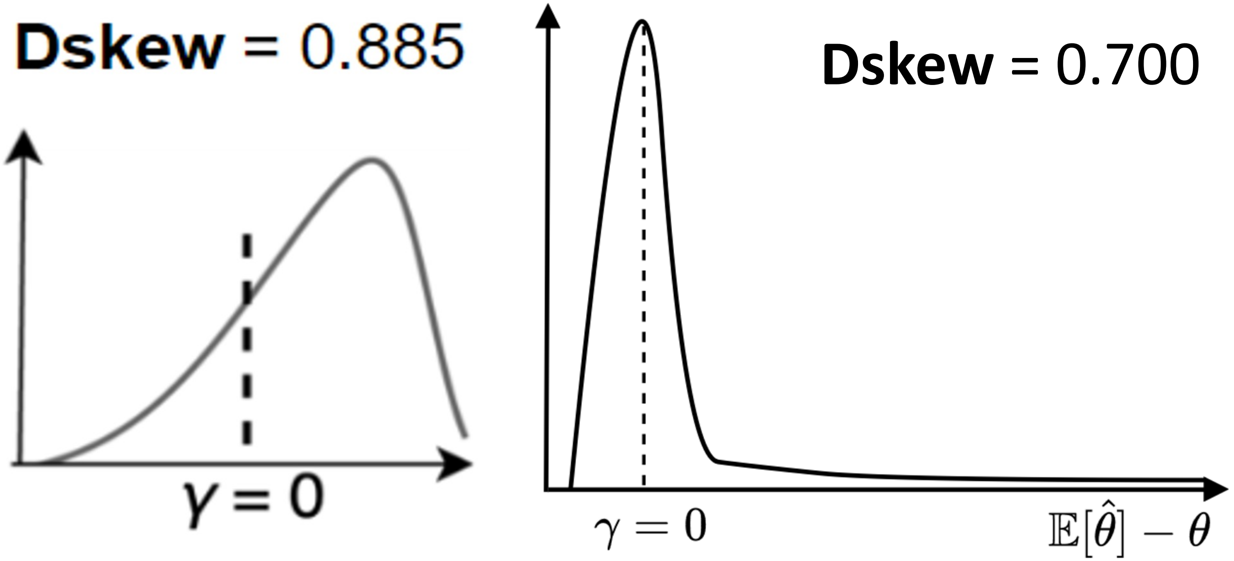


Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

Defining bias in LLM

Distance Skewness estimation

$$d\text{Skew}_n(X) = 1 - \frac{\sum_{i,j} \|x_i - x_j\|}{\sum_{i,j} \|x_i + x_j - 2\gamma\|}$$



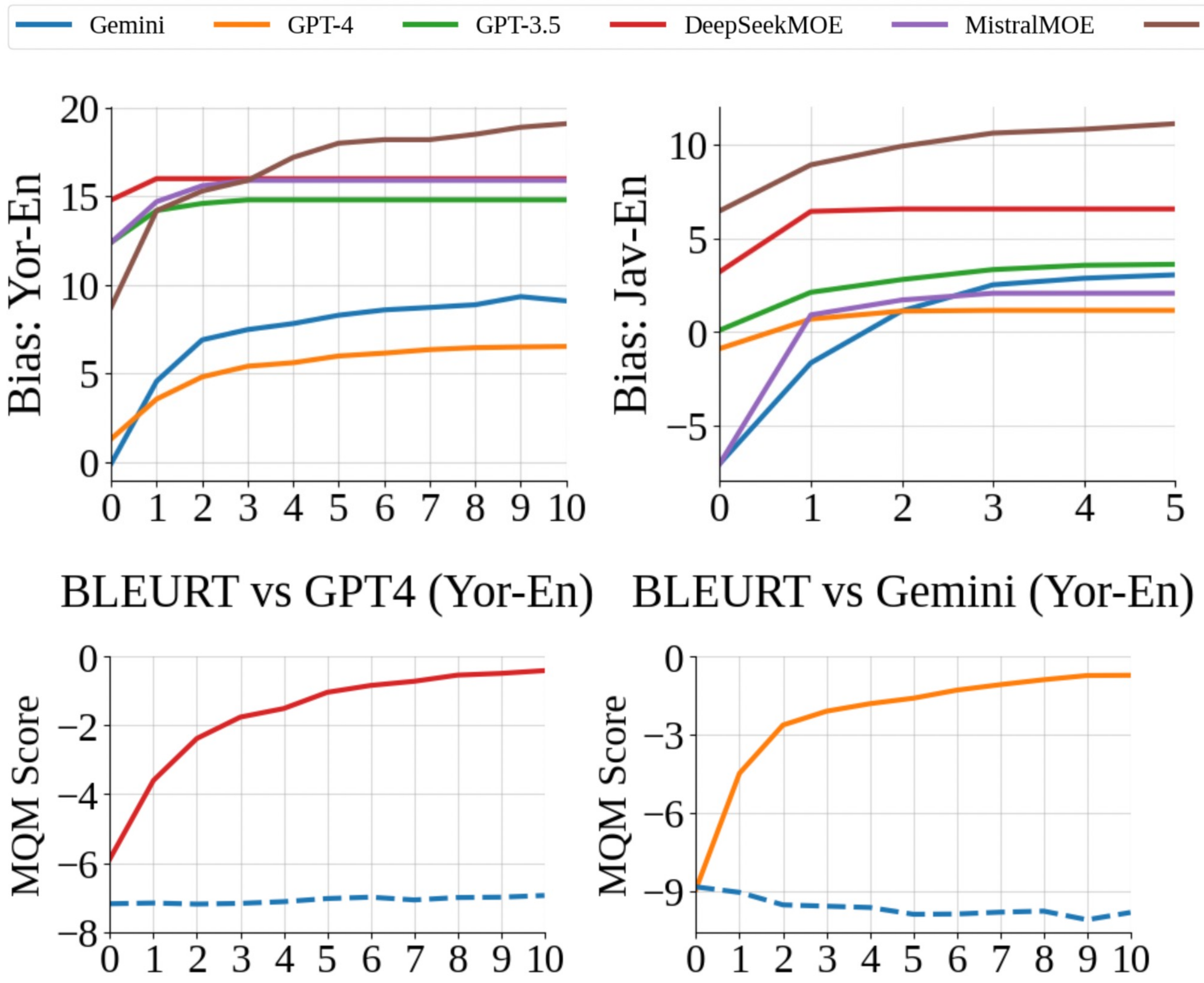
Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. ACL 2024

Quantifying Bias in LLM Evaluators

- Q1: Are LLM self-bias amplified across tasks, languages?
- Q2: What is improved after self-refine?
- Q3: What are factors to alleviate self-bias?

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 Amplification at Translation

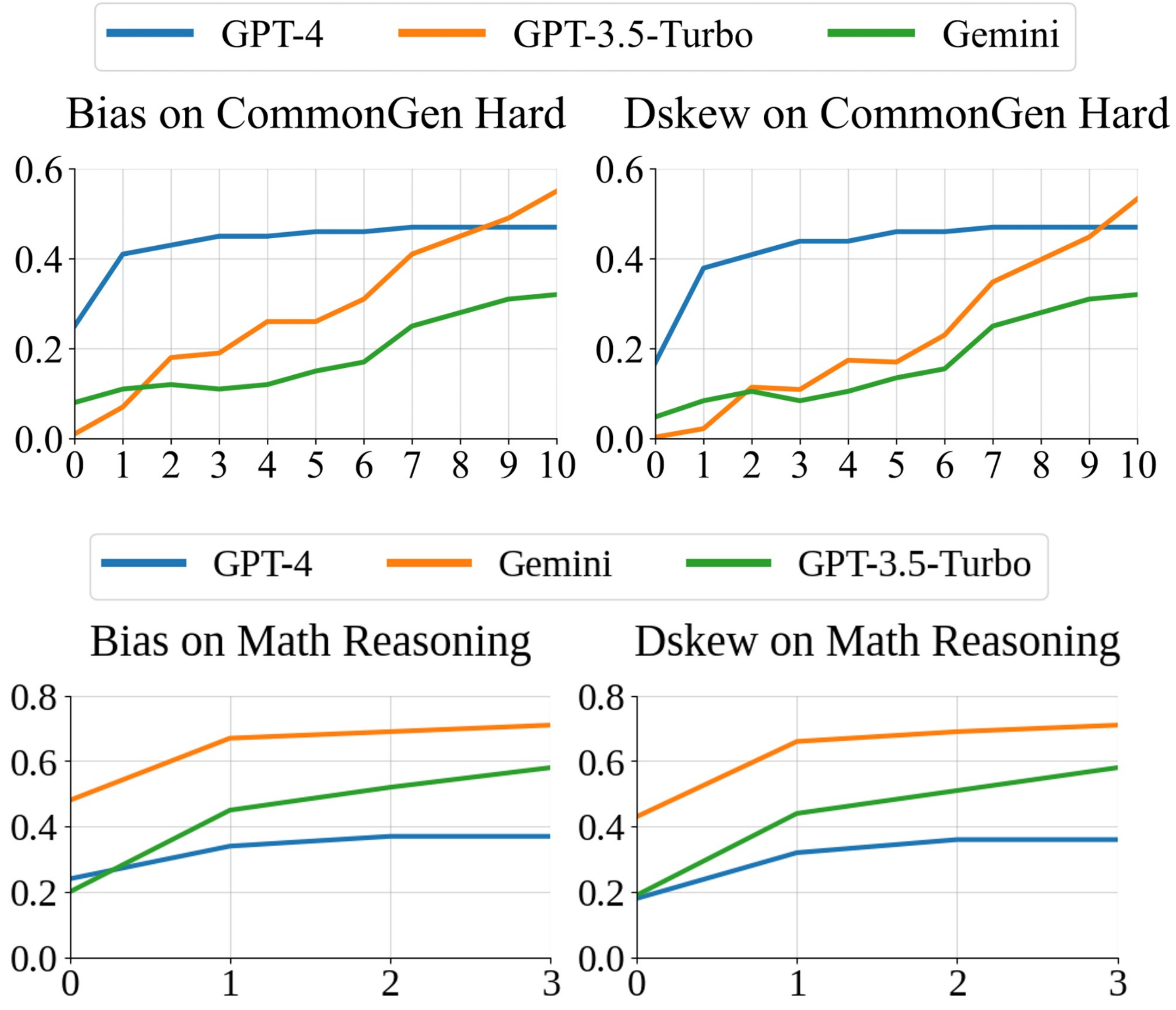


What is the root cause of self-bias amplification?

- 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

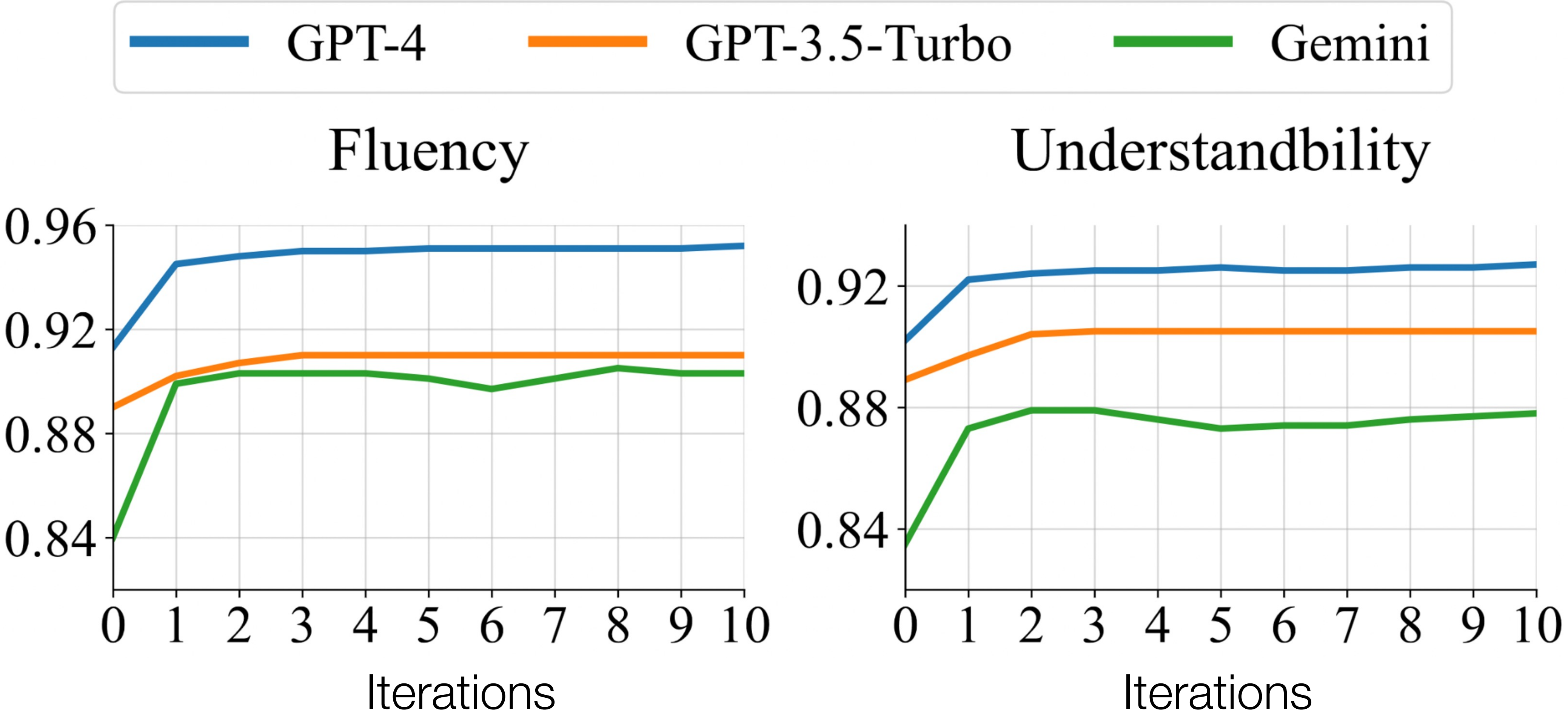
Self-Bias Amplification at Data-to-Text and Math



