Assessing and Improving Large Language Models

Lei Li
Language Technologies Institute
Carnegie Mellon University

Large Language Model Products

Google
Bard
Gemini

OpenAI
ChatGPT
GPT-4

Meta
Llama 2

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

Evaluating Large Language Models

- BLEU for evaluation?
  - 20 year old metric… with obvious limitation.
- But LLM generation requires new metrics
  - diverse output (OOD)
  - BLEU/ROUGE will have significantly decreased correlations with human judgments.

Outline

- InstructScore: Explainable Text Generation Evaluation
- Assessing Knowledge in LLMs (KaRR)
- Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback

When you made a mistake…

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

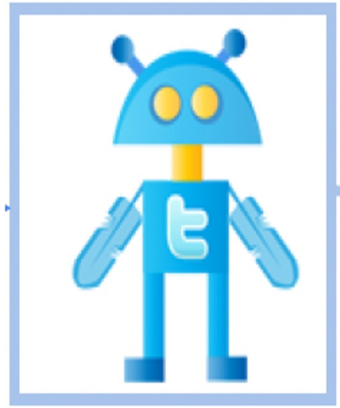
outbreak

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

Limitations of Prior Metrics
- Lack of Interpretation

Reference: The outbreak of the COVID-19 crisis

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".

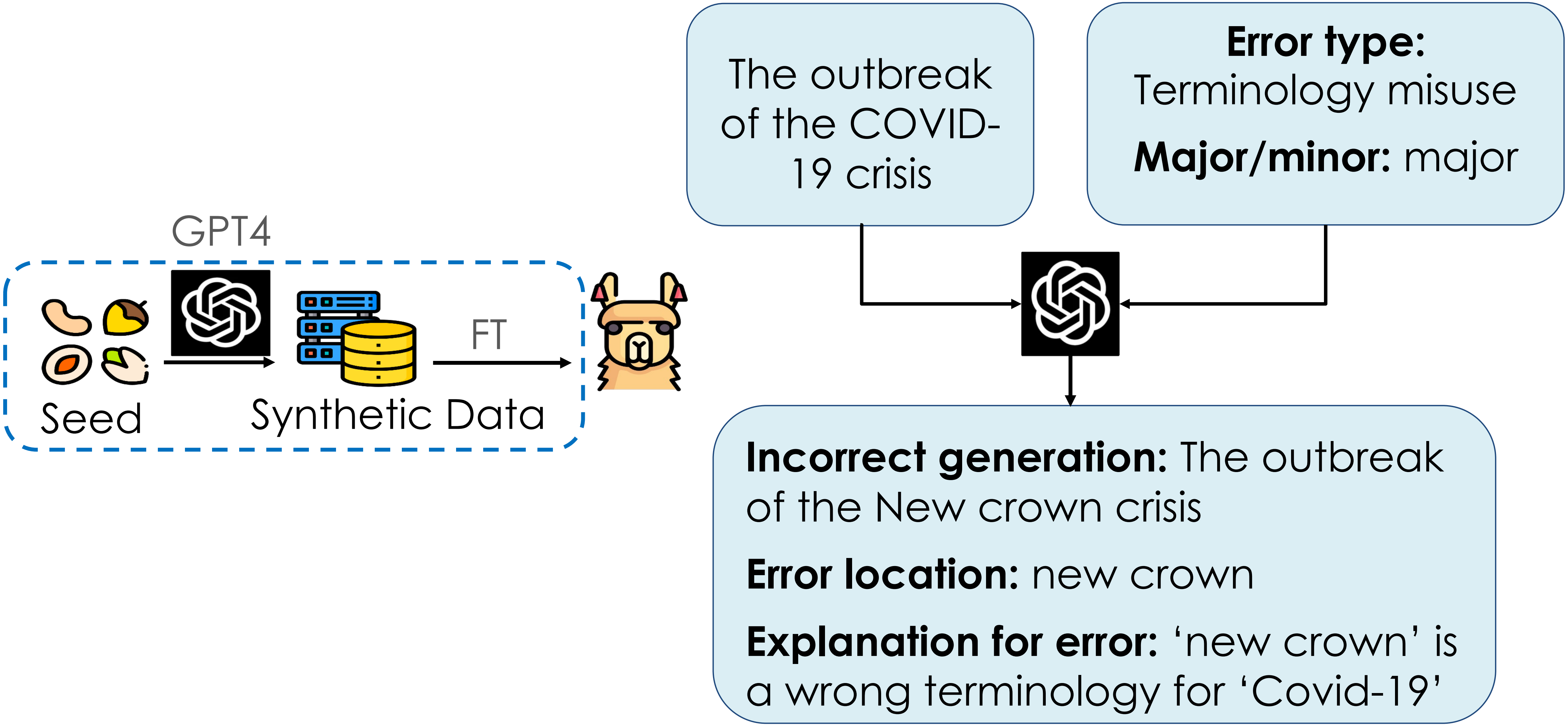
Why is training an explainable metric challenging?