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

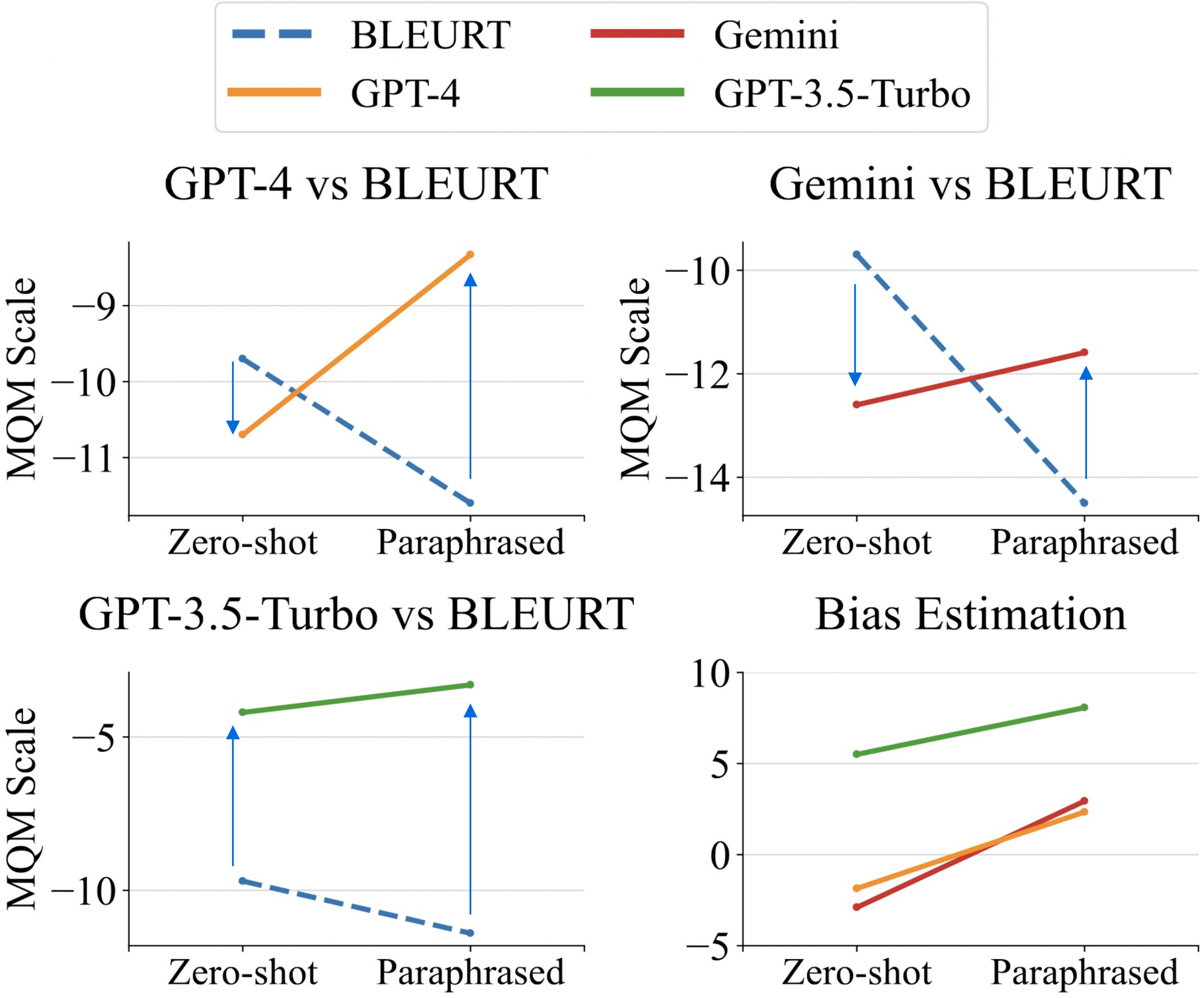
What is improving at Self-refine if not quality

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

LLMs favor texts that follow their style



Paraphrase other LLM (Madlad-400)'s translation can significantly increase bias on LLM's estimation

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 evaluator?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - Interpretable text generation evaluation (InstructScore)
  - Assessing knowledge in LLMs (KaRR)
- Post-training alignment
  - Online Preference Optimization (BPO)
  - Iterative refinement with fine-grained feedback (LLMRefine)

When you made a mistake…

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

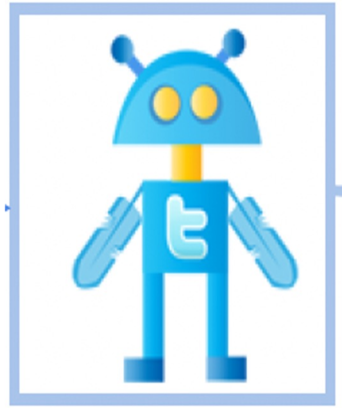
outbreak

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

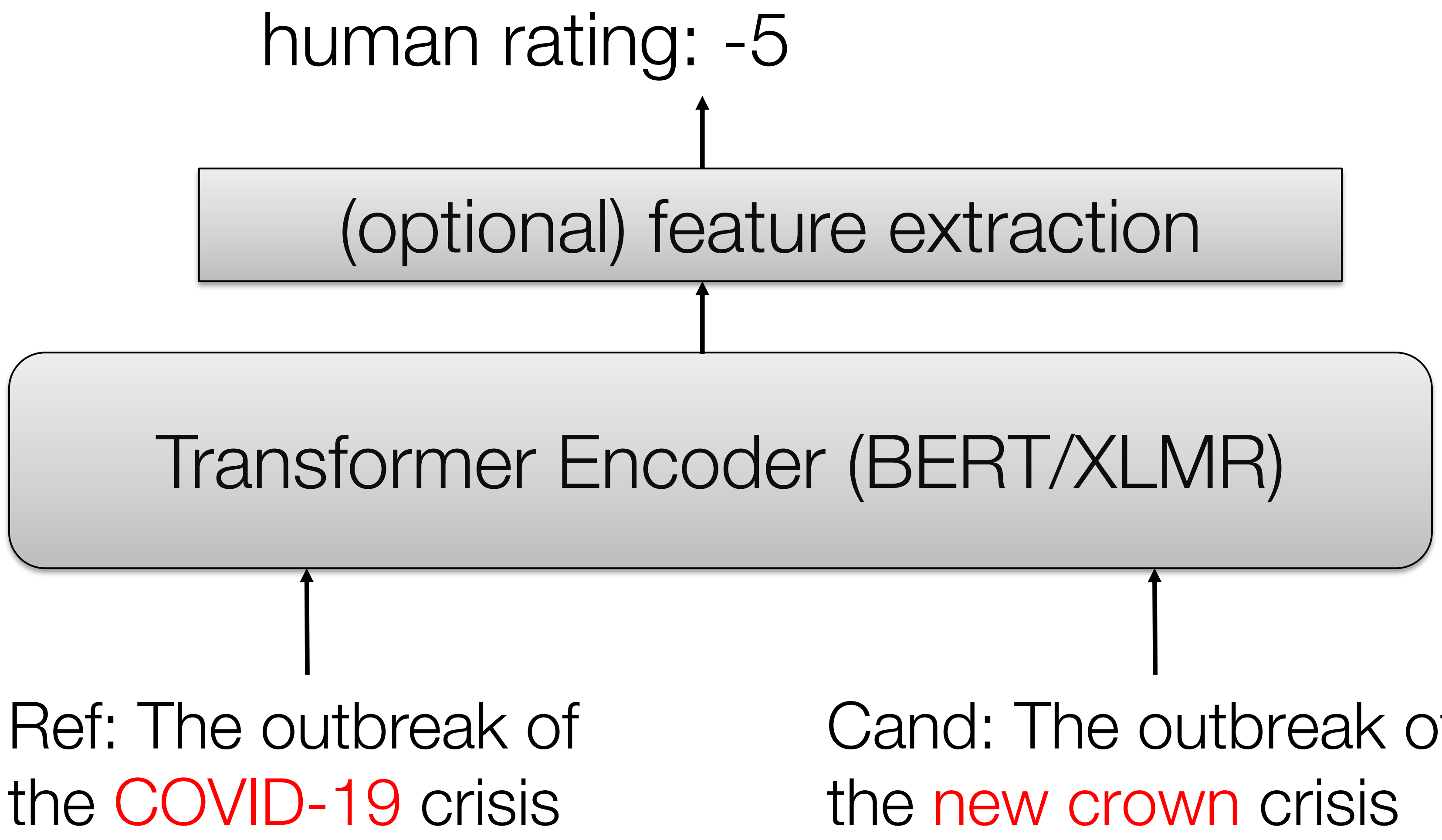
BertScore: 0.925

COMET: 0.711

BLEURT: 0.519

SEScore2: -5.43

Training Reference-based Metrics



COMET
BLEURT
SEScore
SEScore2

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".

Why is training an explainable metric challenging?

- Data Scarcity
- Indirect training objective (Not regression anymore)
- Well Defined Explainability

Ideal Metric

- Highly Aligned with Expert Annotator
- Fine-grained Explainability
- Generalizable

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:

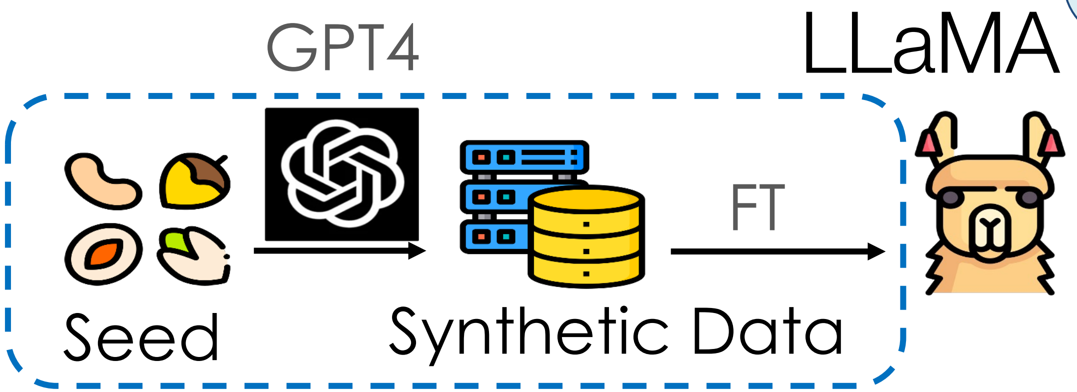
[GPT4 fill in]

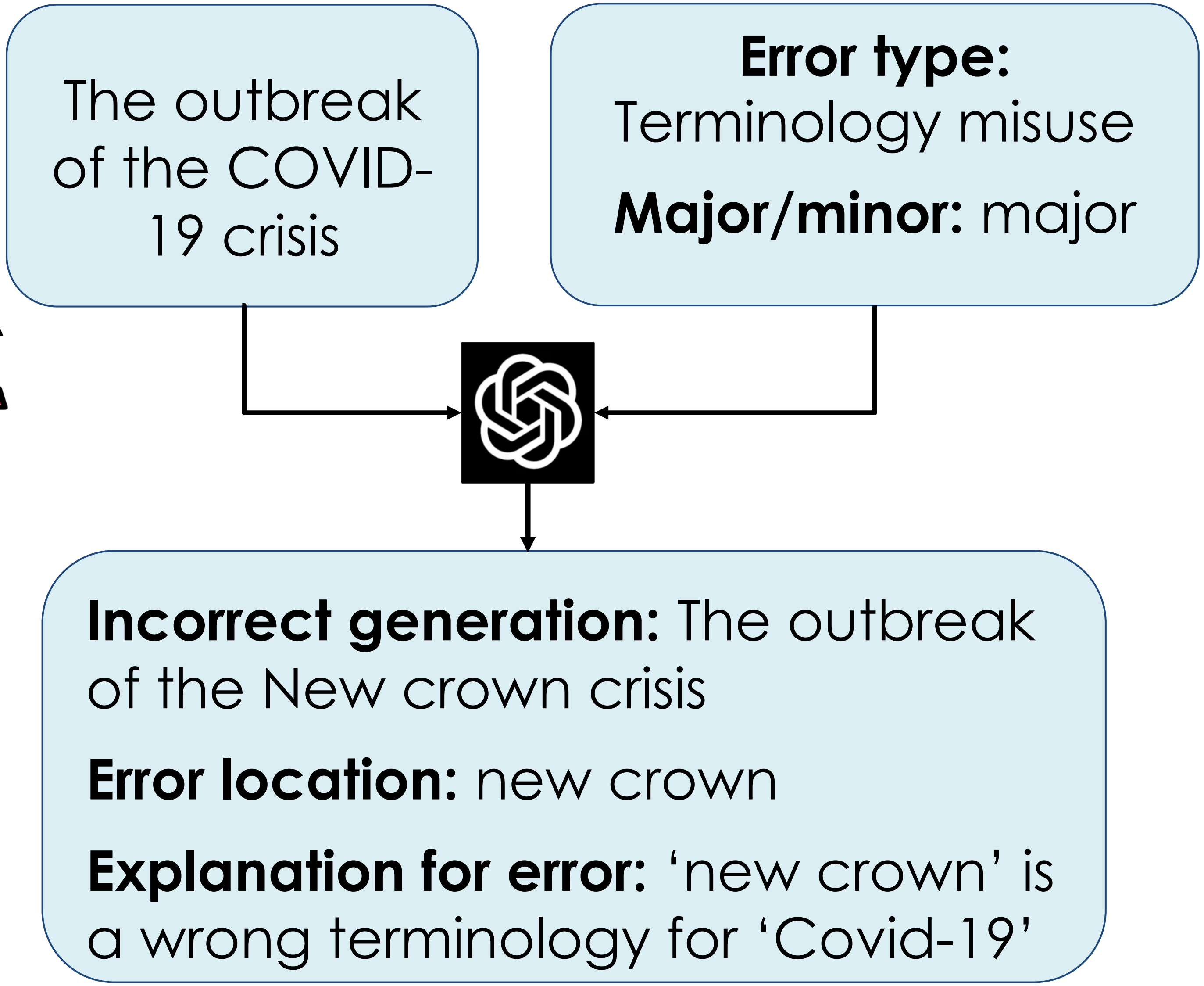
Error location 1: [GPT4 fill in]