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

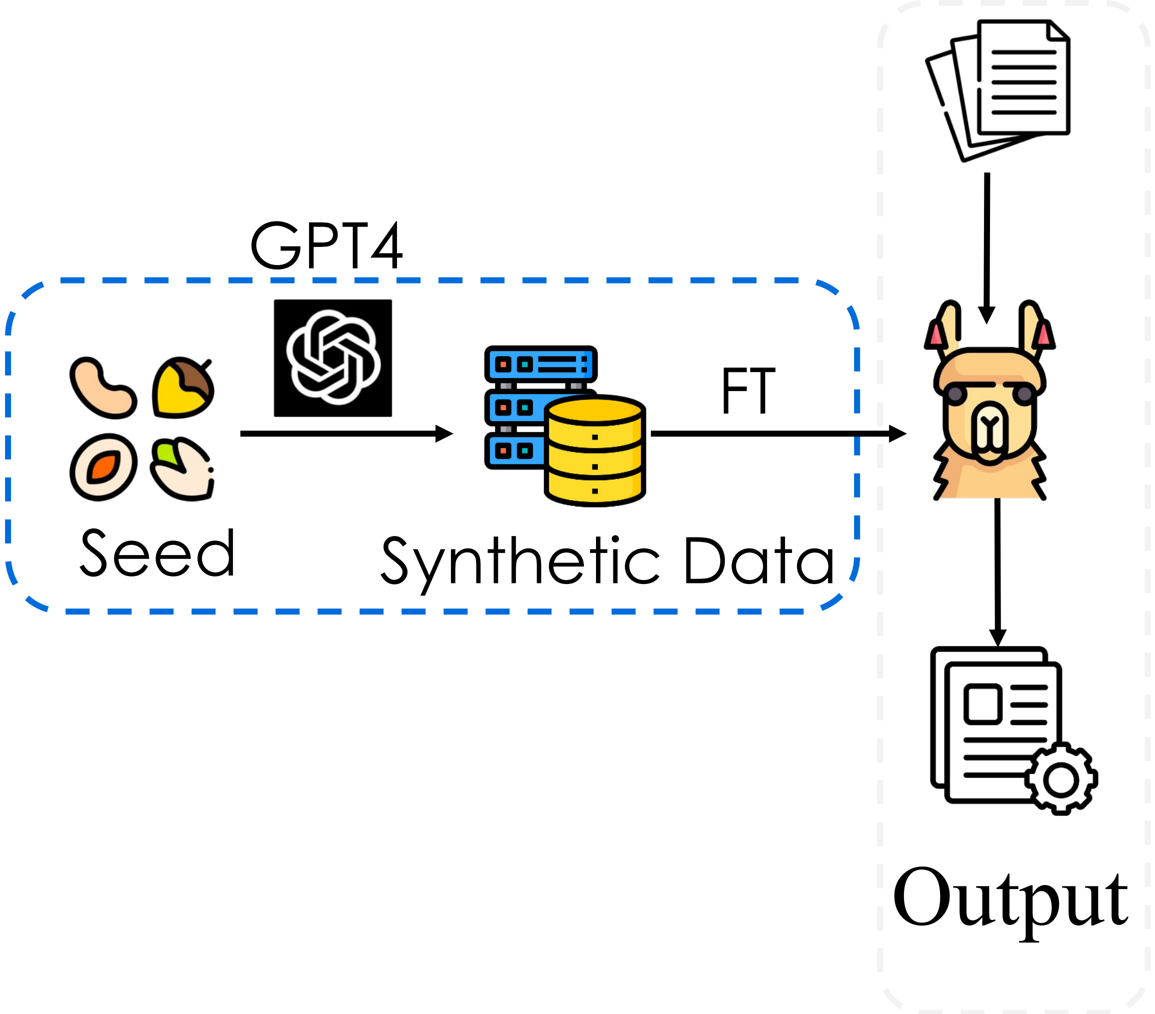
Ideal Metric

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

Direct Prompting



But, failed explanation in GPT4

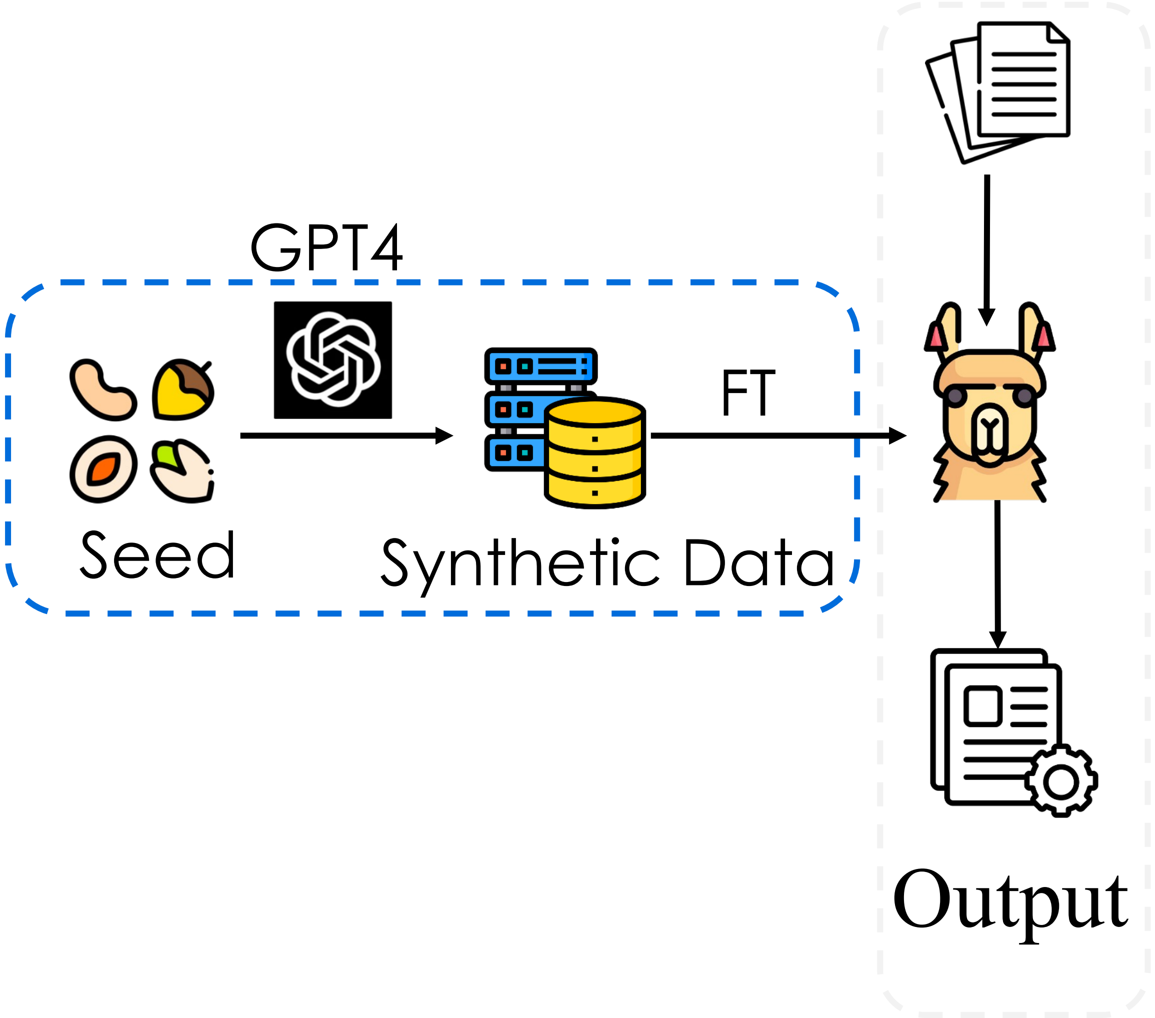


Error type 3: Missing information

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

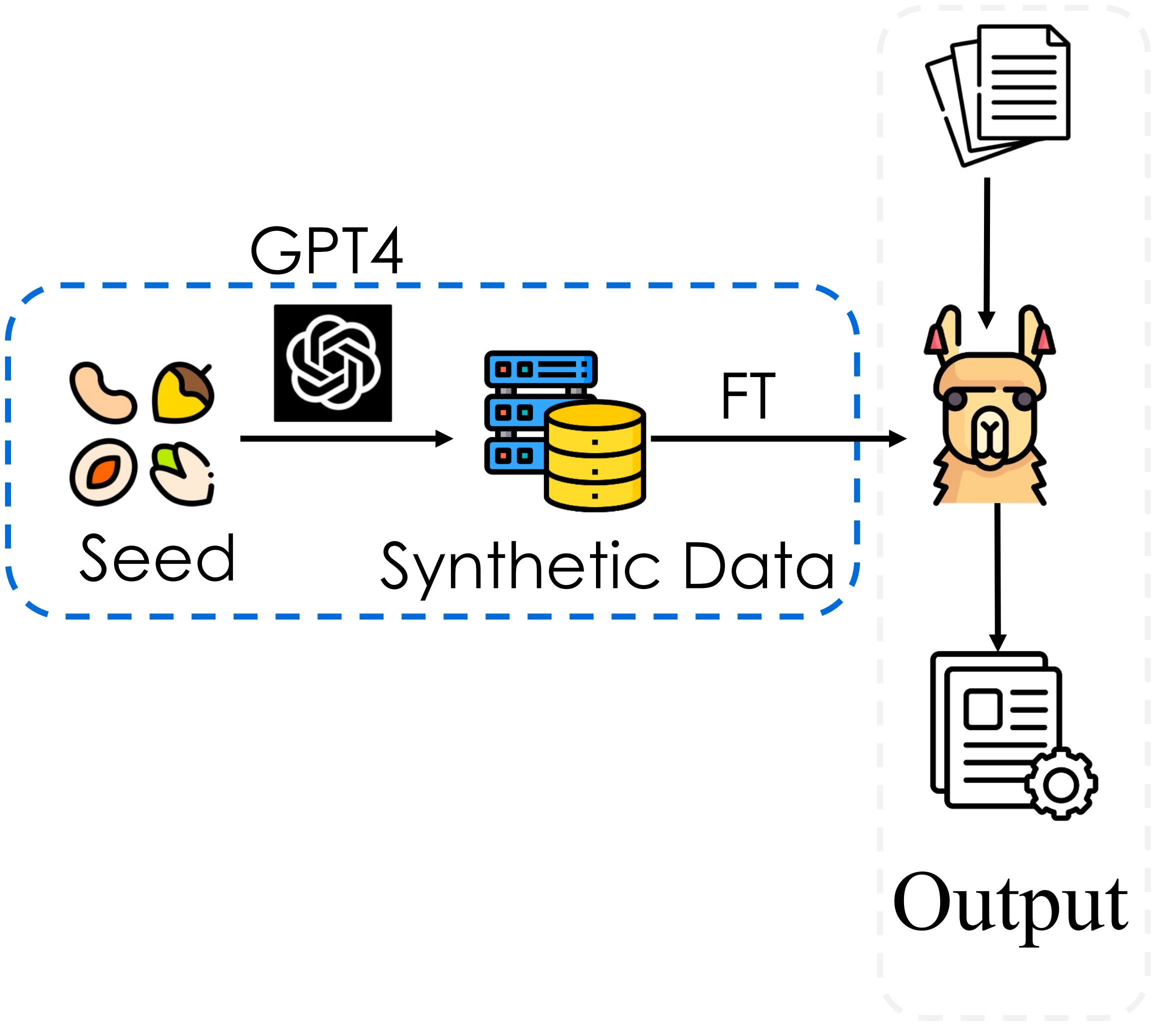


Evaluated text: The outbreak of the new crown crisis

Error location: 'virus'

Hallucination

But, failed explanation in GPT4



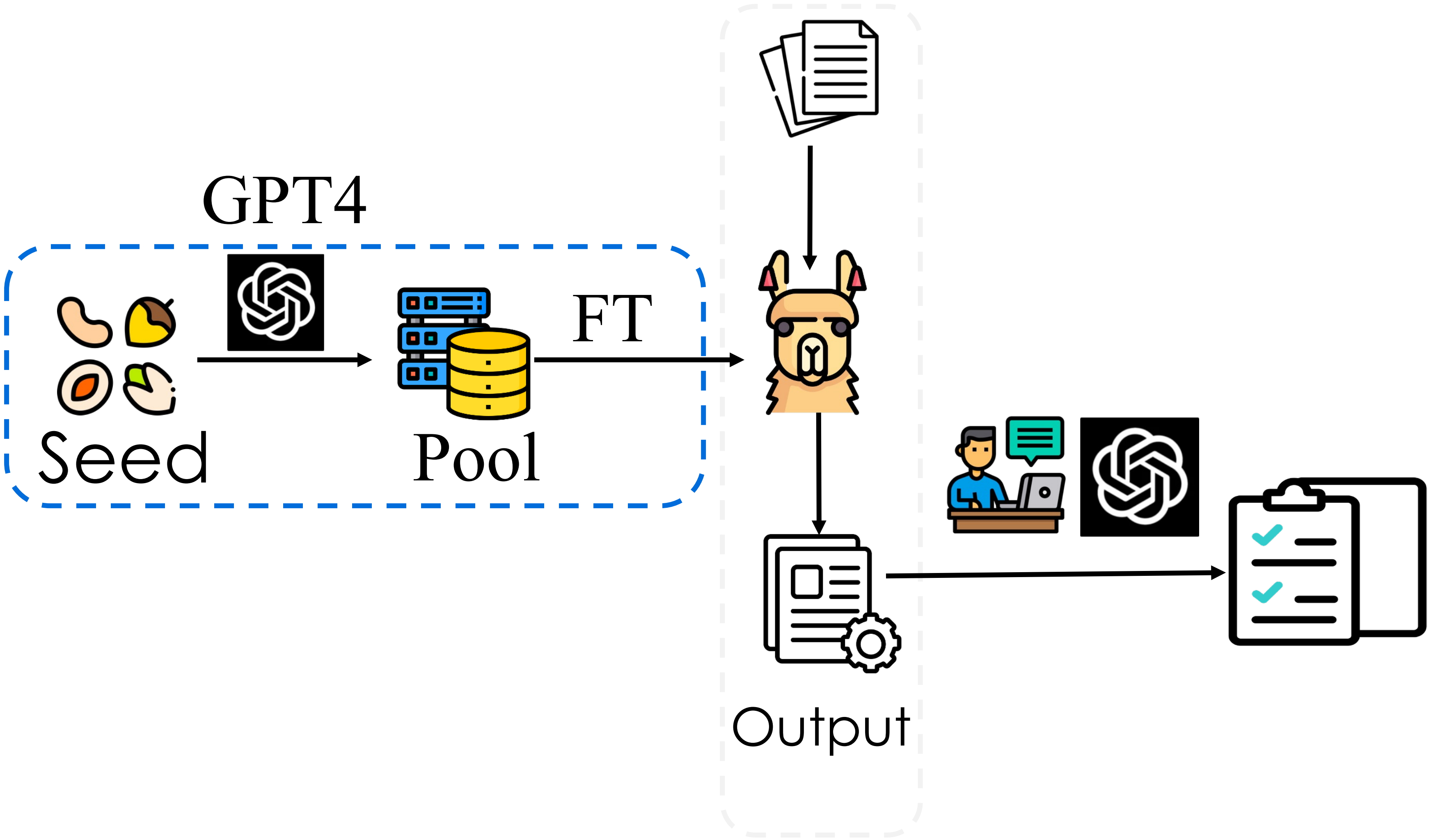
Explanation for error 1: The incorrect translation uses the word "annual" instead of "annual"

Explanation is illogical

Failures of GPT4 generated explanation

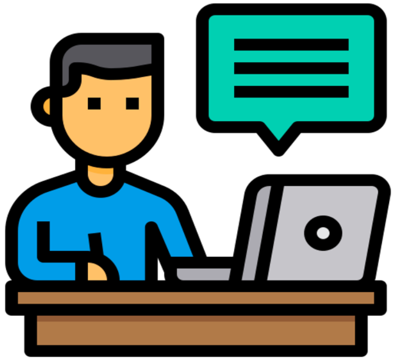
Fields	Failure Mode	Description (M is local failure mode, G is global failure mode)
Error Type	Inconsistency to explanation	M1: Error type is inconsistent with explanation
Error Location	Inconsistency to explanation	M2: Error locations are not consistent with the explanation
	Hallucination	M3: Error 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 reward 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 reward 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?