Explanation for error 1:

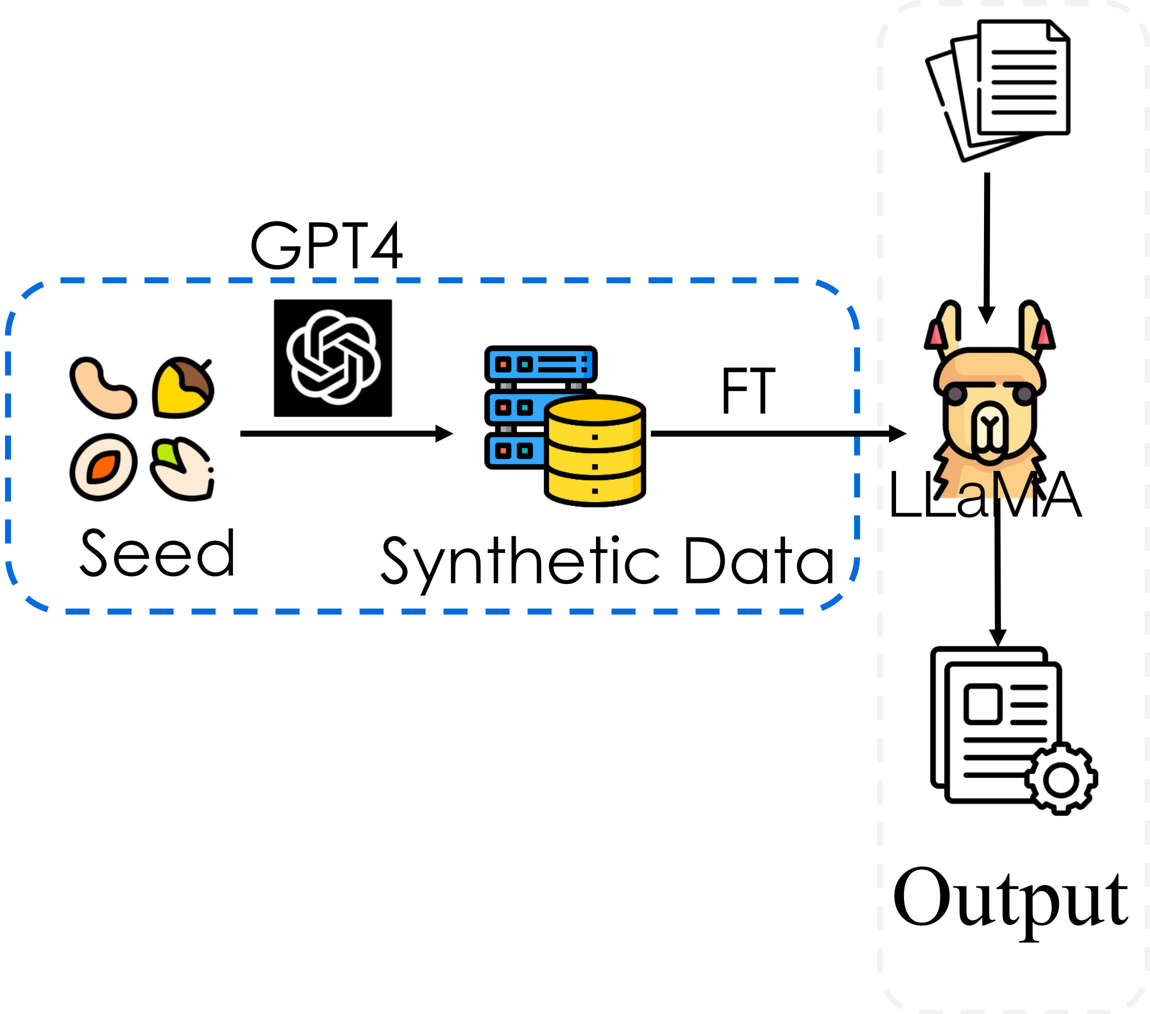
[GPT4 fill in]

Using synthetic data from Direct Prompting





But, failed explanation in GPT4

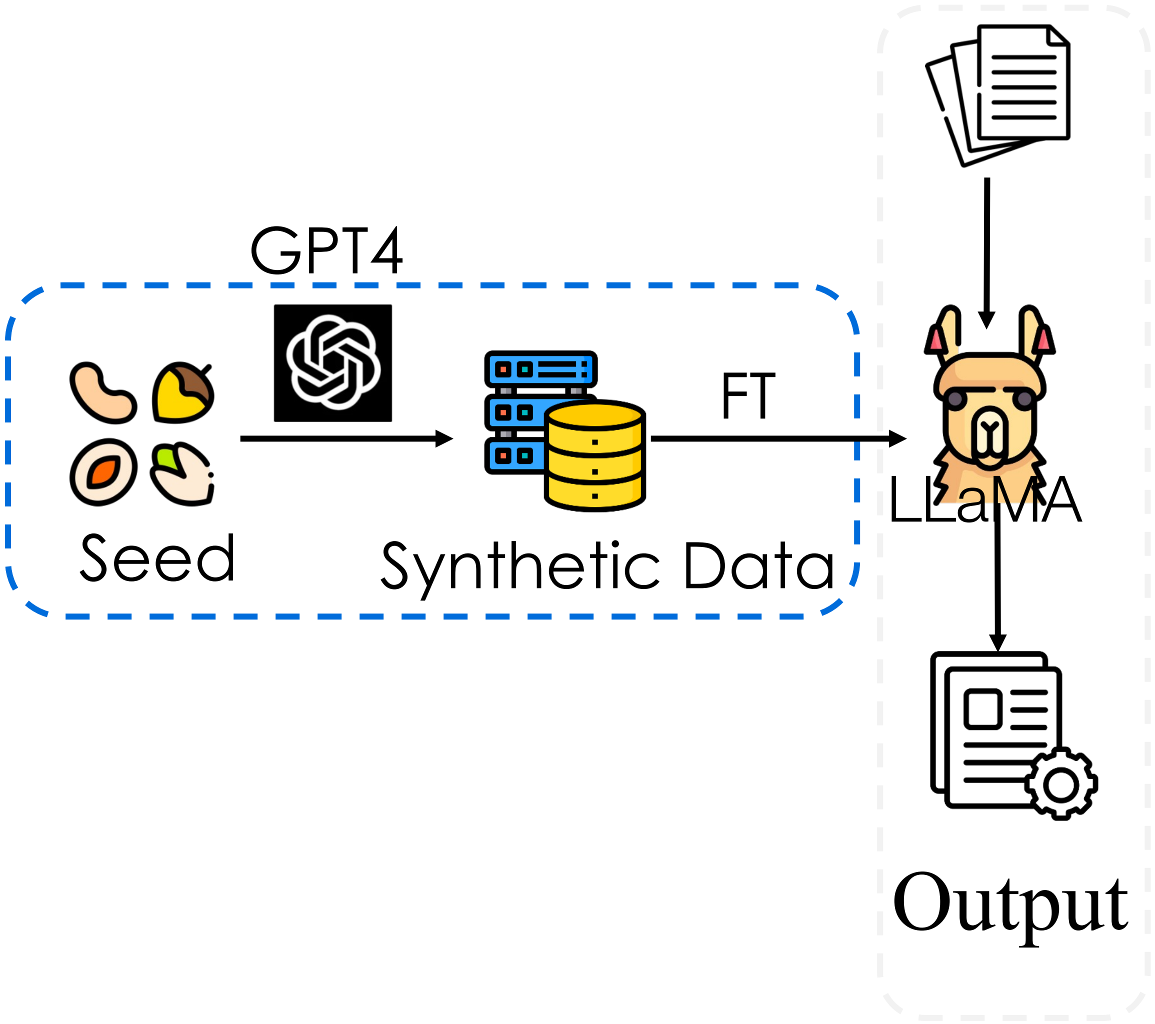


Error type 3: Missing information

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

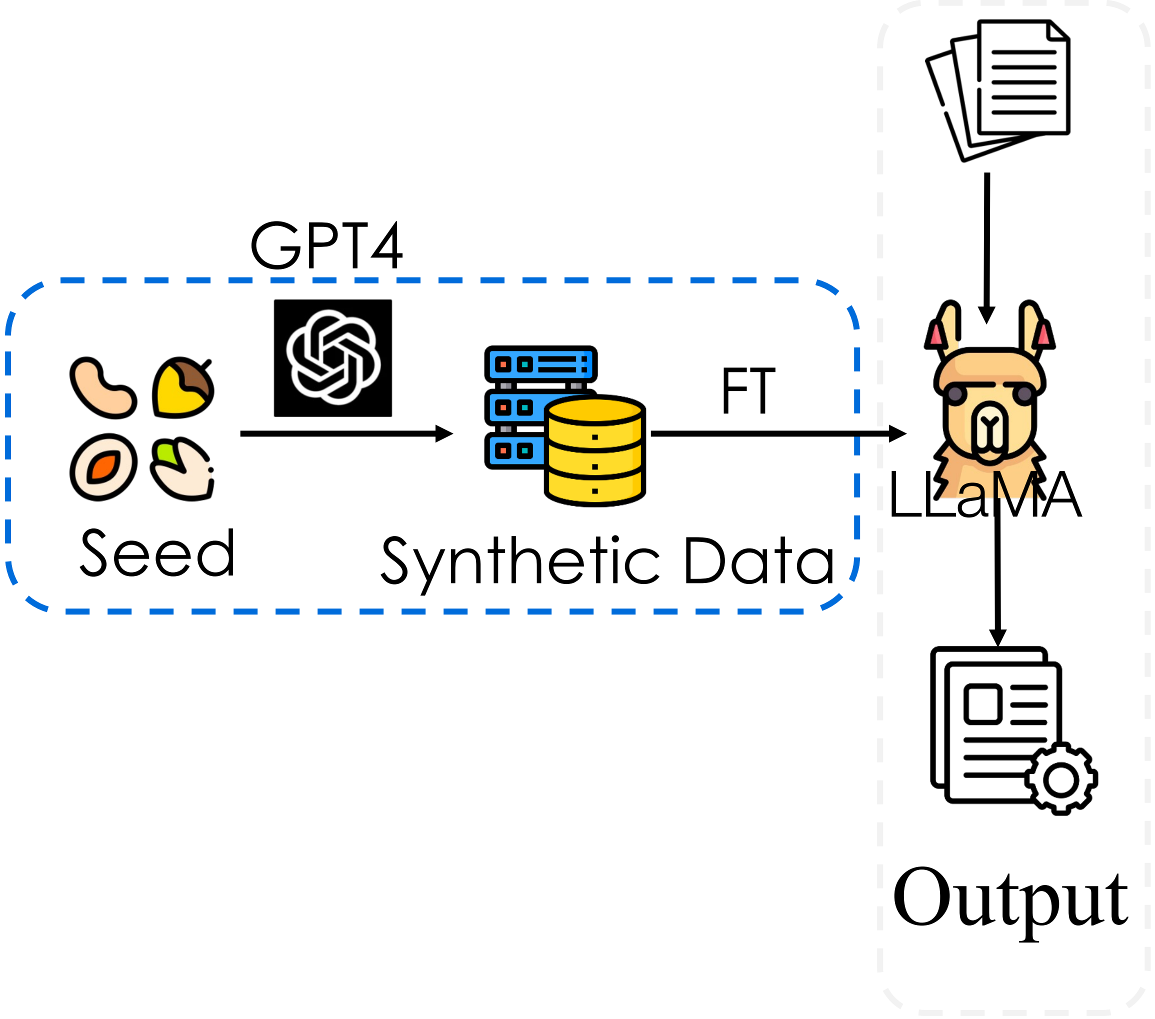


Evaluated text: The outbreak of the new crown crisis

Error location: 'virus'

Hallucination

But, failed explanation in GPT4



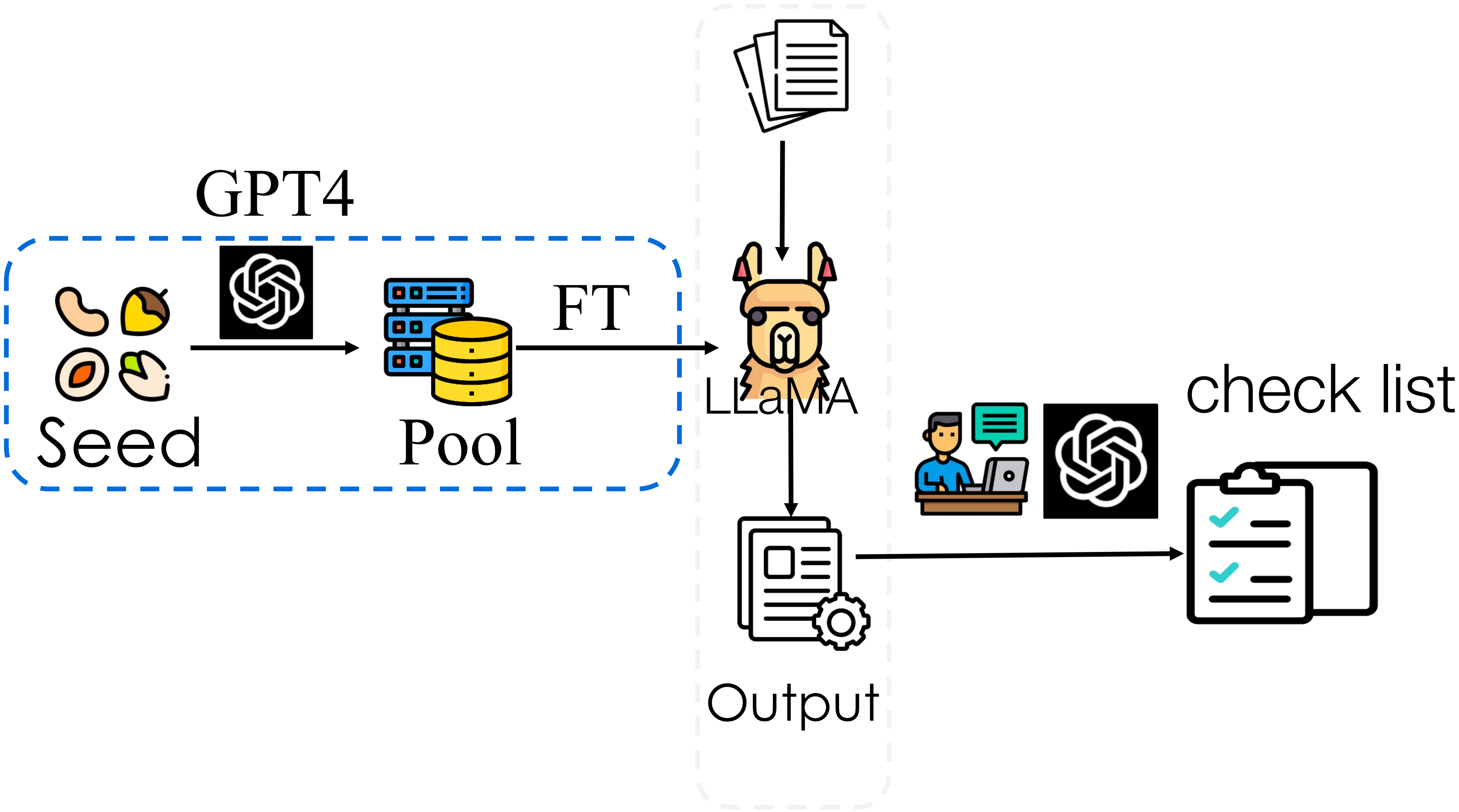
Explanation for error 1: The incorrect translation uses the word "annual" instead of "annual"