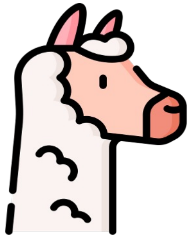
Are two phrase aligned?

InstructScore: Automatic Feedback

Reference
Candidate

Error location1
Error Type1
Major/Minor
Explanation1

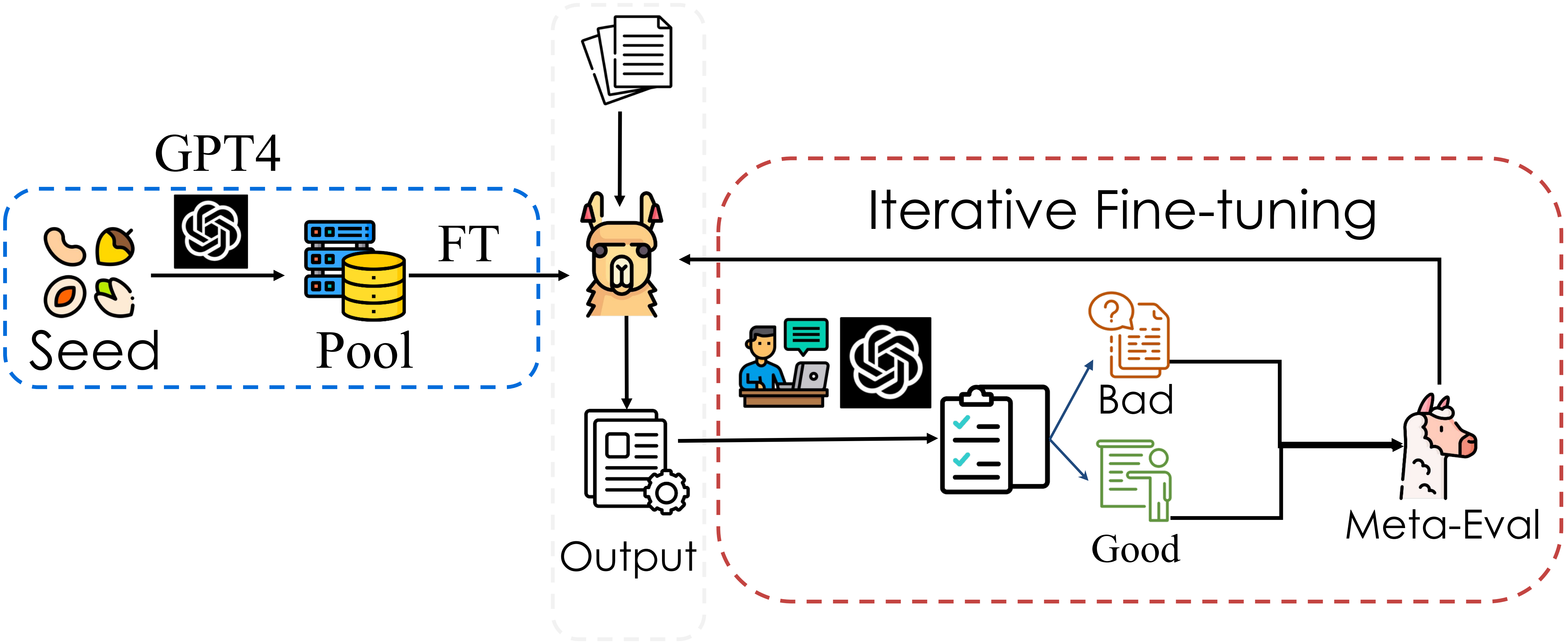
Error location2
Error Type2
Major/Minor
Explanation2



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

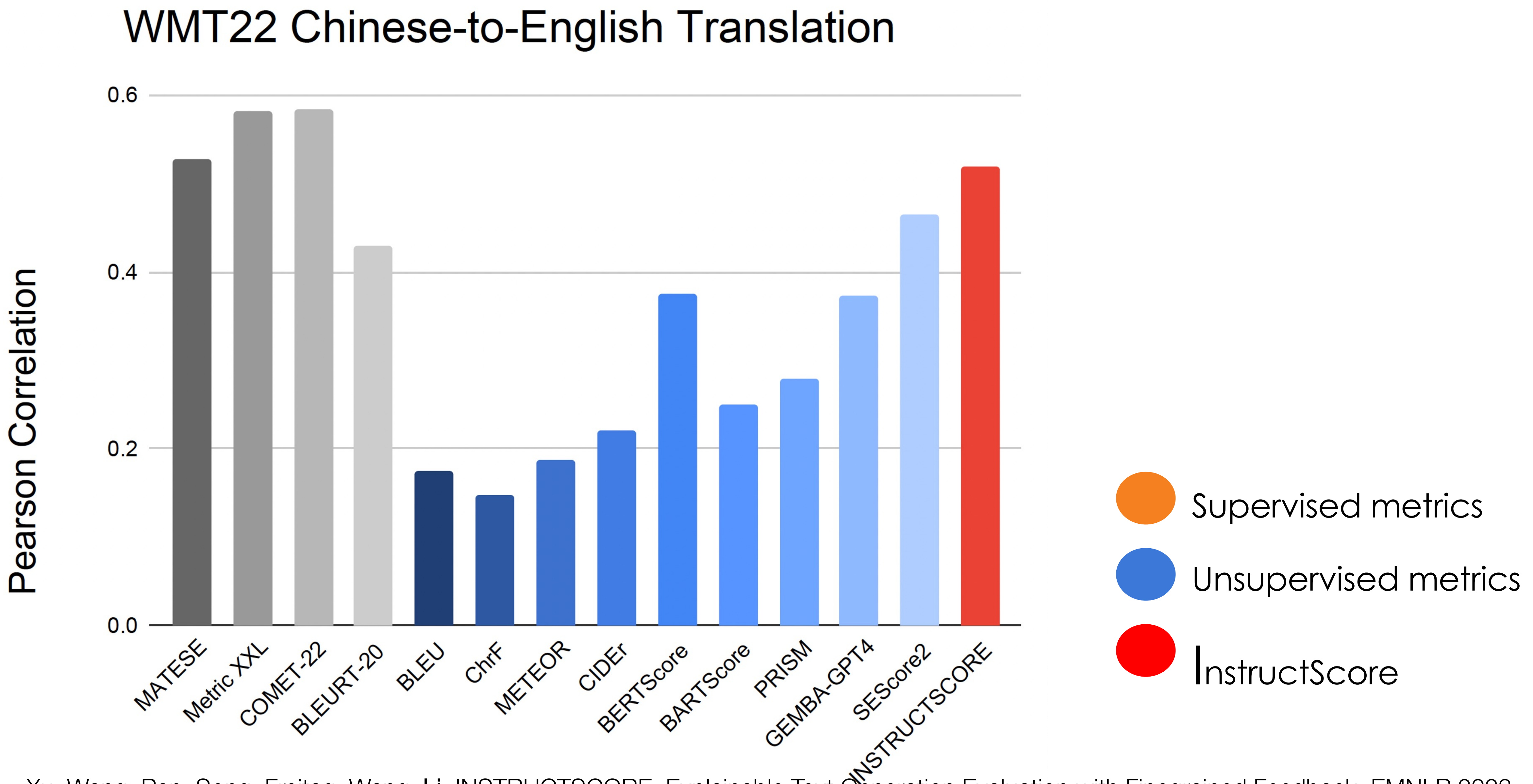
Alignment Score: 7/8

InstructScore: Refinement



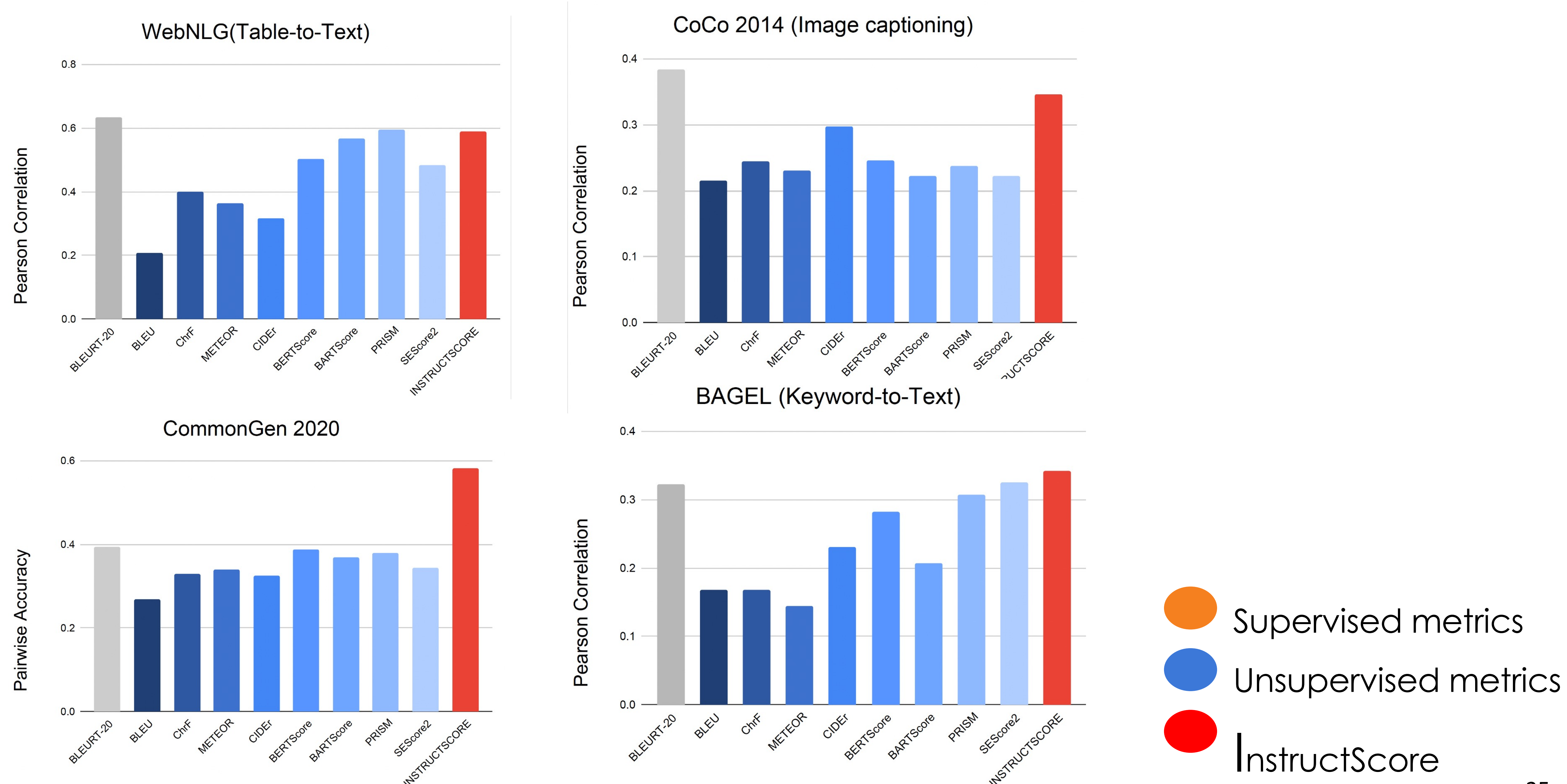
Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

InstructScore can judge machine translation!



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

InstructScore can do well in other tasks as



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?"

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

2X

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

- InstructScore: Explainable Text Generation Evaluation
- Assessing Knowledge in LLMs (KaRR)
- Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback

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

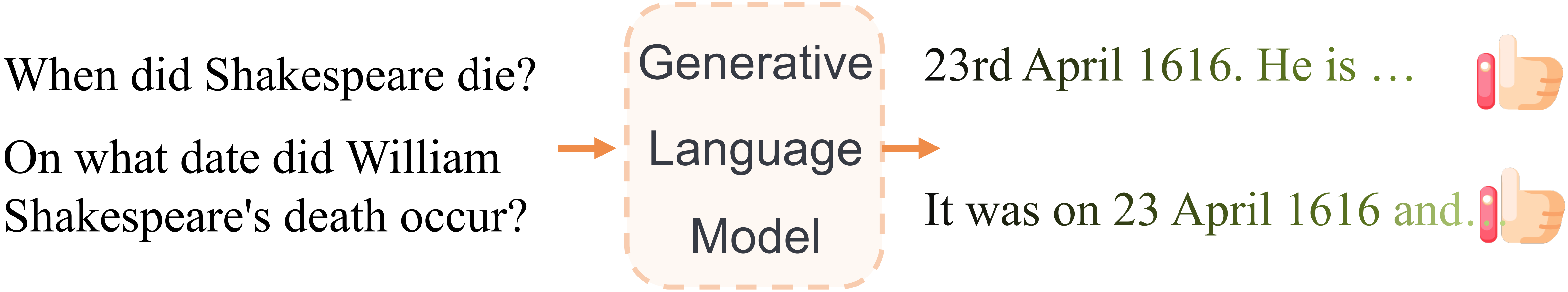
1. Distinguish if text generation stems from genuine knowledge or just high co-occurrence with given text.

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

    $\boxed{\text{(a random name)}}$'s job is a writer.

Assessing LLM's Knowledge

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



Reliable?

Why Do We Need Knowledge Assessment?

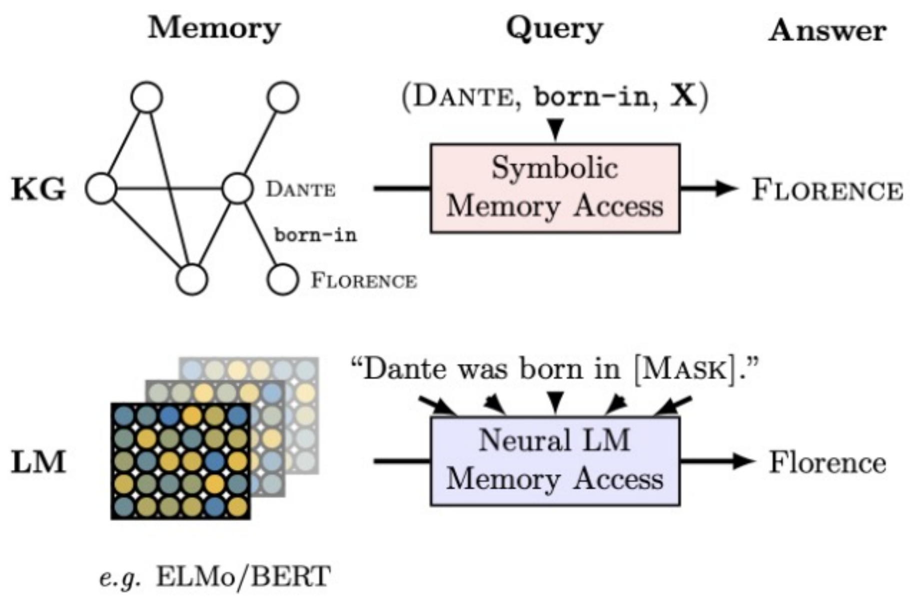
- The assessment results directly affect the people's trust in the LLM generated content.

- 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.*

Challenges in Knowledge Assessment

- **Accuracy v.s. Reliability:** Previous studies primarily assess accuracy, not reliability.

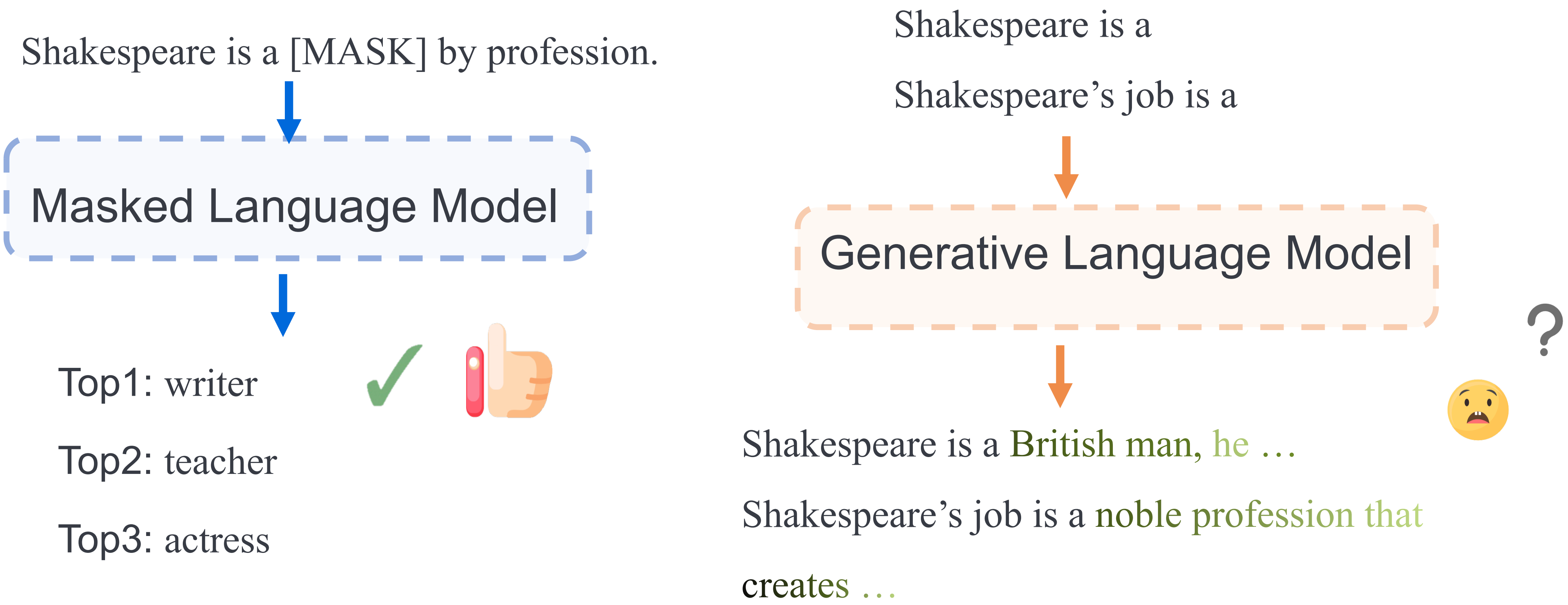


Probing method for MLM[1]

[1]*Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In Proceedings of EMNLP-IJCNLP, 2019.*

Challenges in 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.