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

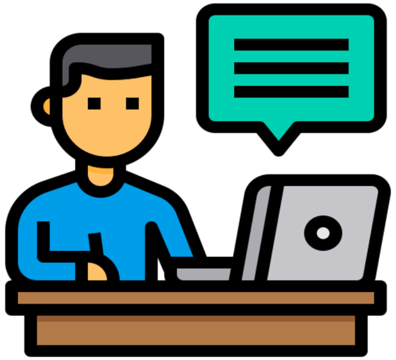
Introducing InstructScore



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

Use GPT-4 as a checking Model

Human defines all failure modes



Formulate them into a checklist

Perform checklist by asking GPT4 to perform simpler tasks (QA, information extraction etc)

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?

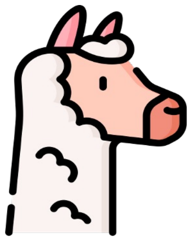
Are two phrase aligned?

InstructScore: Automatic Feedback

Reference
Candidate

Error location1
Error Type1
Major/Minor
Explanation1

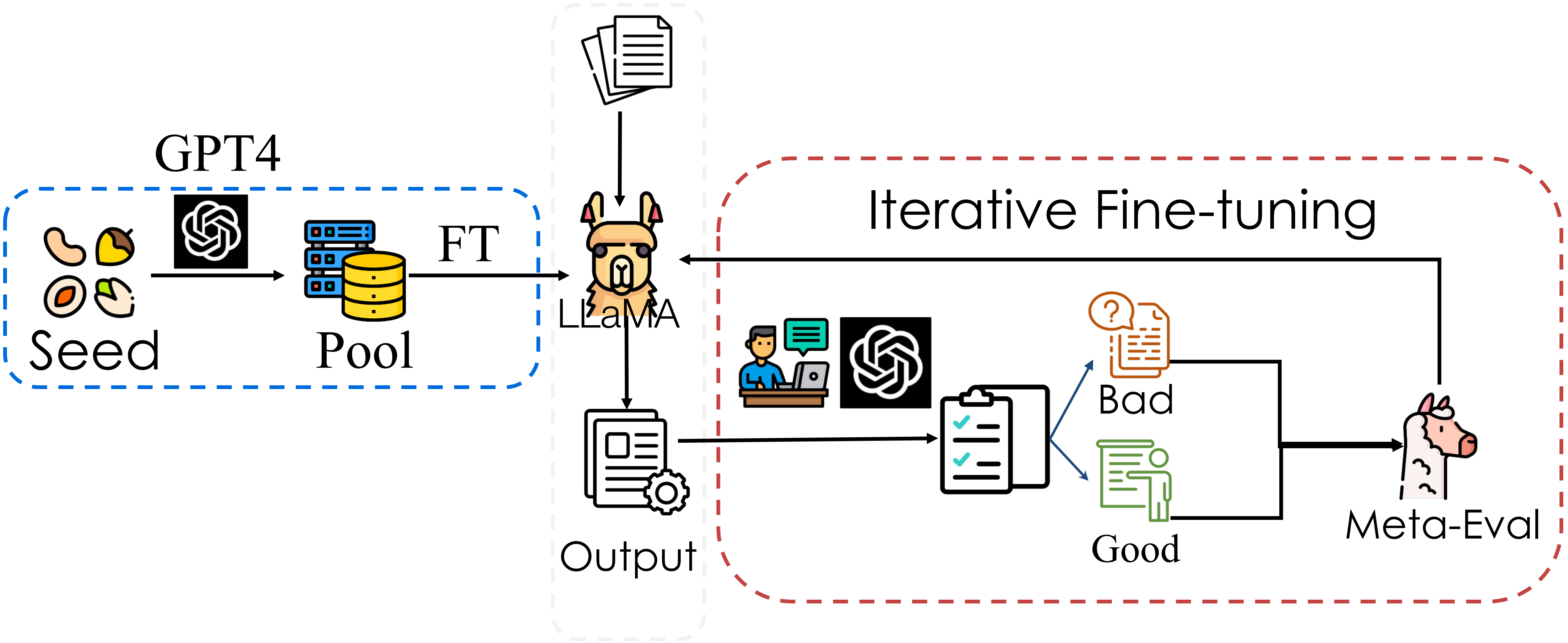
Error location2
Error Type2
Major/Minor
Explanation2



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

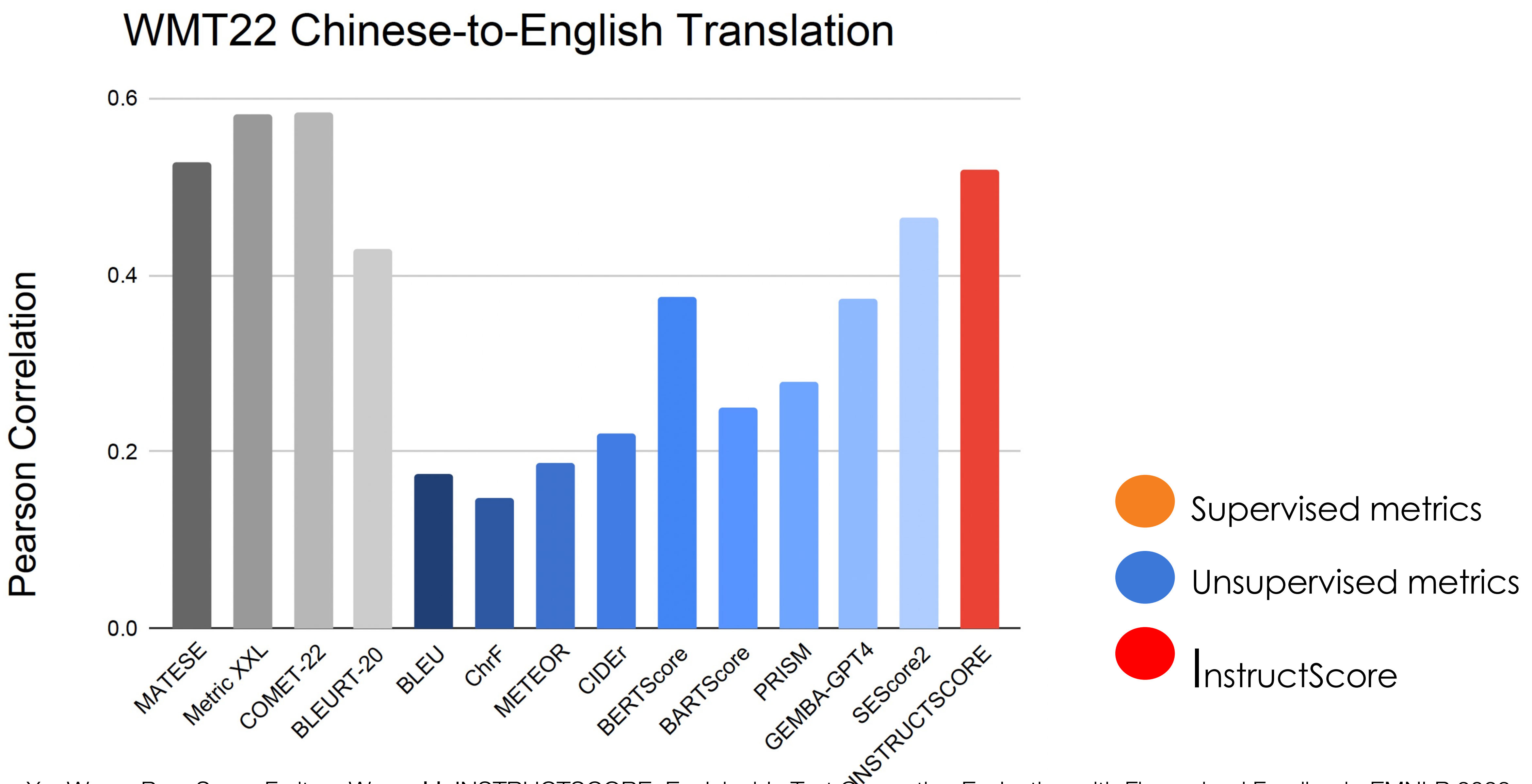
Alignment Score: 7/8

InstructScore: Refinement



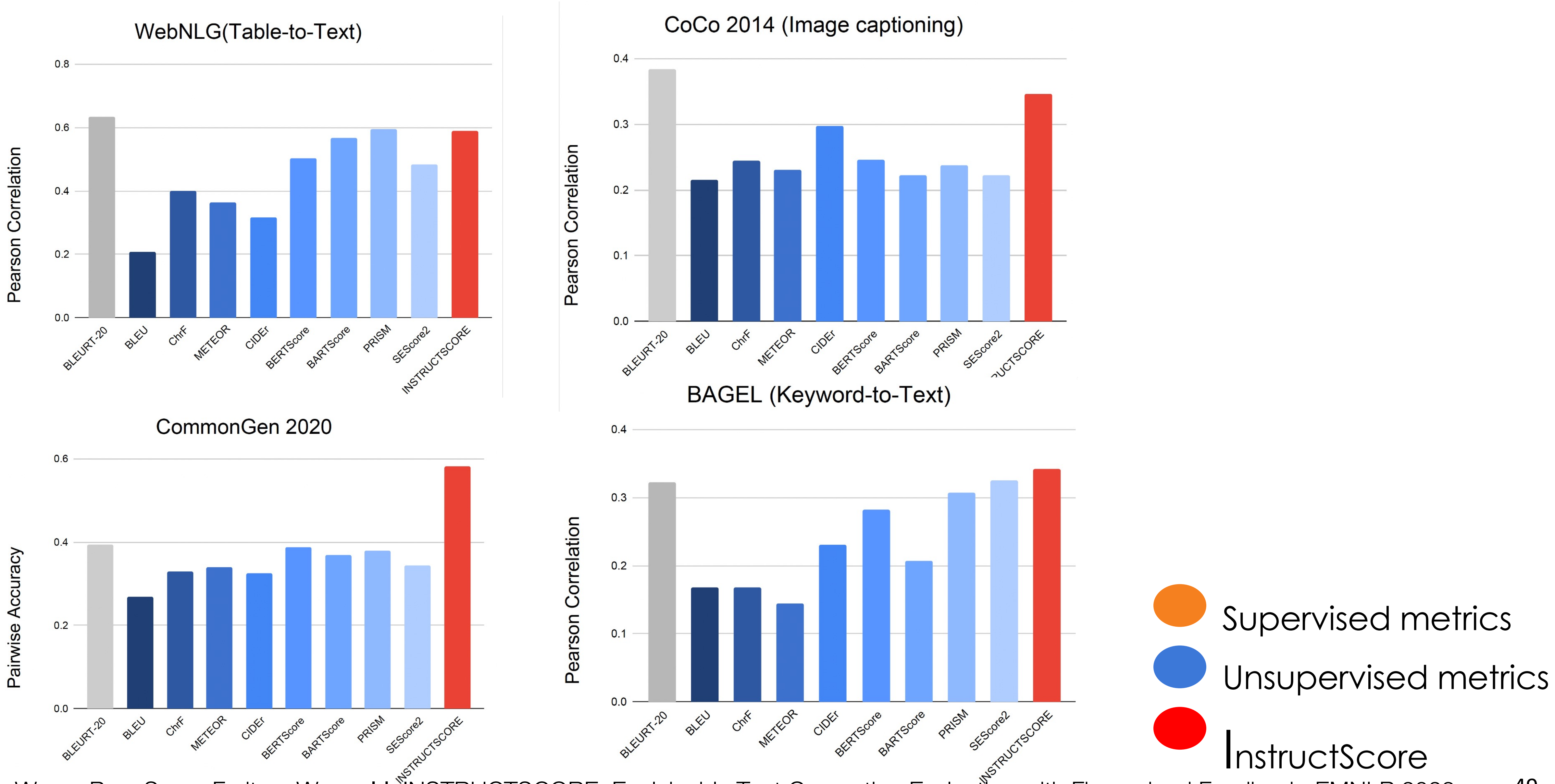
Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

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 evaluator?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - Interpretable text generation evaluation (InstructScore)
  - Assessing knowledge in LLMs (KaRR)
- Post-training alignment
  - Online Preference Optimization (BPO)
  - Iterative refinement with fine-grained feedback (LLMRefine)

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.

   $\boxed{\text{William Shakespeare's}}$ job is a writer.

   $\boxed{\text{John Smith's job}}$ is a writer.

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

Why Do We Need Knowledge Assessment?

- The assessment results directly affect the people's trust in the LLM generated content.