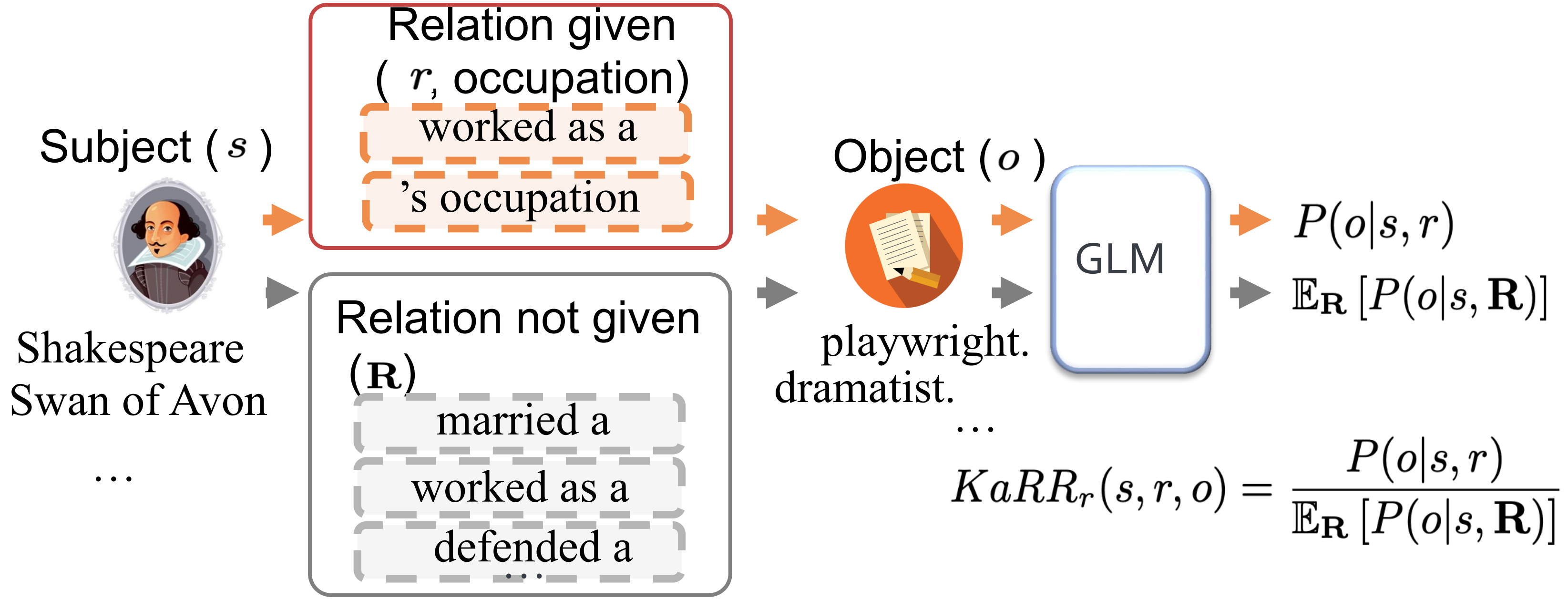
- Example: a disease noted by D, and no disease noted by $\neg D$, exposure noted by E, and no exposure noted by $\neg E$. The risk ratio can be written as:

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

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

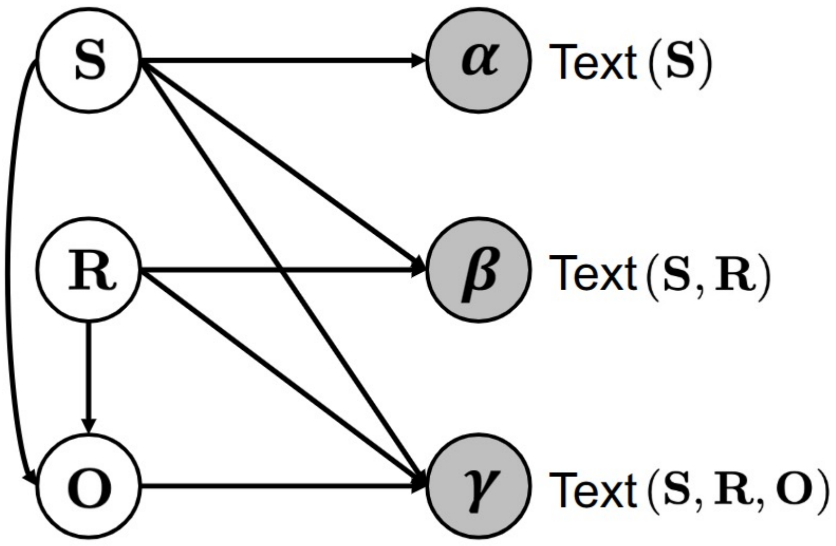
Knowledge Assessment Risk Ratio

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



Graphical Model for Knowledge Assessment

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



hollow circles: latent variables
shaded circles: observed variables

Establish the connection between symbols and text forms.

Goal: estimate the model knowledge on **symbols** through the observable model probability across diverse corresponding **textual forms**.

Calculating KaRR

KaRR is formulated based on knowledge symbols. The graphical model facilitates the implementation by employing model probabilities on the text.

E.g., we can use the graphical model to help calculate the numerator of KaRR_s and KaRR_r:

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

Further, we use $P_{\mathcal{M}}$ to denote the generation probability of model $\mathcal{M}$ then,

$$P(o \mid s, r, \beta_k) = \sum_{j=1}^{|\gamma|} P\left(o, \gamma_j \mid s, r, \beta_k\right) = \sum_{j=1}^{|\gamma|} P_{\mathcal{M}}\left(\gamma_j \mid s, r, \beta_k\right) P\left(o \mid \gamma_j\right)$$

KaRR Dataset

- Good coverage -- 994,123 entities and 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].", …

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.

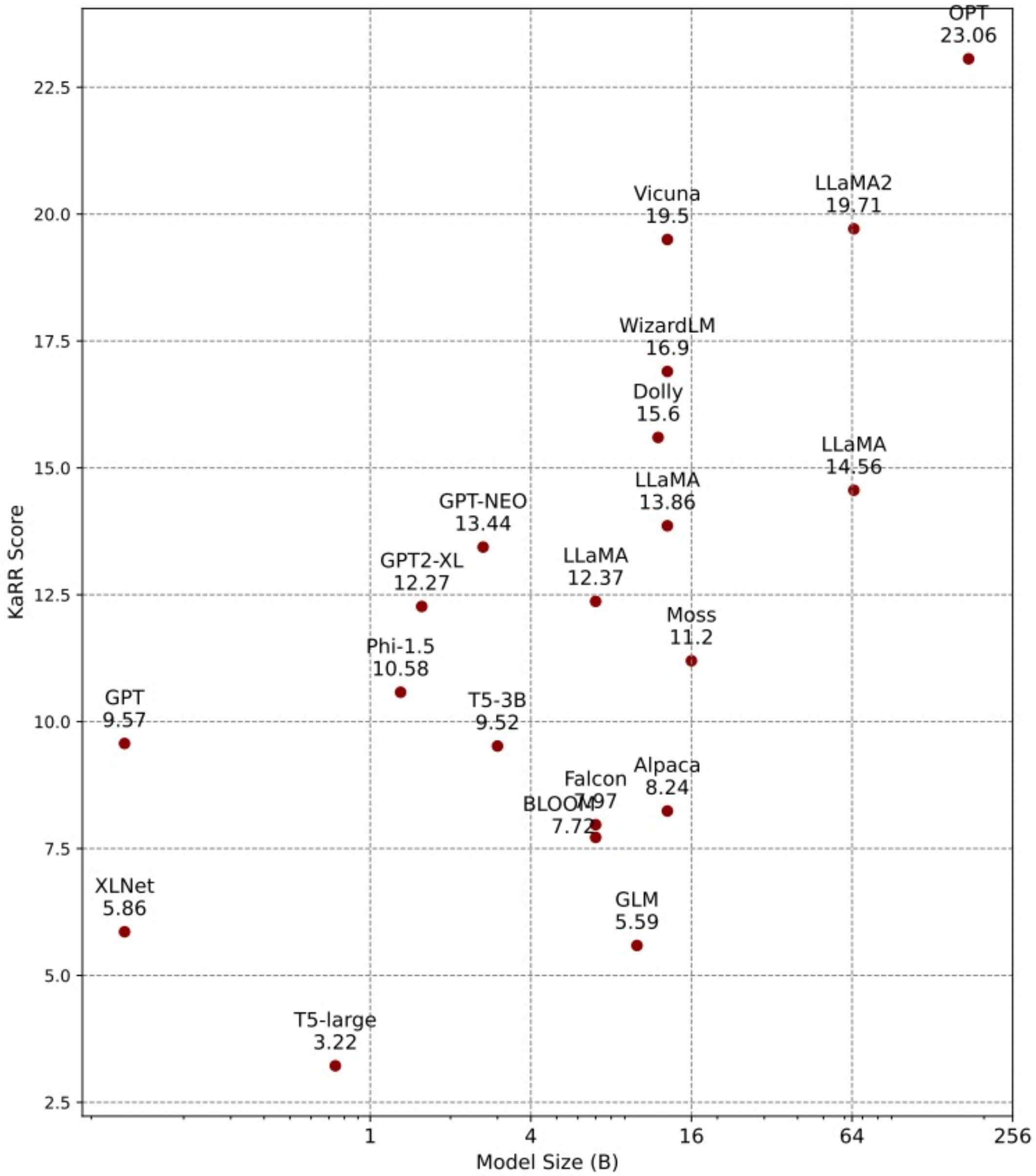
  2) Rating: another 3 annotators to rate the knowledge (0 or 1) in model according to the generations.

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.