- Once we identify inconsistency of LLM generation, we could potentially correct such knowledge in LLMs[1].

[1] *Nicola De Cao, Wilker Aziz, and Ivan Titov. Editing factual knowledge in language models. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021.*

Risk Ratio

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

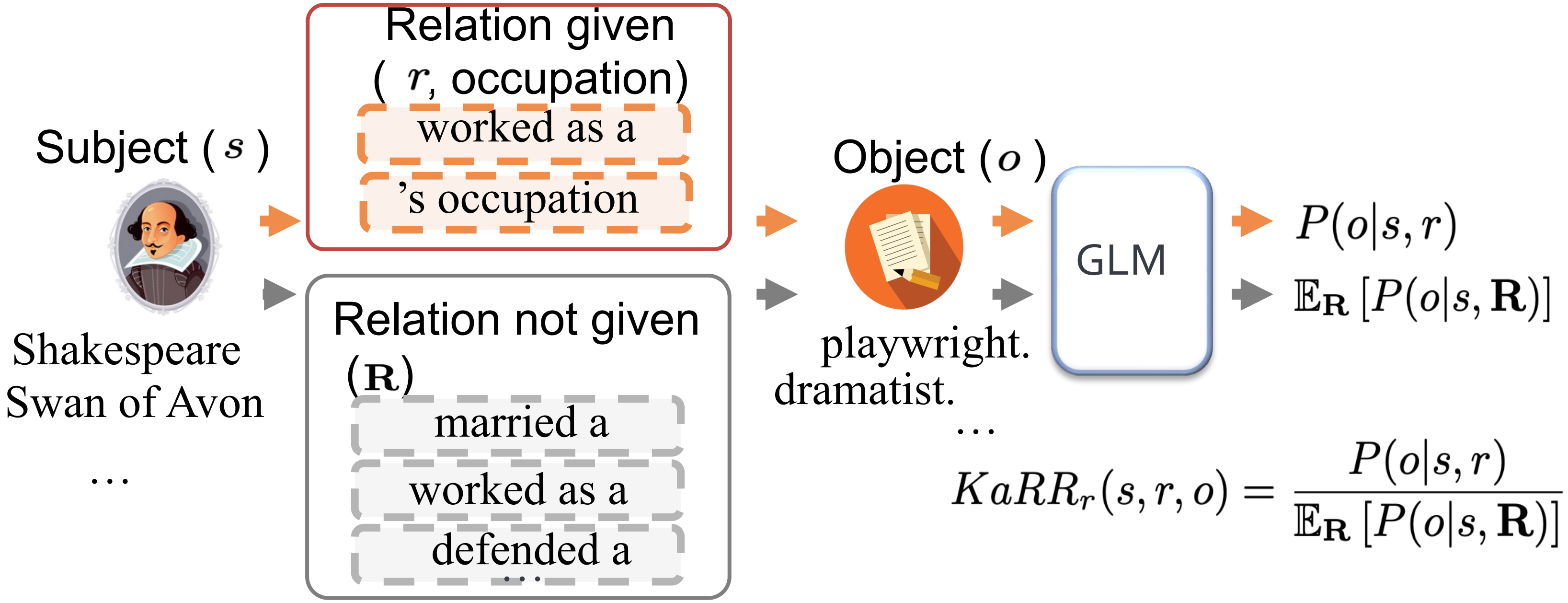
- Example: a disease noted by D, and no disease noted by $\neg D$, exposure noted by E, and no exposure noted by $\neg E$. The risk ratio can be written as:

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

	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)

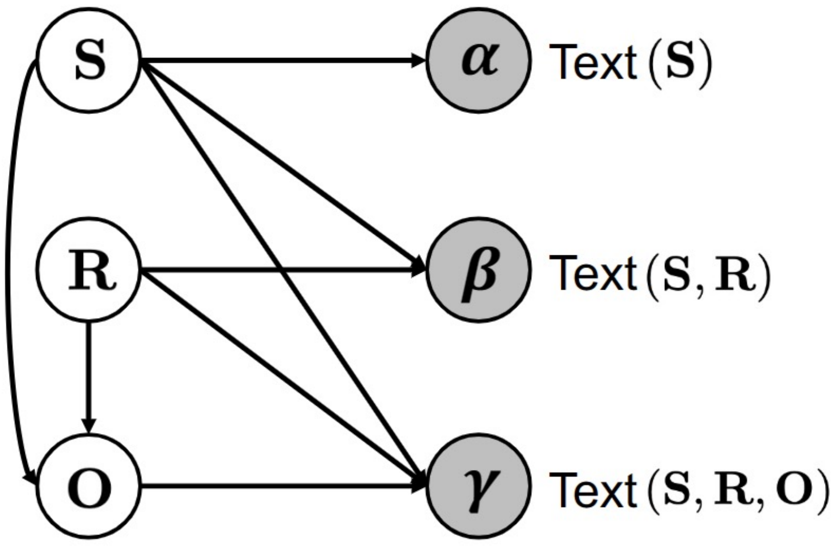
- Assesses the joint impact of subject and relation symbols on the LLM's ability to generate the object symbol.



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

KaRR via graphical model

To evaluate LLM knowledge reliably, we decompose the knowledge symbols and text forms.



hollow circles: latent variables
shaded circles: observed variables

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

$$P(o \mid s, r) = \sum_{k=1}^{|\beta|} P(o, \beta_k \mid s, r)$$

$$= \sum_{k=1}^{|\beta|} P(\beta_k \mid s, r) \cdot P(o \mid s, r, \beta_k)$$

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

Results of Human Assessment

- Human annotation:

  1) Annotating: 3 annotators each write 3 prompts to probe the model knowledge, refine the prompts based on the generations until the generations are aliases of the target answer.

  2) Rating: another 3 annotators to rate the knowledge (0 or 1) in model according to the generations.

Method	Recall	Kendall's τ	p-value
LAMA@1	83.25%	0.17	0.10
LAMA@10	65.81%	0.08	0.23
ParaRel	69.15%	0.22	0.02
K-Prompts	78.00 %	0.32	0.03
KaRR	**95.18%**	**0.43**	0.03

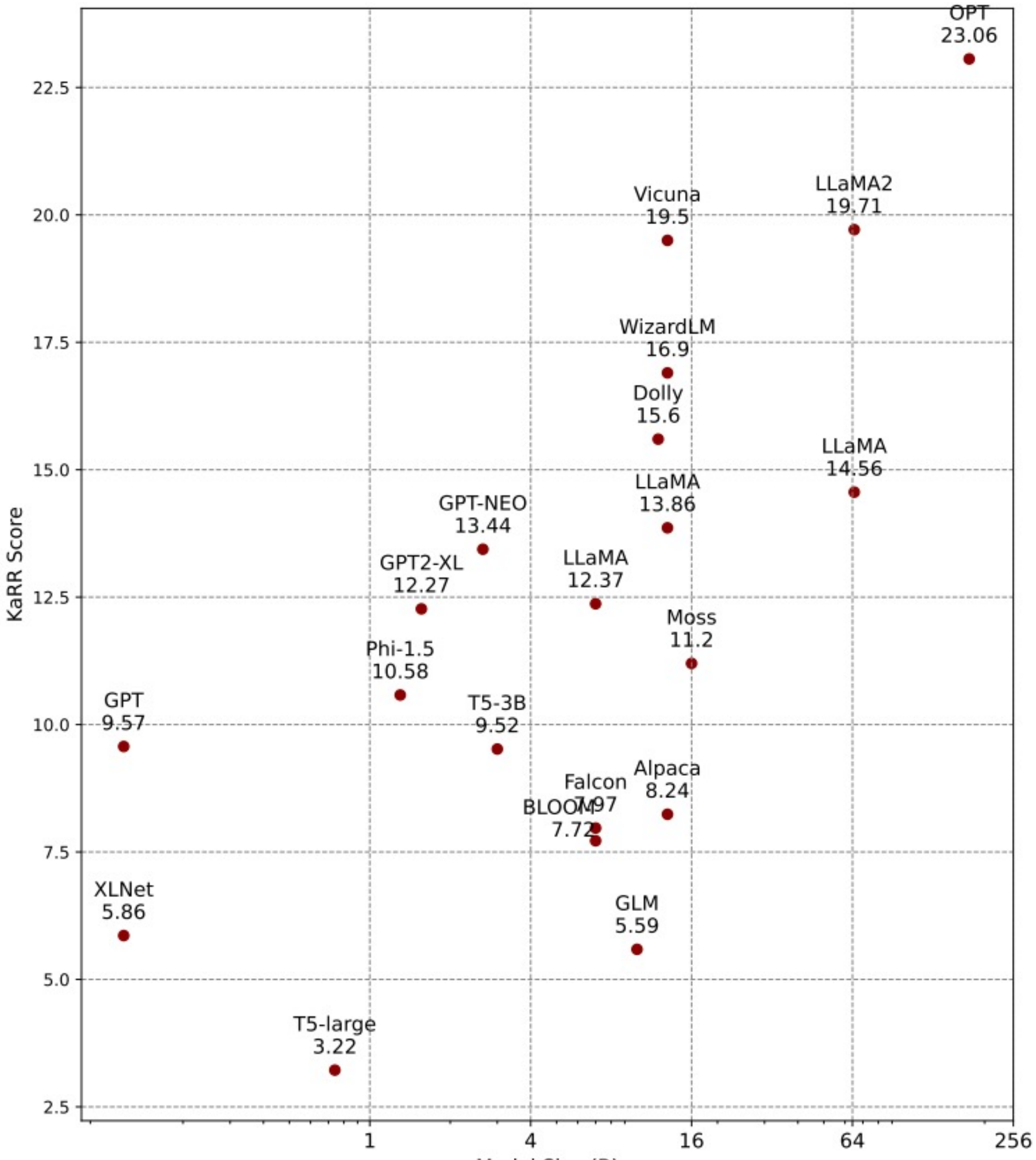
We calculate the Kendall tau correlation between scores fr om various methods and hu man evaluation rankings for f actual knowledge.

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

KaRR Scores for 20 LLMs

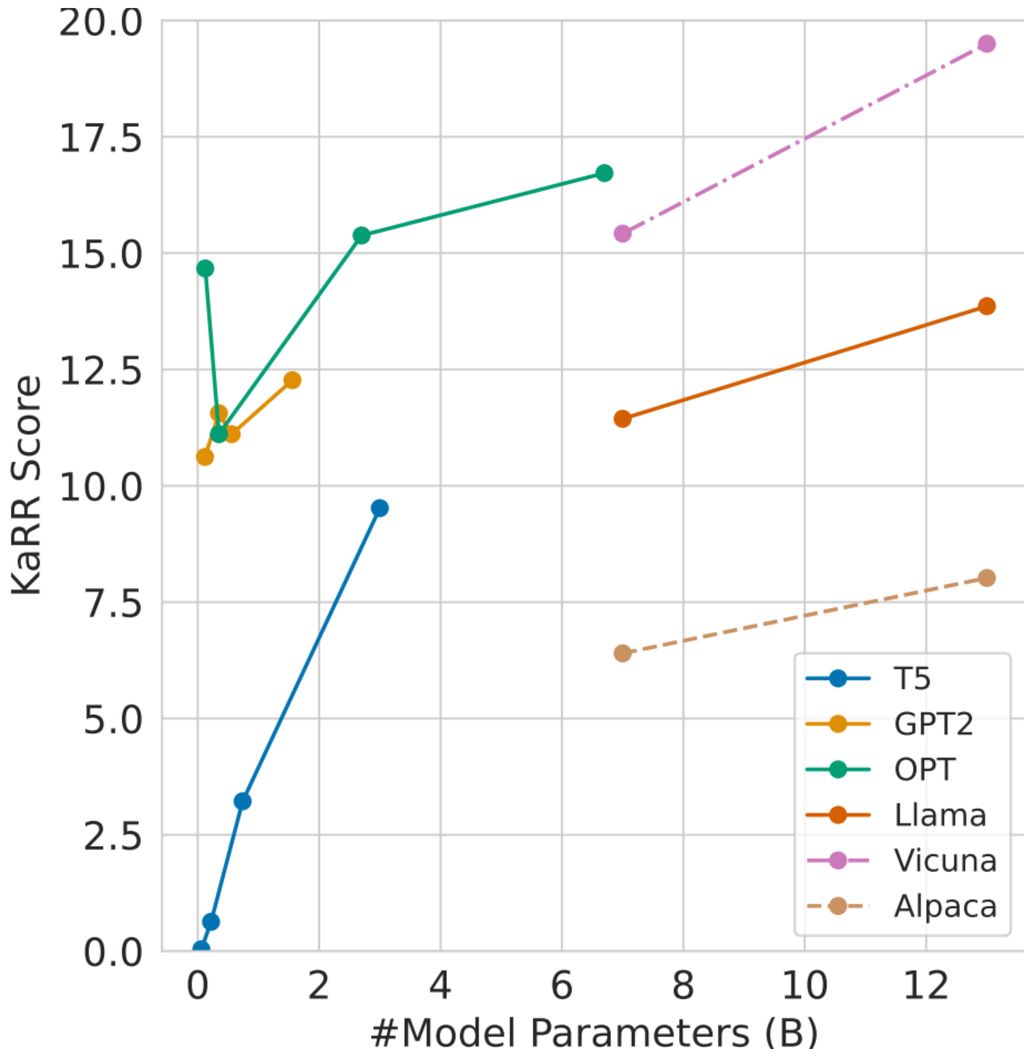
- Small and medium-sized LLMs struggle with generating correct facts consistently.

- Finetuning LLMs with data from more knowledgeable models can enhance knowledge.



Scaling Effect on Knowledge

- larger models generally hold more factual knowledge.
- Scaling benefits vary among models. E.g., T5-small to T5-3B.