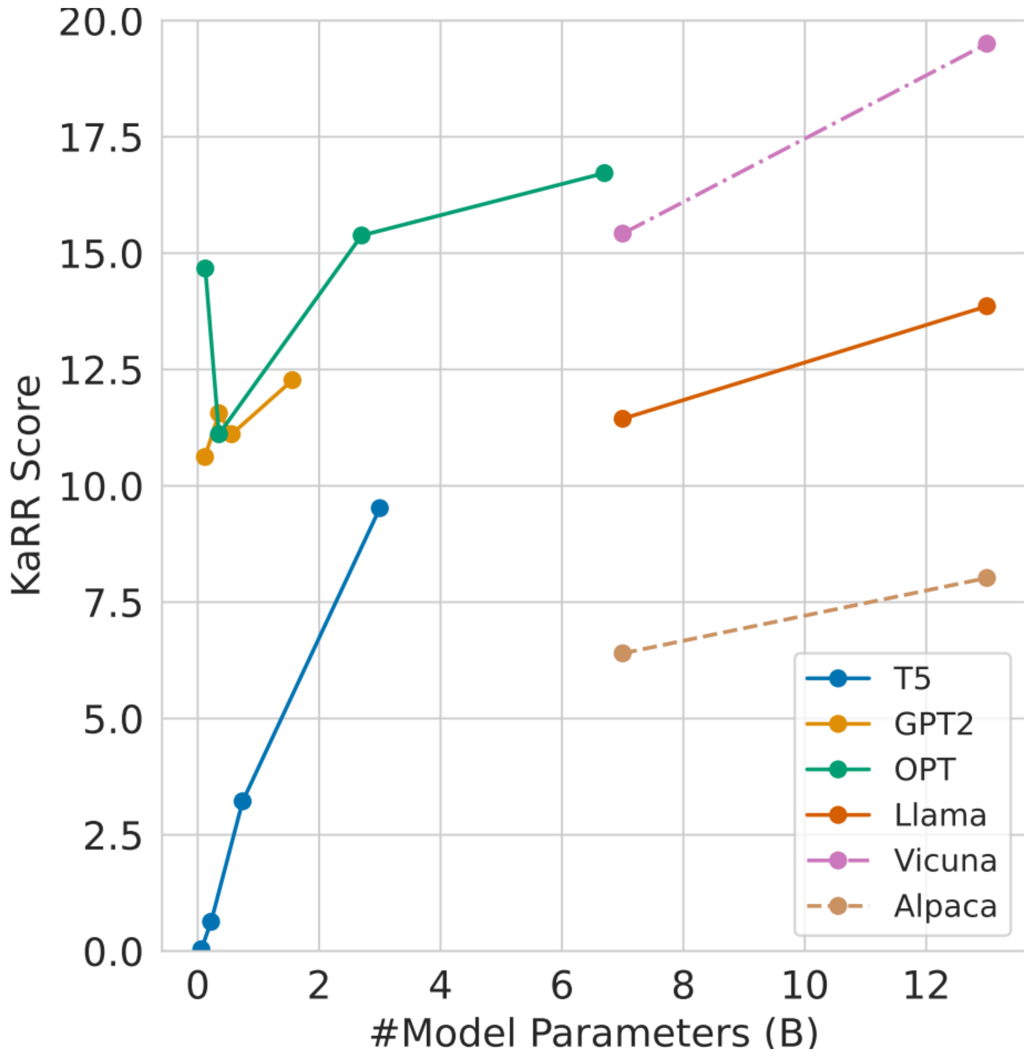
KaRR Scores on 20 LLMs

- 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

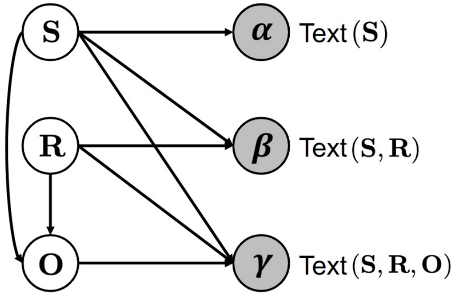
- larger models generally hold more factual knowledge.
- Scaling benefits vary among models. E.g., T5-small to T5-3B.

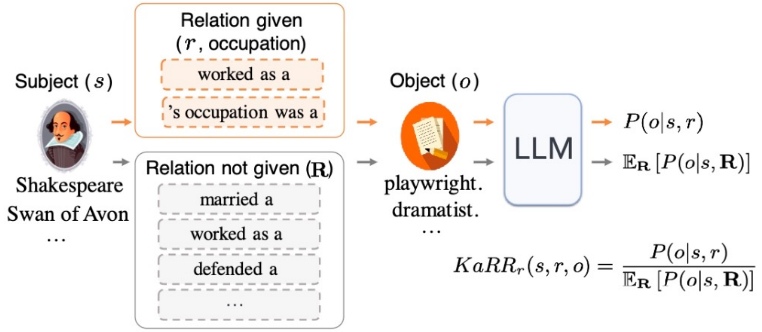


Summary of LLM Knowledge Assessment

- 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

- InstructScore: Explainable Text Generation Evaluation
- Assessing Knowledge in LLMs (KaRR)
- Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback

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

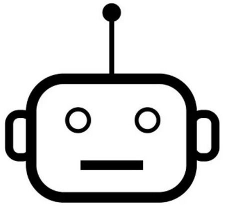
Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

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}



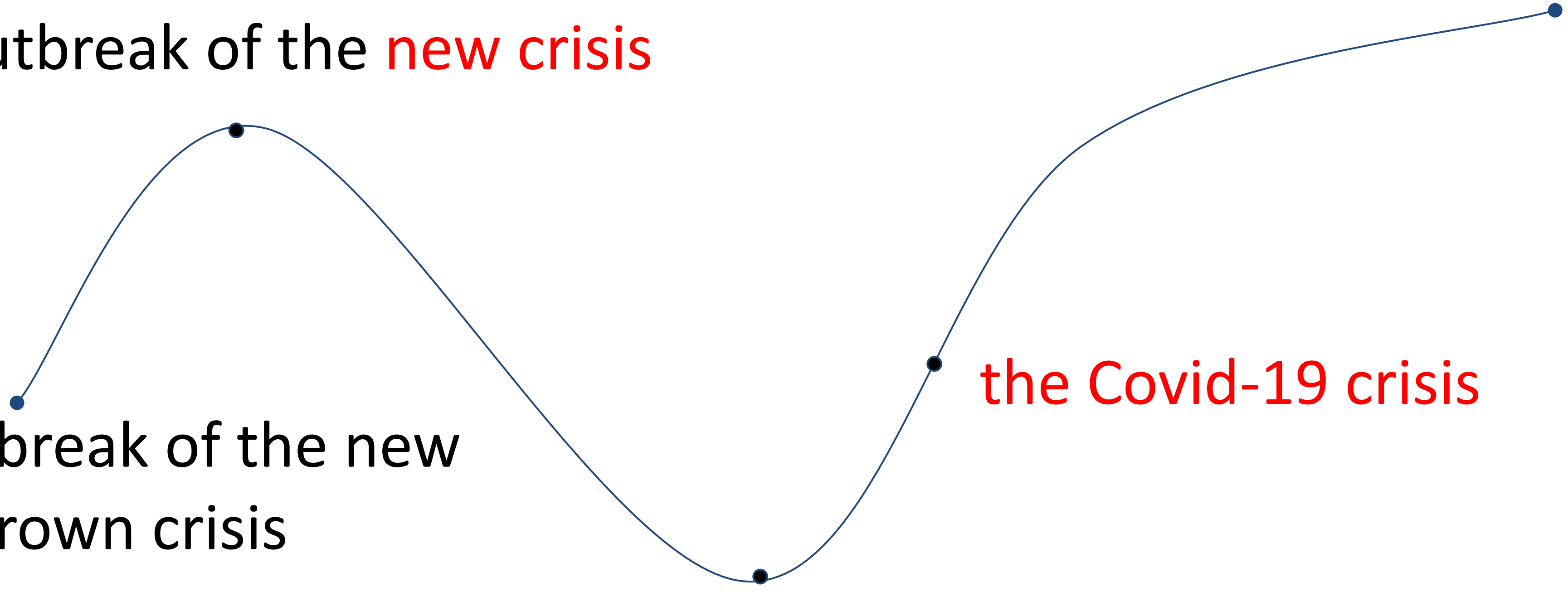
$$P_{accept} = \min(1, e^{\frac{s(F(c_i)) - s(F(y\_i))}{n * T_i}})$$

Accept new revision

Keep the last step candidate

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

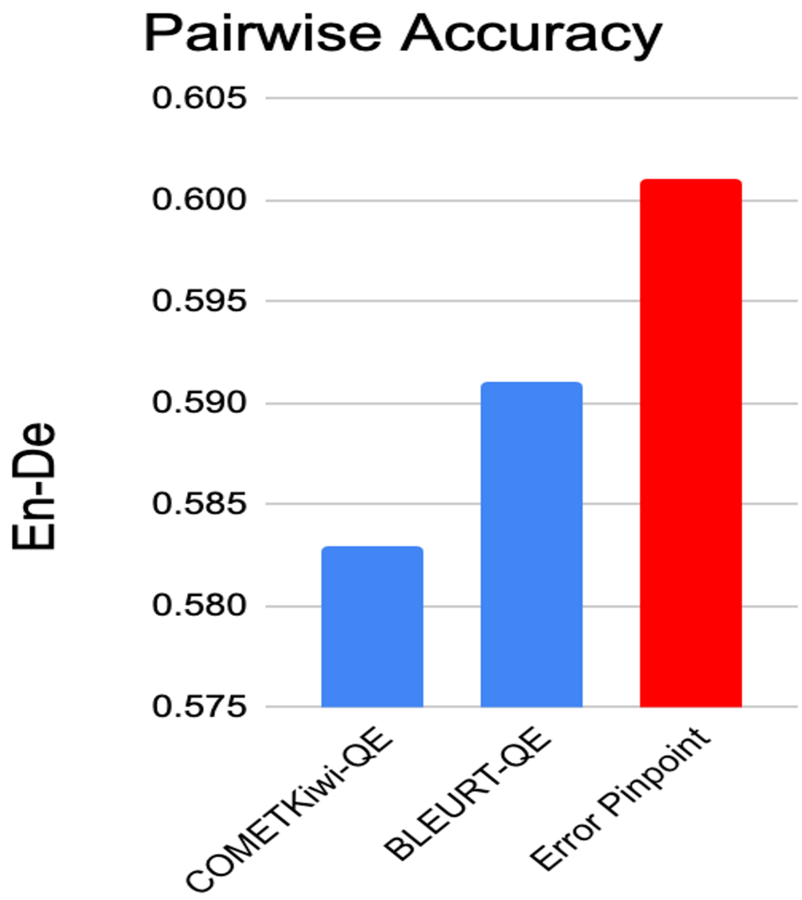
Source Translation: 新冠疫情危机爆发

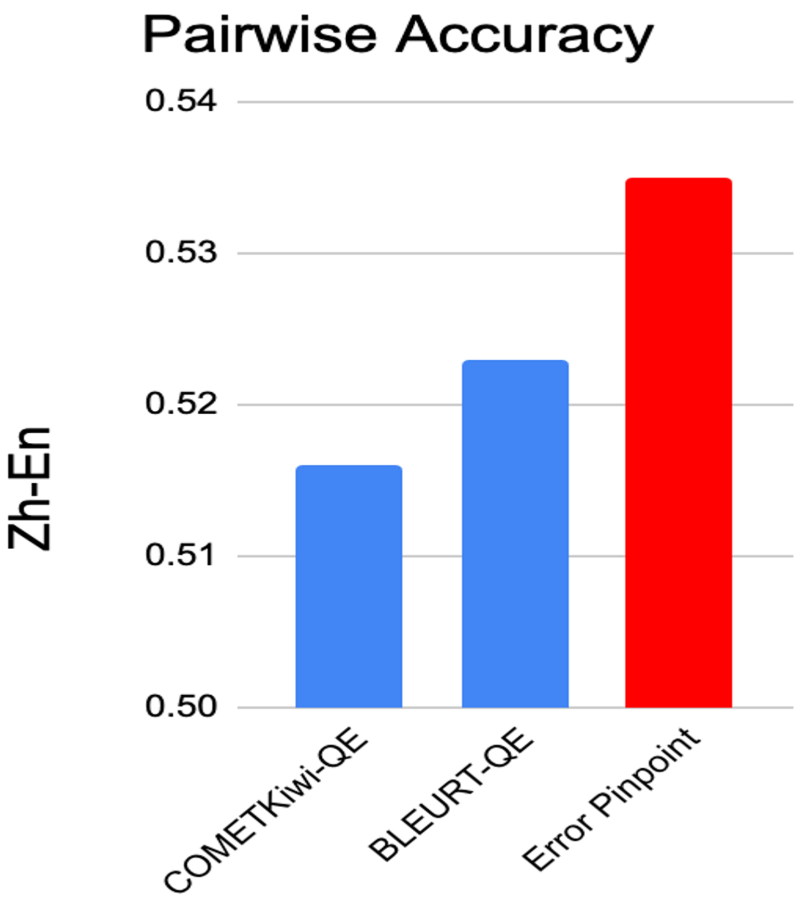


"the new crisis" is a major mistranslation error. The correct translation should be: " the Covid-19 crisis"

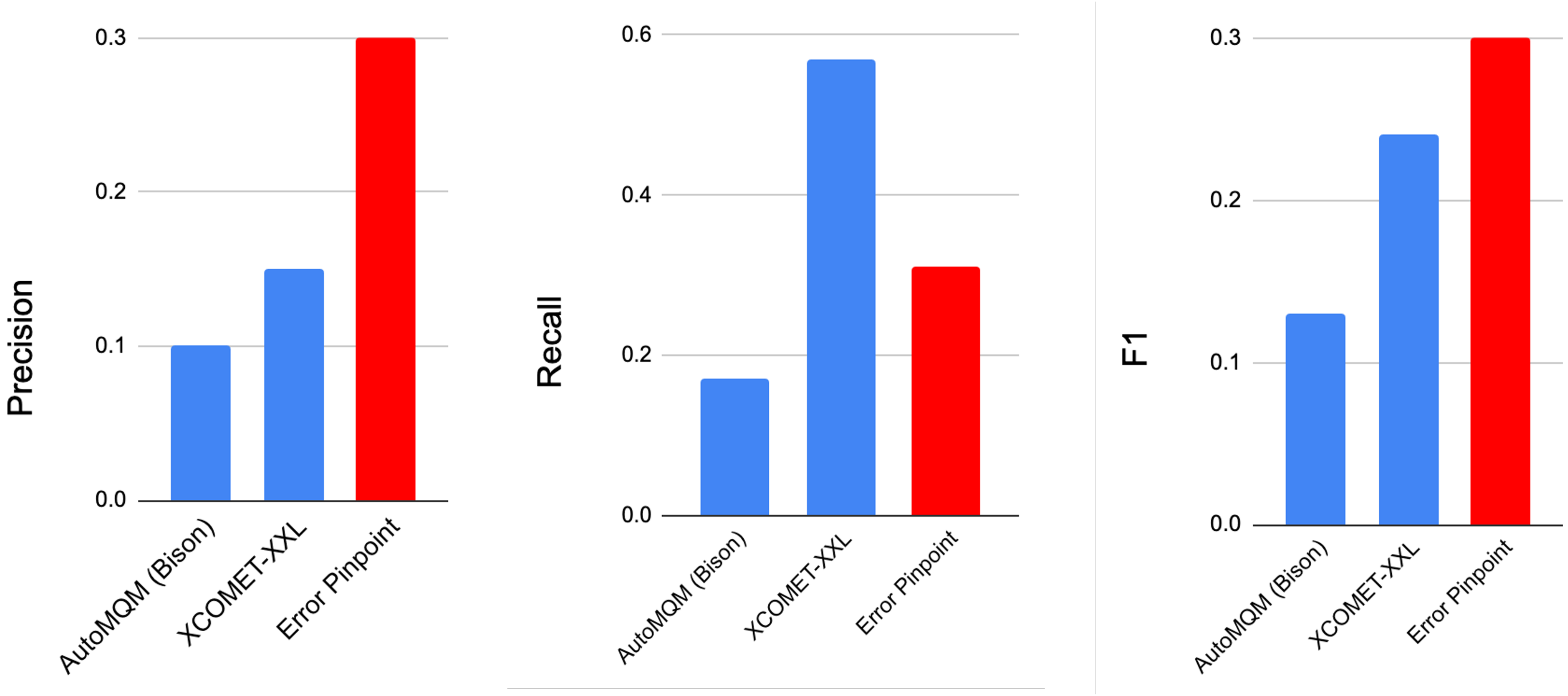
Wenda Xu, Daniel Deutsch, Mara Finkelstein, JurajJuraska, Biao Zhang, Zhongtao Liu, William Yang Wang, Lei Li, and Markus Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024

RQ1: How well does our error pinpoint model align with human annotations of generation quality?