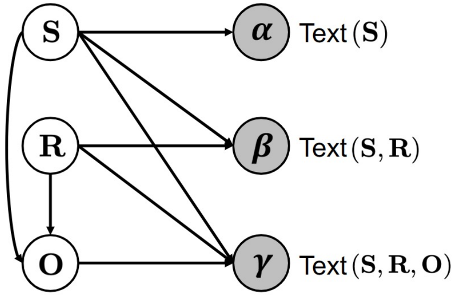


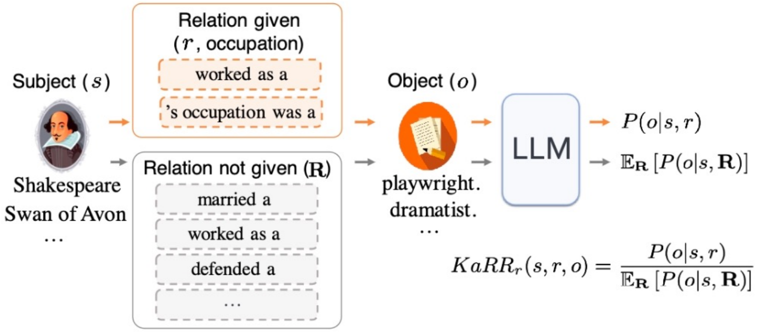
Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Zhifang Sui, Lei Li. Statistical Knowledge Assessment for LLMs. Neurips 2023

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)](https://github.com/dqxiu/KAssess)





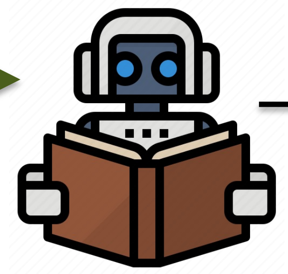
Outline

- Can we trust LLM evaluator?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - Interpretable text generation evaluation (InstructScore)
  - Assessing knowledge in LLMs (KaRR)
- Post-training alignment
  - Online Preference Optimization (BPO)
  - Iterative refinement with fine-grained feedback (LLMRefine)

Learning from Human Feedback

SFTed LLM

Question: Why is the sky blue?



The sky appears blue because …

The sky is not always blue …

Preference annotation by human

The sky appears blue because …

The sky is not always blue …

$$(x, y_w, y_l)$$

Preferred Dispreferred

Reward modeling in RLHF

$(x, y_w, y_l) \longrightarrow$ $\boxed{\text{Reward Model}}$

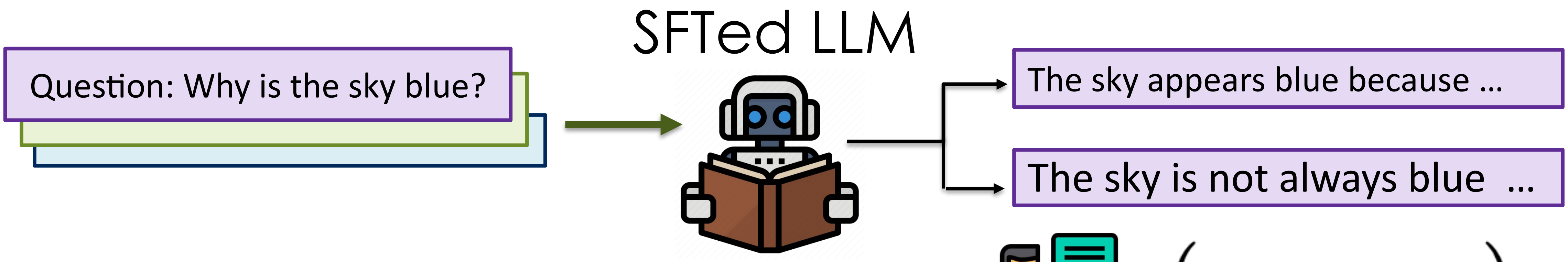
$$p^*(y_1 \succ y_2 \mid x) = \frac{\exp\left(r^*(x, y_1)\right)}{\exp\left(r^*(x, y_1)\right) + \exp\left(r^*(x, y_2)\right)}.$$

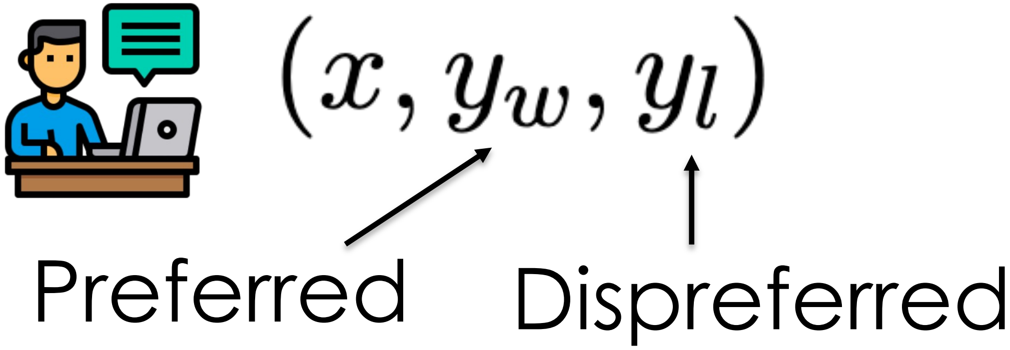
Bradley-Terry Model

$$\mathcal{L}_R(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}}\left[\log \sigma(r_\phi(x, y_w) - r_\phi(x, y_l))\right]$$

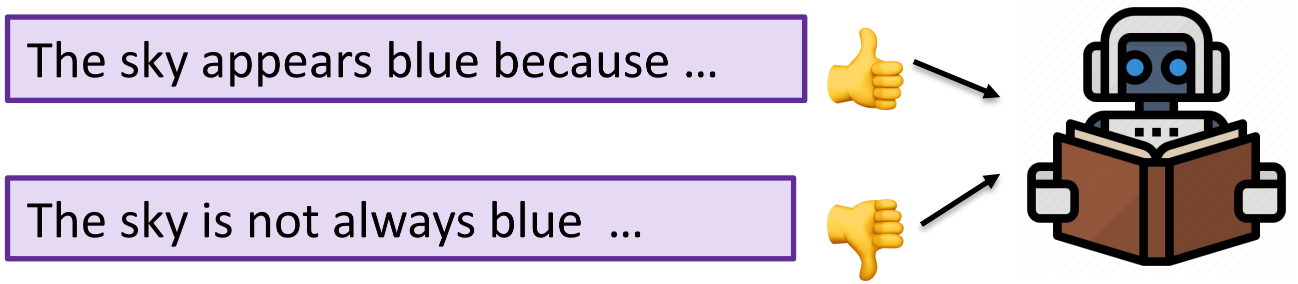
Training language models to follow instructions with human feedback

Direct Preference Optimization





We can skip reward model using DPO



Offline DPO variants

All DPO variants follow this

$$r_\phi(y_w) - r_\phi(y_l) = \beta\left(\log\frac{\pi^*_\theta(y_w)}{\pi_{\text{ref}}(y_w)} - \log\frac{\pi^*_\theta(y_l)}{\pi_{\text{ref}}(y_l)}\right).$$

DPO loss:

$$-\log\sigma\left(\beta\log\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^-|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^-|\boldsymbol{x})}\right)$$

IPO loss:

$$\left(\log\left(\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^-|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^-|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^+|\boldsymbol{x})}\right) - \frac{1}{2\beta}\right)^2$$

Avoids the overfitting from DPO (Squared loss)

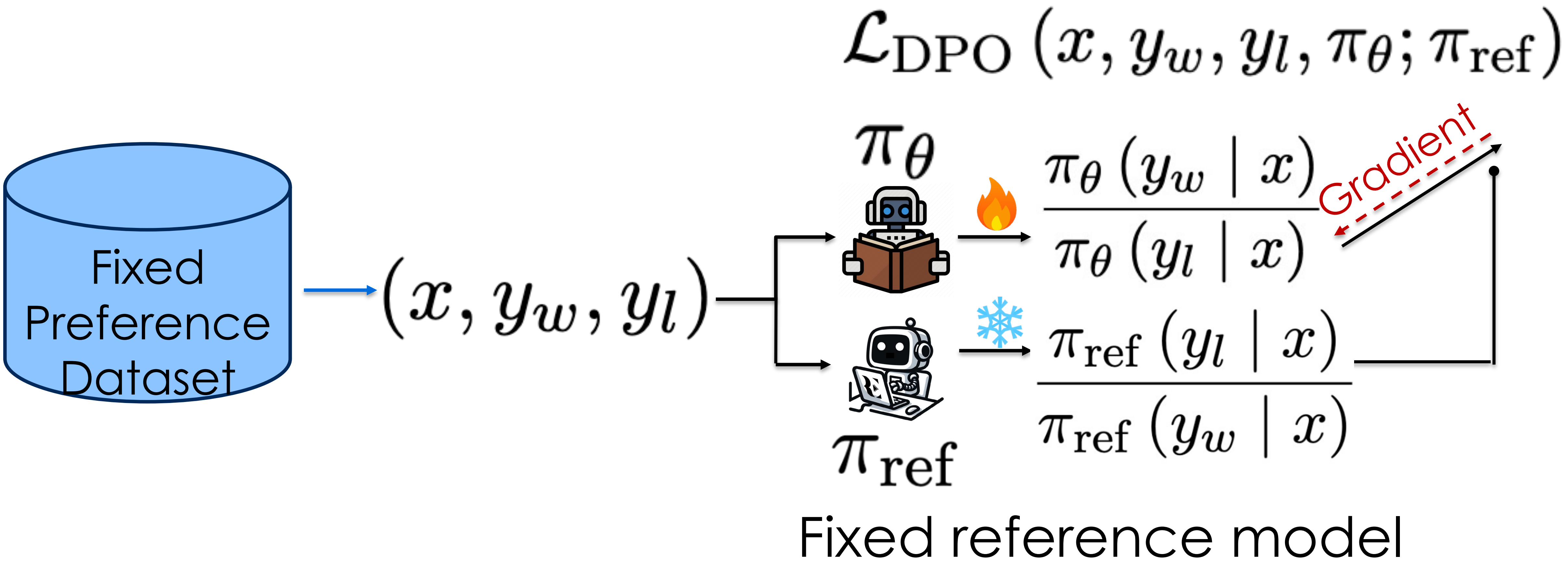
SLiC loss:

$$\max\left(0, 1 - \beta\log\left(\frac{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^+|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^-|\boldsymbol{x})}{\pi_{\boldsymbol{\theta}}(\boldsymbol{y}^-|\boldsymbol{x})\pi_{\boldsymbol{\theta}^0}(\boldsymbol{y}^+|\boldsymbol{x})}\right)\right)$$

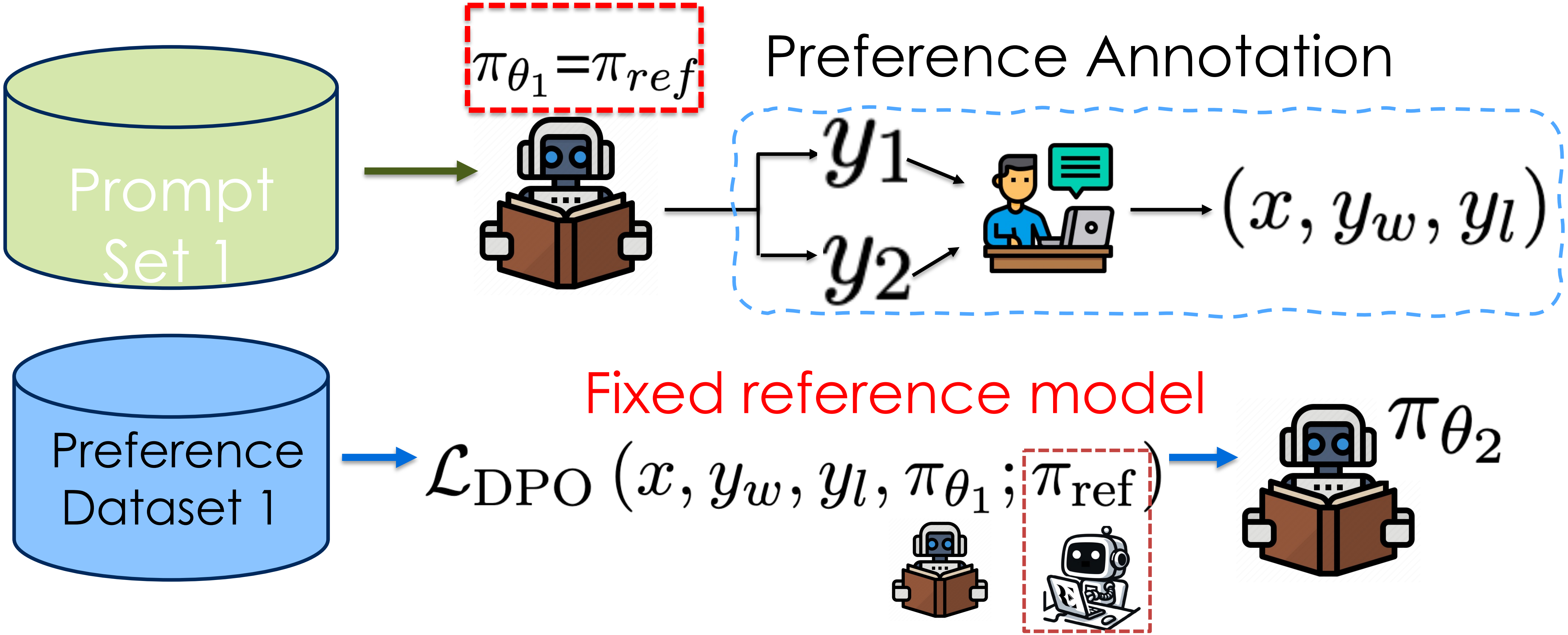
Hinge loss

Generalized Preference Optimization: A Unified Approach to Offline Alignment

Illustration of DPO



Limitation of offline DPO (and online DPO)