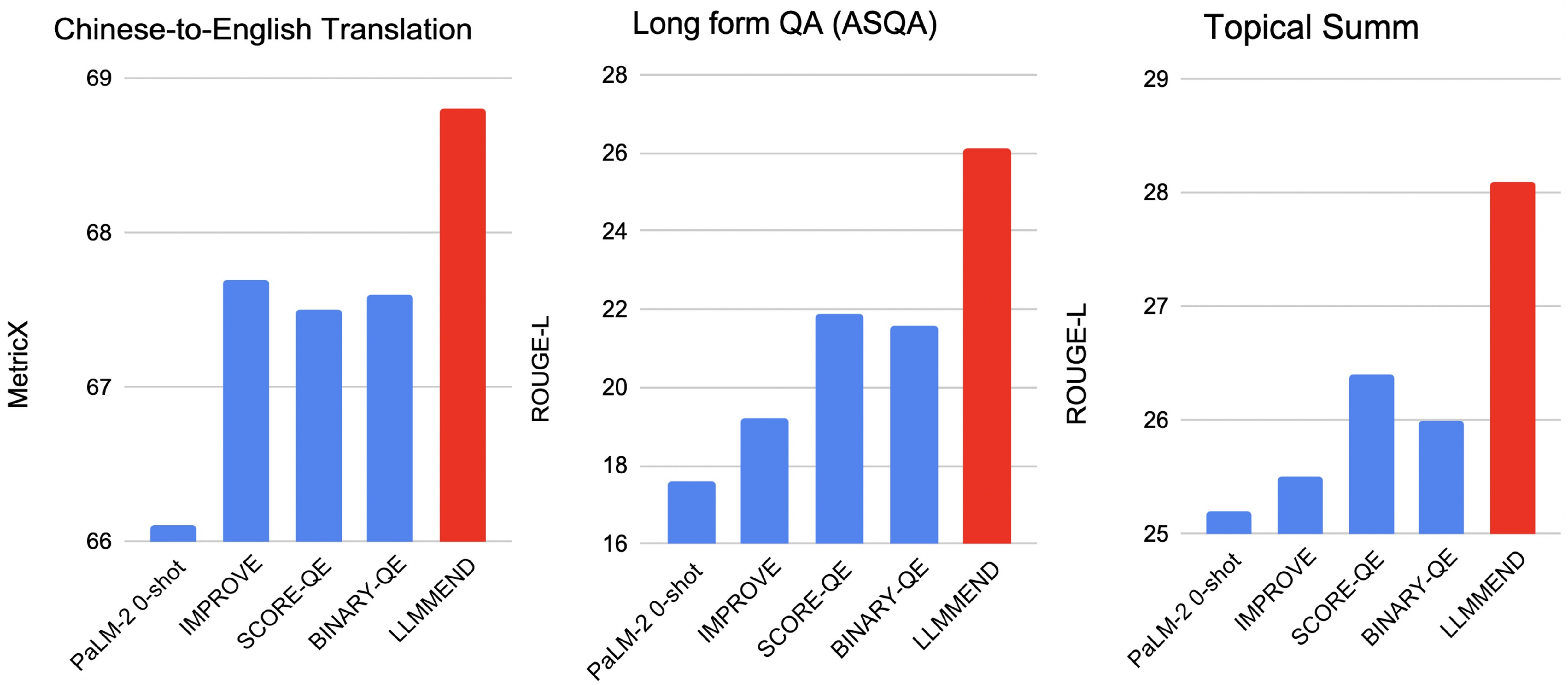




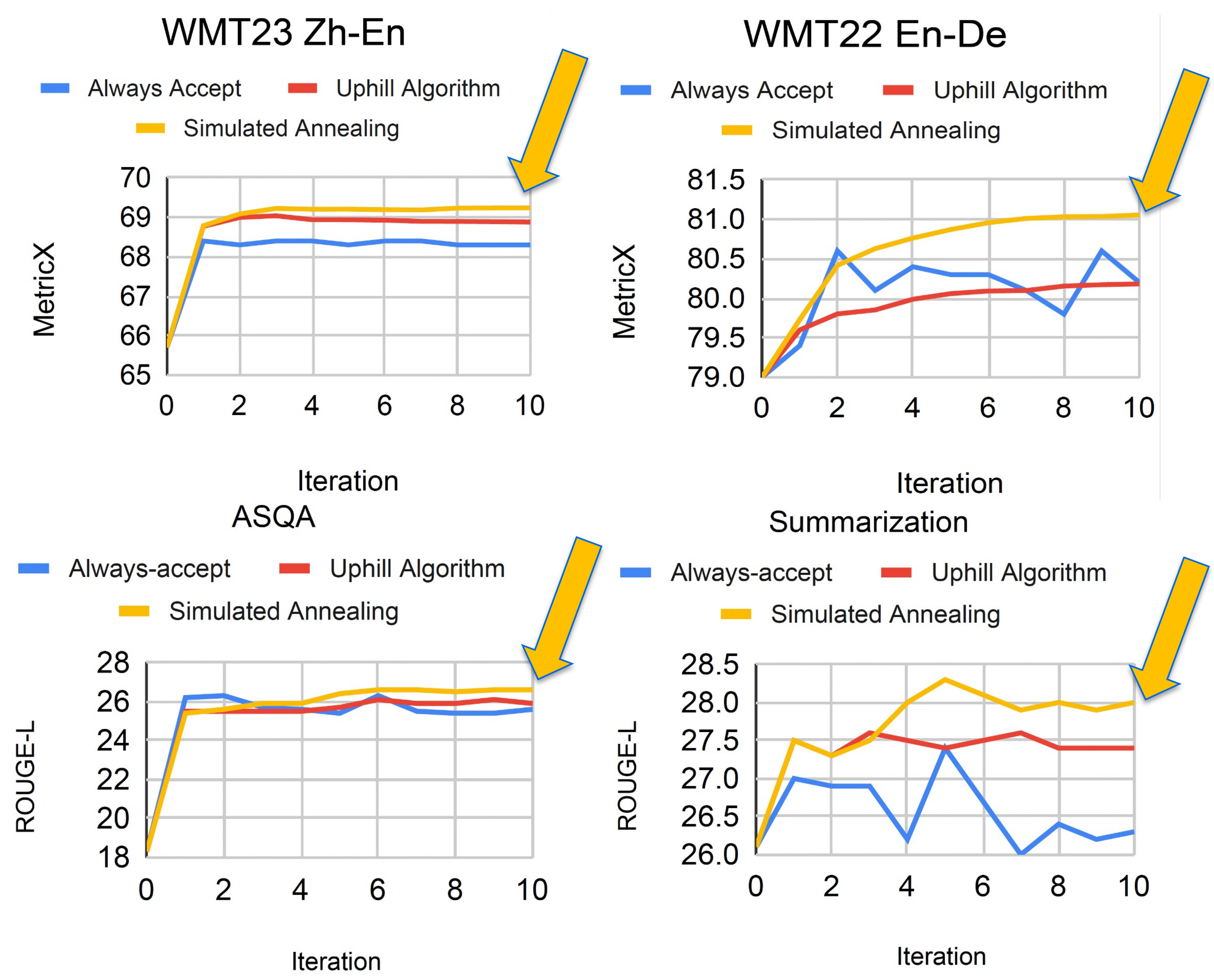
RQ1: How well does our error pinpoint model align with human annotations of translation quality?



RQ2: Does fine-grained feedback result in better downstream translations than more coarse feedback?

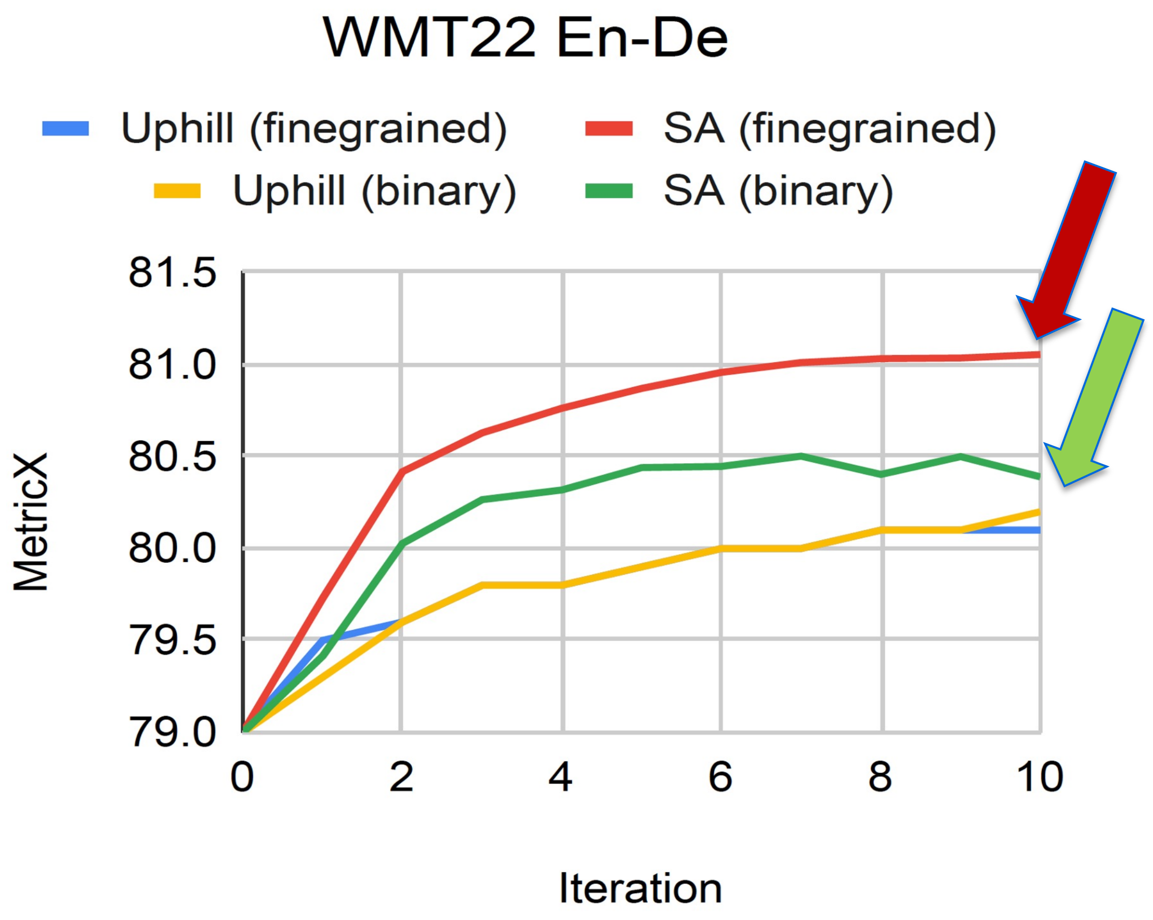


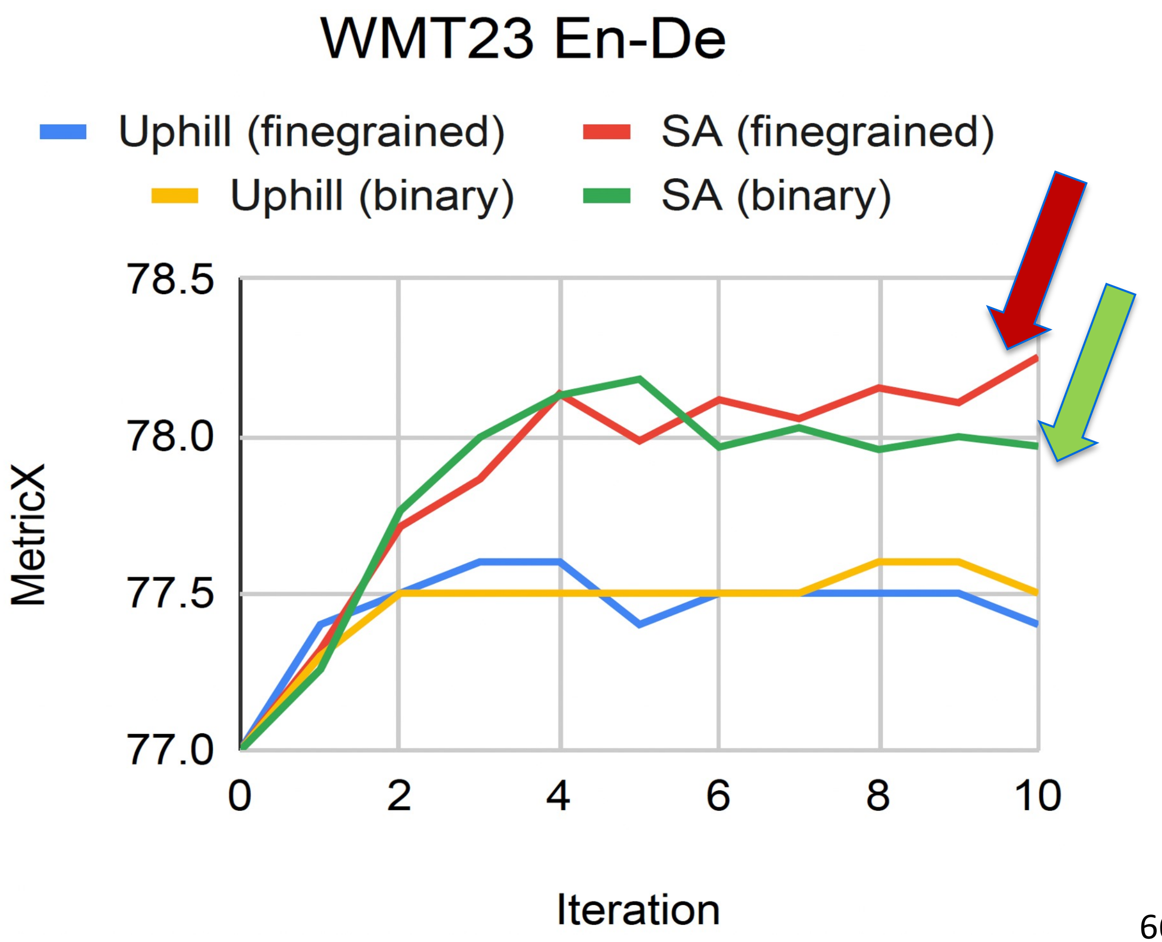
Simulated Annealing can boost refinement



Translation
Summarization
Long form QA

Simulated annealing can boost performance of both coarse and fine-grained feedback





Human Evaluation further validates our results

Our fine-grained has all win/lose ratios greater than 1

WMT22 En-De	Win/lose ratio
0-shot	2.34
Improve	2.44
BLEURT-Score-QE	2.79
BLEURT-Binary-QE	1.76
Score-QE	1.23
Binary-QE	1.84

Our SA has all win/lose ratios greater than 1

WMT22 En-De	Win/lose ratio
Always-Accept	1.56
Greedy Uphill	1.38

Summary

- InstructScore: Explainable Text Generation Evaluation
- Assessing Knowledge in LLMs (KaRR)
- Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback

Reference

- Xu, Wang, Pan, Song, Freitag, Wang, Li. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023. https://arxiv.org/abs/2305.14282

- Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023. https://arxiv.org/abs/2305.10519

- Xu, Deutsch, Finkelstein, Juraska, Zhang, Liu, Wang, Li, Freitag. LLMRefine: Pinpointing and Refining Large Language Models via Fine-Grained Actionable Feedback. NAACL 2024. https://arxiv.org/abs/2311.09336