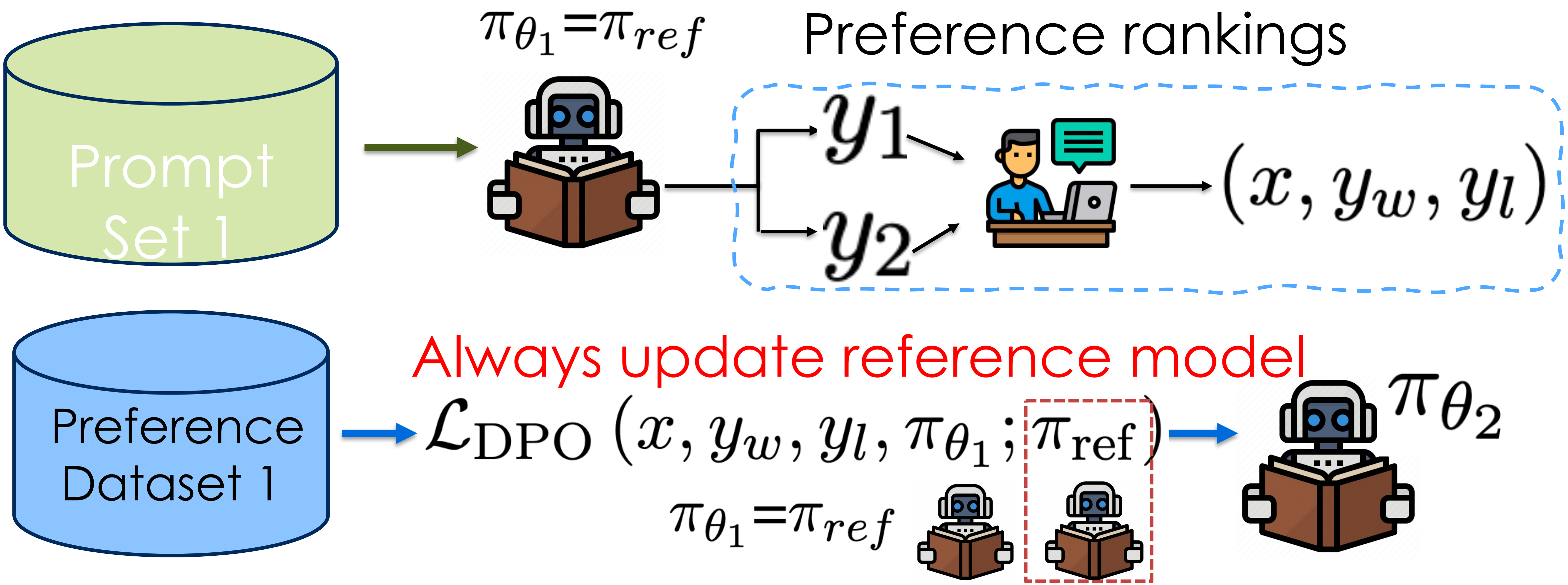


New Algorithm: BPO (B=Behavior)

- Data collection needs to be online
- The reference model needs to be updated and has to be close to the behavior LLM

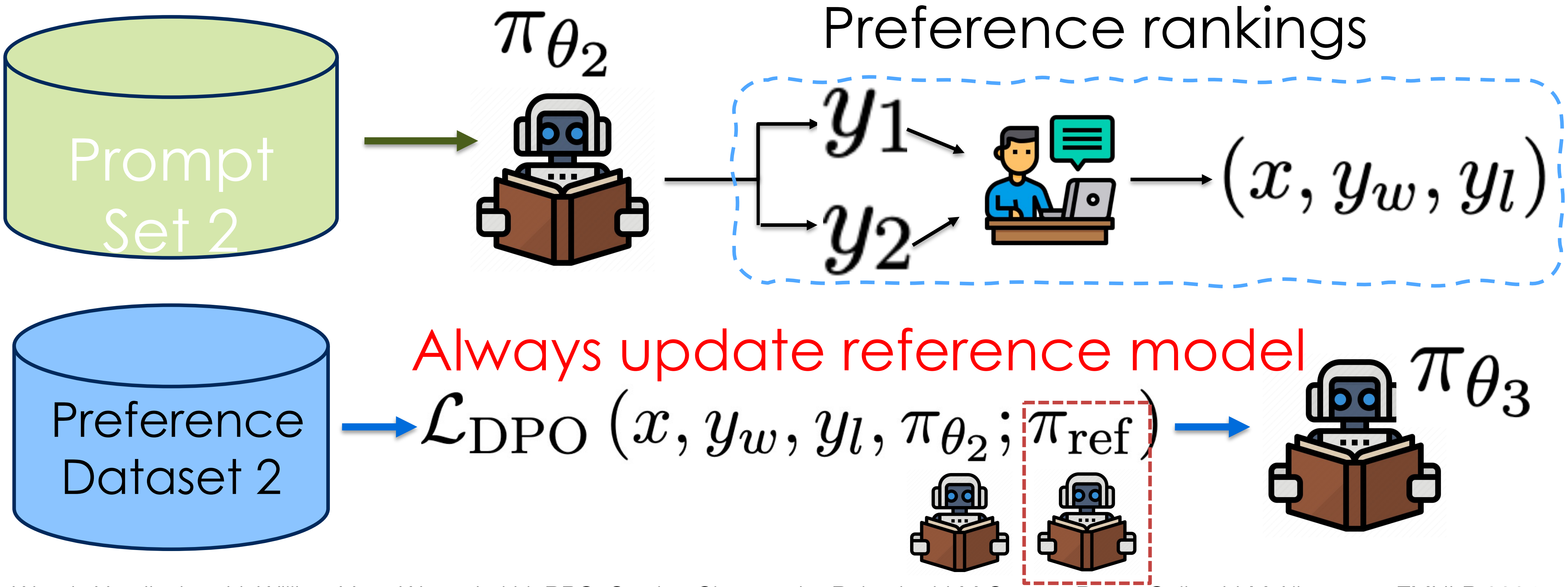
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO



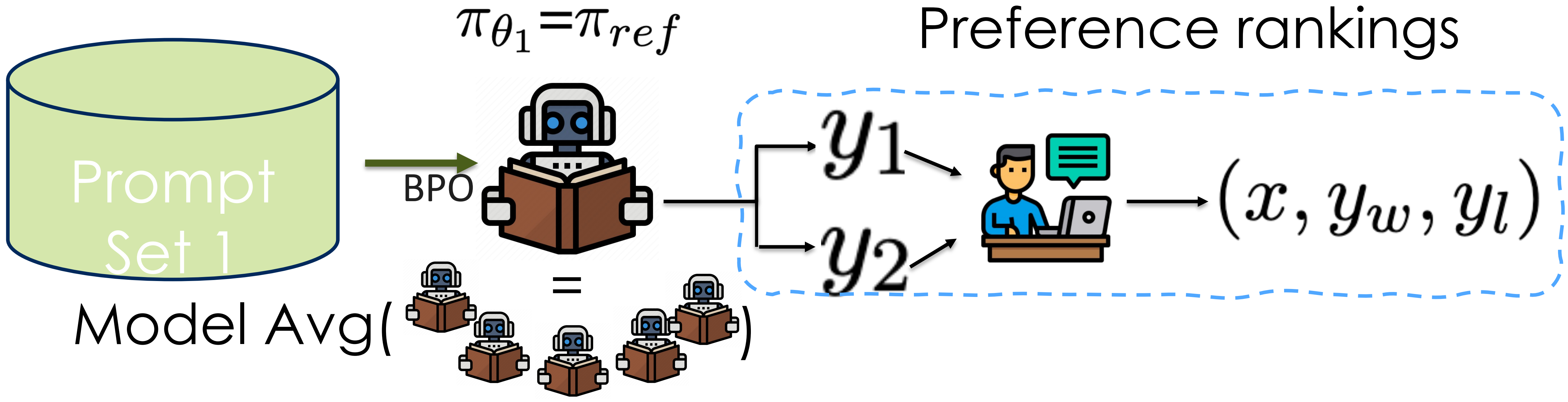
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO



Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

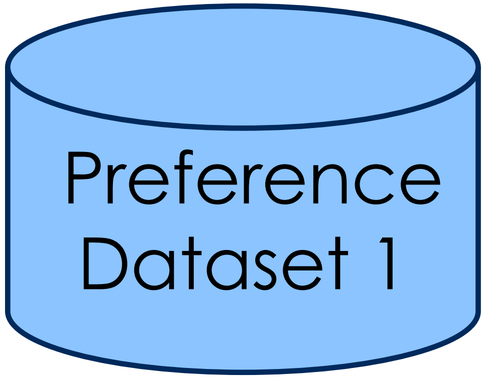
Practical implementation of BPO (Lora ensemble)



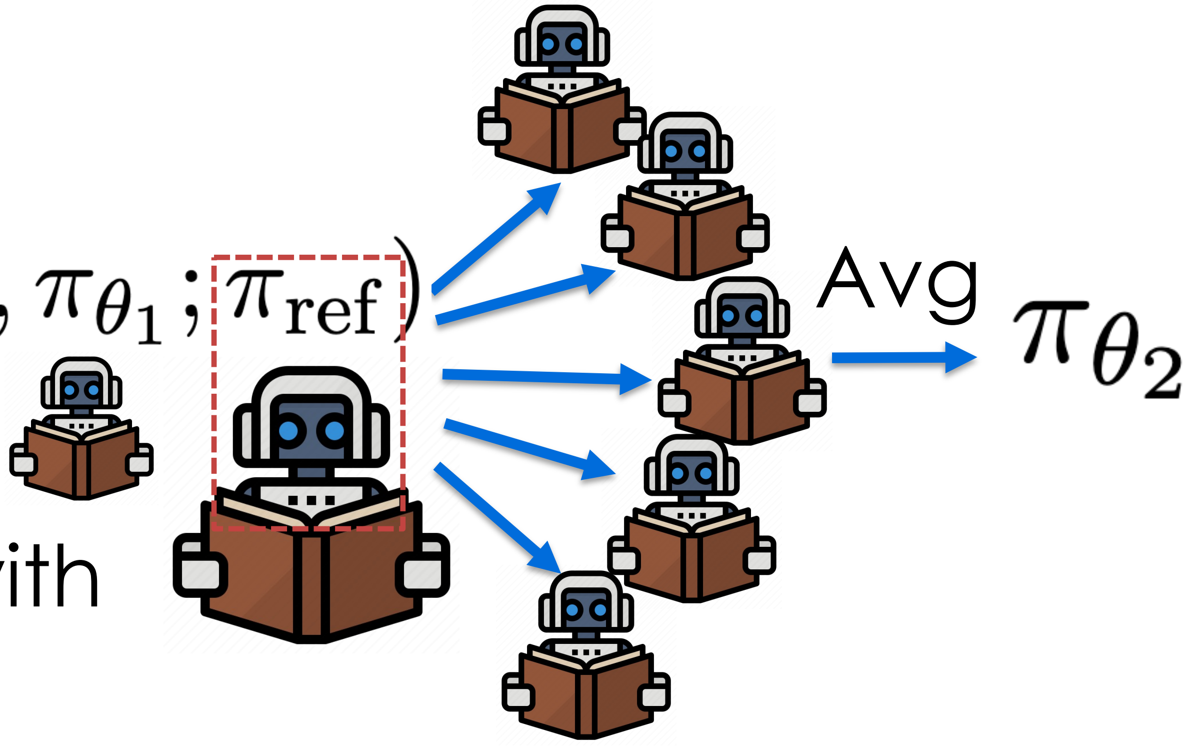
We use model averaged lora weights to perform sampling

Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

Practical implementation of BPO (Lora ensemble)



$$\mathcal{L}_{\mathrm{DPO}}\left(x, y_w, y_l, \pi_{\theta_1}; \pi_{\mathrm{ref}}\right)$$

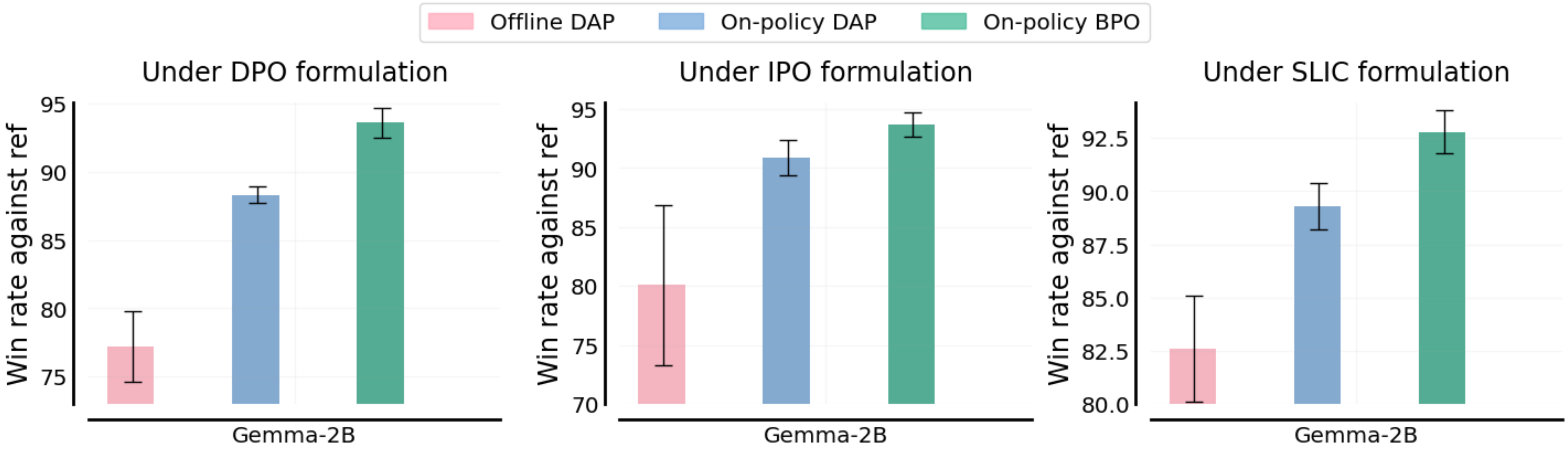


We update reference model with Model averaged behavior LLM

Each lora weight is updated independently

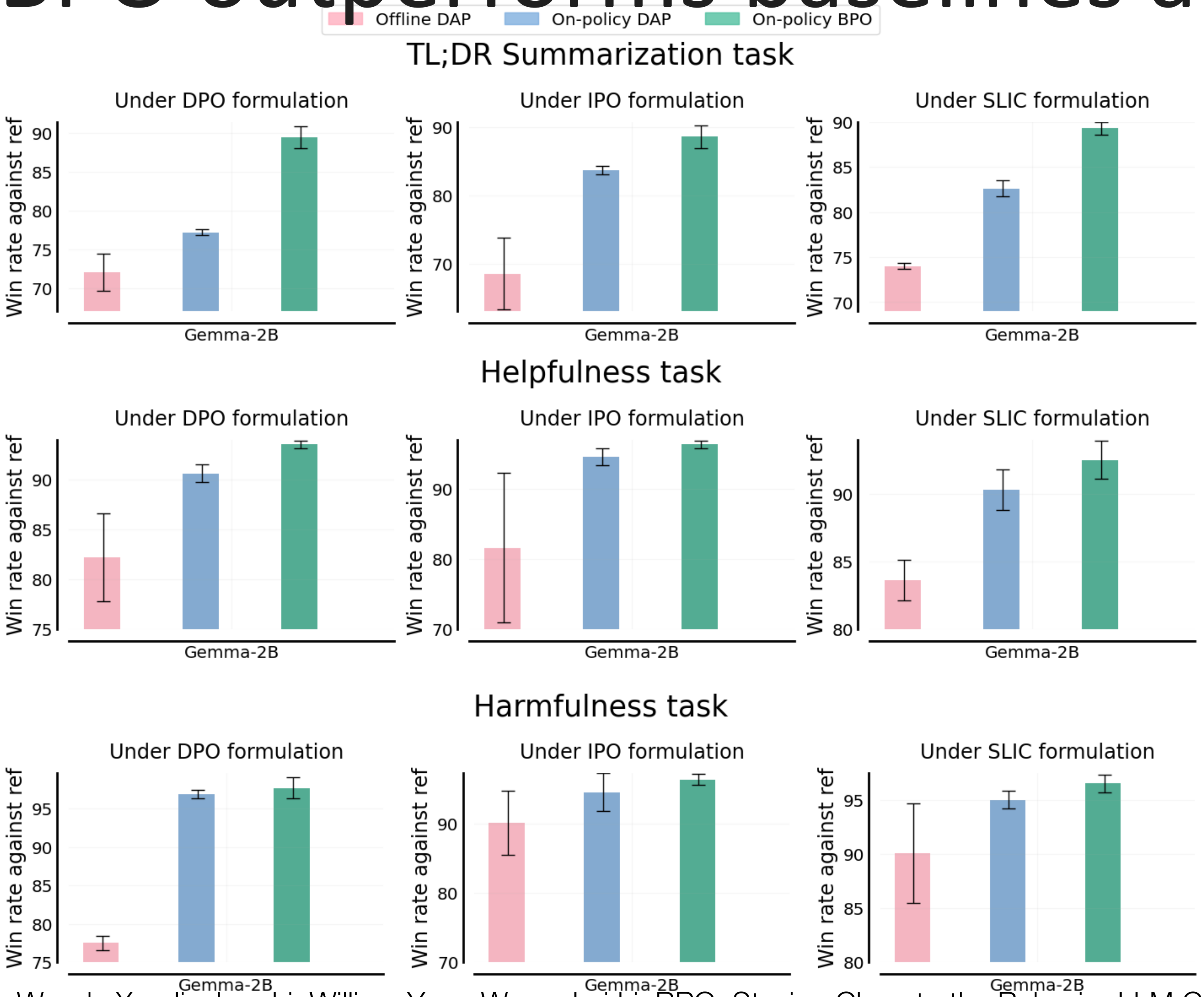
Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO outperforms online and offline alignment methods



Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO outperforms baselines across three tasks



Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

BPO Highlight

- Reference model should stay close to the behavior LLM and create better online LLM alignment

- Practical applicability: We empirically show our online BPO with >=2 data collection steps can significantly improve offline baselines

- The effectiveness of BPO stems from proximity to the behavior model, rather than improvements in the reference model's quality.

Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.

Outline

- Can we trust LLM evaluator?
    - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
    - Interpretable text generation evaluation (InstructScore)
    - Assessing knowledge in LLMs (KaRR)
- Post-training alignment
    - Online Preference Optimization (BPO)
    - Iterative refinement with fine-grained feedback (LLMRefine)

Can we use fine-grained feedback to guide LLM?

Input: Translate " 新冠疫情危机爆发 " into English.

LLM's output:

the outbreak of the new crown crisis

What feedback can we give to LLM?

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM's output:

the outbreak of the new crown crisis

Ask LLM to improve?

Source:新冠疫情危机爆发

Translation: the outbreak of the new crown crisis

Please Improve current translation.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM's output:

the outbreak of the new crown crisis

Use binary feedback to guide LLM?

Source:新冠疫情危机爆发
Translation: the outbreak of the new crown crisis
Your translation contains errors. Please improve current translation.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM's output:

the outbreak of the new crown crisis

Use scalar feedback to guide LLM?

Source:新冠疫情危机爆发

Translation: the outbreak of the new crown crisis

Your translation has score of 70/100. Please improve current translation.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2023. Iterative translation refinement with large language models.

Can we use fine-grained feedback to guide LLM?

Input: Translate "新冠疫情危机爆发" into English.

LLM's output:
the outbreak of the new crown crisis

Use fine-grained feedback to guide LLM!

Source:新冠疫情危机爆发
Translation: the outbreak of the new crown crisis
" new crown" is a major terminology error. Please improve current translation.

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

When can we accept refined proposal?

Source:新冠疫情危机爆发
Translation: the outbreak of the new crown crisis
" new crown" is a major terminology error. Please improve current translation.

LLM's proposal:
the outbreak of the new crisis

Reject

resample from LLM

Accept

Repeat above steps for n iterations

LLM's final output:
the outbreak of the Covid-19 crisis

Source Translation: 新冠疫情危机爆发

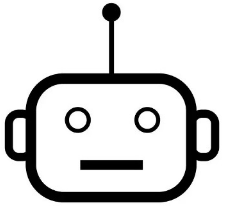
Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

LLMRefine Algorithm

Repeat n times

Obtain feedback F_i from error pinpoint

Sample revision c_i based on feedback f_i and last generation y_{i-1}



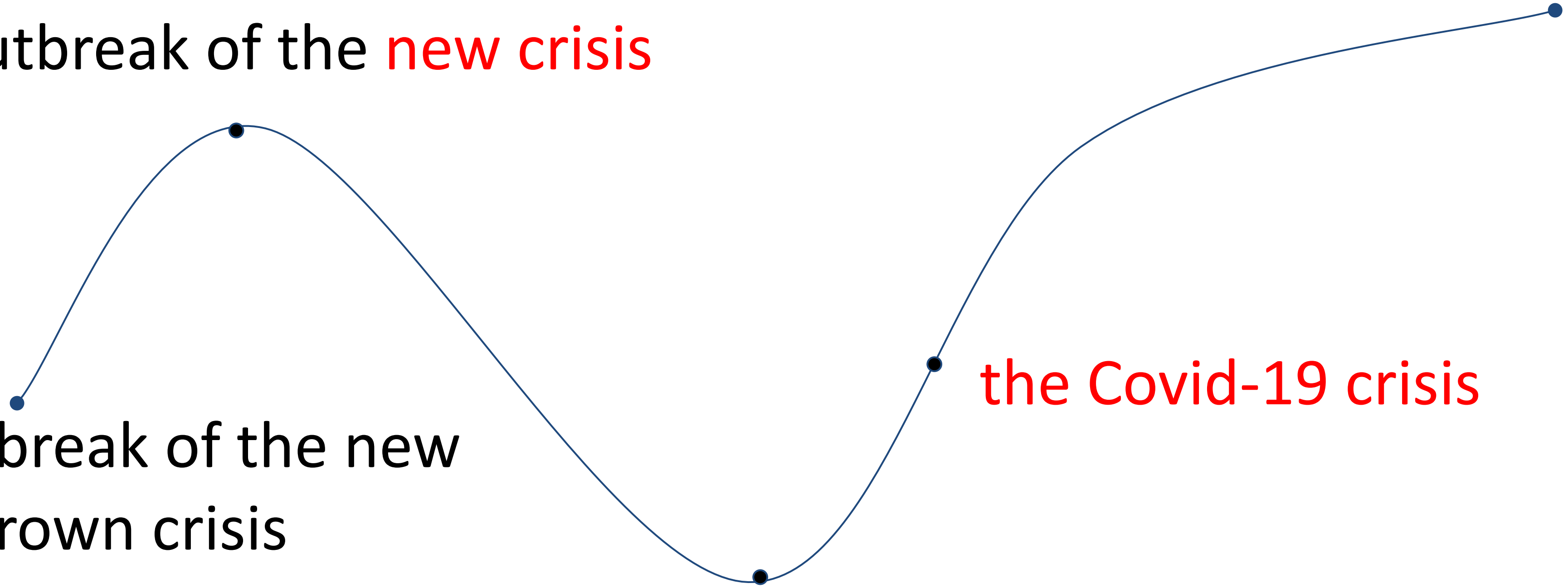
$$P_{accept} = \min(1, e^{\frac{s(F(c_i)) - s(F(y\_i))}{n * T_i}})$$

Accept new revision

Keep the last step candidate

$$T_{i+1} = max(T_i - c * T_i, 0)$$

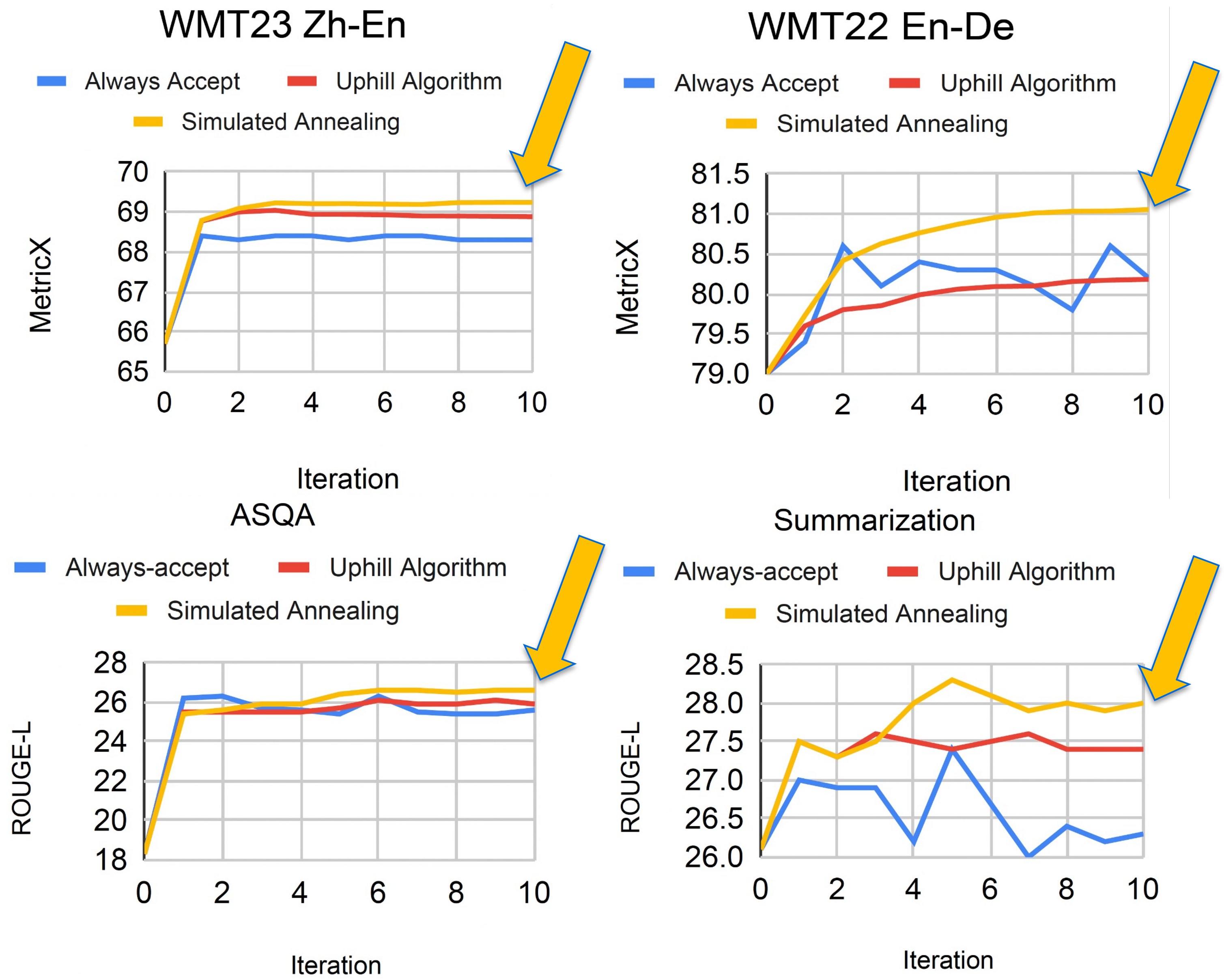
Source Translation: 新冠疫情危机爆发



"the new crisis" is a major mistranslation error. The correct translation should be: " the Covid-19 crisis"

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

Simulated Annealing can boost refinement



Translation
Summarization
Long form QA

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL24

Key insights of LLMRefine

- Binary feedback is not enough
- Fine-grained feedback is better
- Algorithmic iterative refinement is superb

Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL24

Summary

- Can we trust LLM evaluator?
  - Self-bias in LLM Evaluators (source-based)
- Evaluating LLM Generation Quality
  - Interpretable text generation evaluation (InstructScore)
  - Assessing knowledge in LLMs (KaRR)
- Post-training alignment
  - Online Preference Optimization (BPO)
  - Iterative refinement with fine-grained feedback (LLMRefine)

Future thoughts

- Evaluating
  - complex knowledge
  - LLM RAG
  - LLM Agent
- Evaluation for open-end generation
  - PerSE at EMNLP 2024
- Better/robust alignment learning

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.
- Wenda Xu, Jiachen Li, William Yang Wang, Lei Li. BPO: Staying Close to the Behavior LLM Creates Better Online LLM Alignment. EMNLP 2024.
- Xu, Deutsch, Finkelstein, Juraska, Zhang, Liu, Wang, Li, Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024.