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

ANTHROP\C

LE **You**

create a funny viral tweet about my talk at Michigan State University with the title "Assessing and Improving Large Language Models"

ChatGPT

🤖✨ Excited to announce my upcoming talk at Michigan State University titled 'Assessing and Improving Large Language Models'! Join me as I reveal the secret lives of AI models - spoiler alert: they're just like us, but with an endless appetite for data and a slight obsession with cats. 🐱💻 #AIHumor #MichiganStateUniversity

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: $\mathrm{P}(\mathrm{x}_{1..\mathrm{T}}) = \prod_{t=1}^{\mathrm{T}} \underline{\mathrm{P}(\mathrm{x}_{t+1}|\mathrm{x}_{1..t})}$

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…

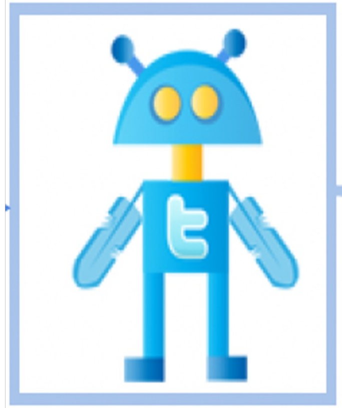
Teacher 1:
You have a bad translation. You get score of 20/100

Teacher 2:
Wifecake != 'Sweetheart cake'. This is a major mistranslation error. Score: 20/100

Limitations of Prior Metrics
- Lack of Interpretation

Reference: Is there a wife in the wifecake?

Candidate: Is there a wife in the sweetheart cake?



BLEU: 0.661

BertScore: 0.925

COMET: 0.711

BLEURT: 0.519

SEScore2: -5.43

Ideal Metric: Fine-grained Explanation

Reference: Is there a wife in the sweetheart cake?

Candidate: Is there a wife in the wifecake?

Error location: wifecake

Error type: Terminology is used inconsistently

Major/Minor: Major

Explanation: The term "wife cake" is not the standard term for this food, which is "sweetheart cake".

Why is training an explainable metric challenging?

- Fine-grained Data Scarcity
- Deviation of Human rating
- Well Defined Explainability

Ideal Metric

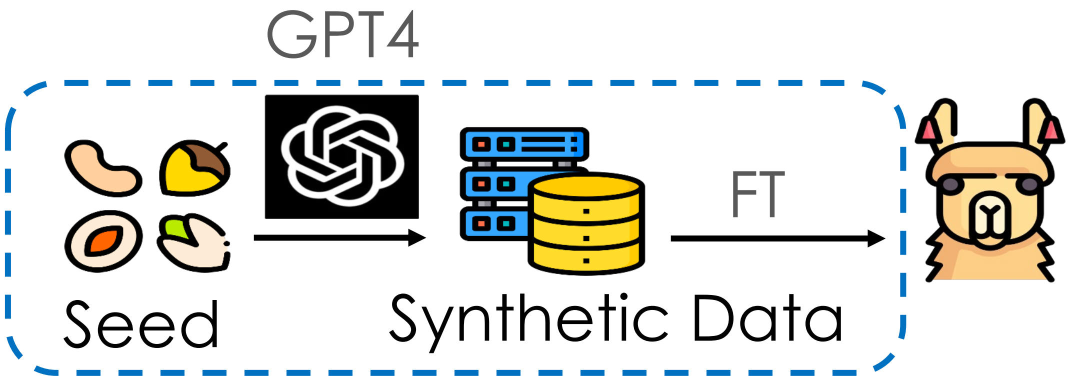
Highly Aligned with Expert Annotator

Fine-grained Explainability

Generalizable

Naive solution

Guided error-and-explanation synthesis



Derive synthetic data

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]

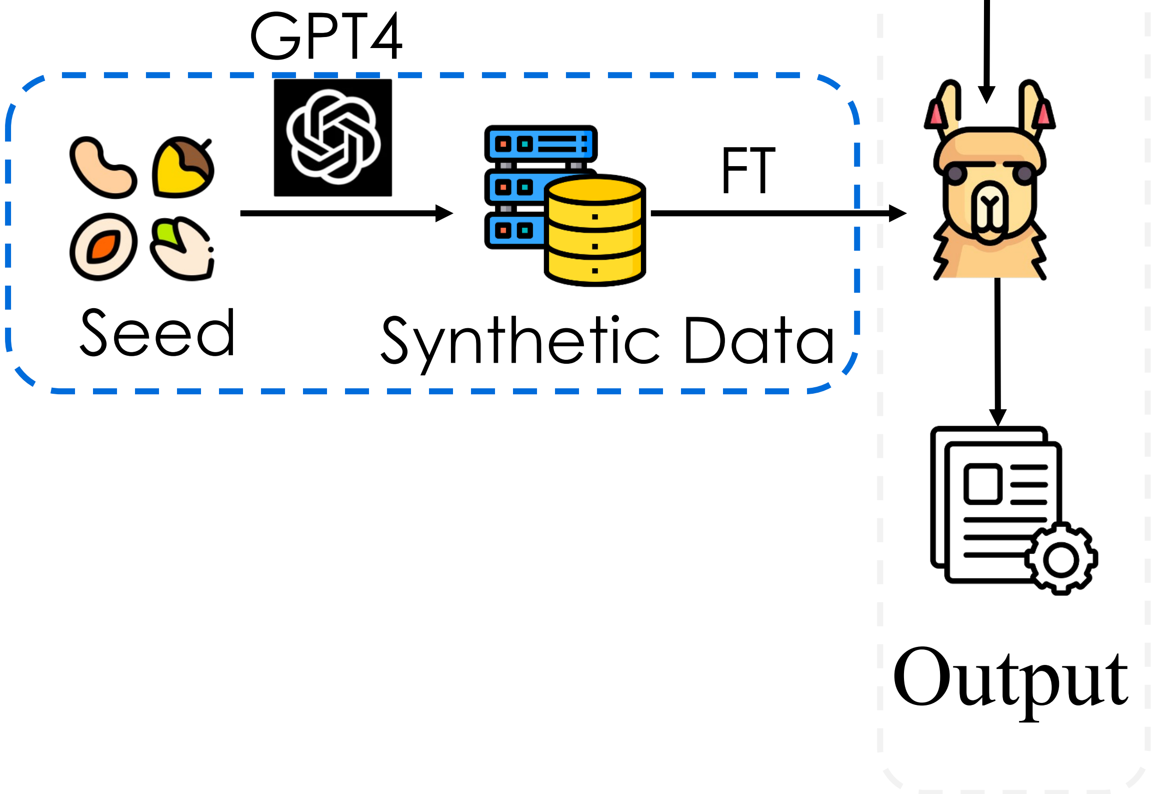
Error location 1: [GPT4 fill in]

Explanation for error 1:

[GPT4 fill in]

But, failed explanation in GPT4

Guided error-and-explanation synthesis



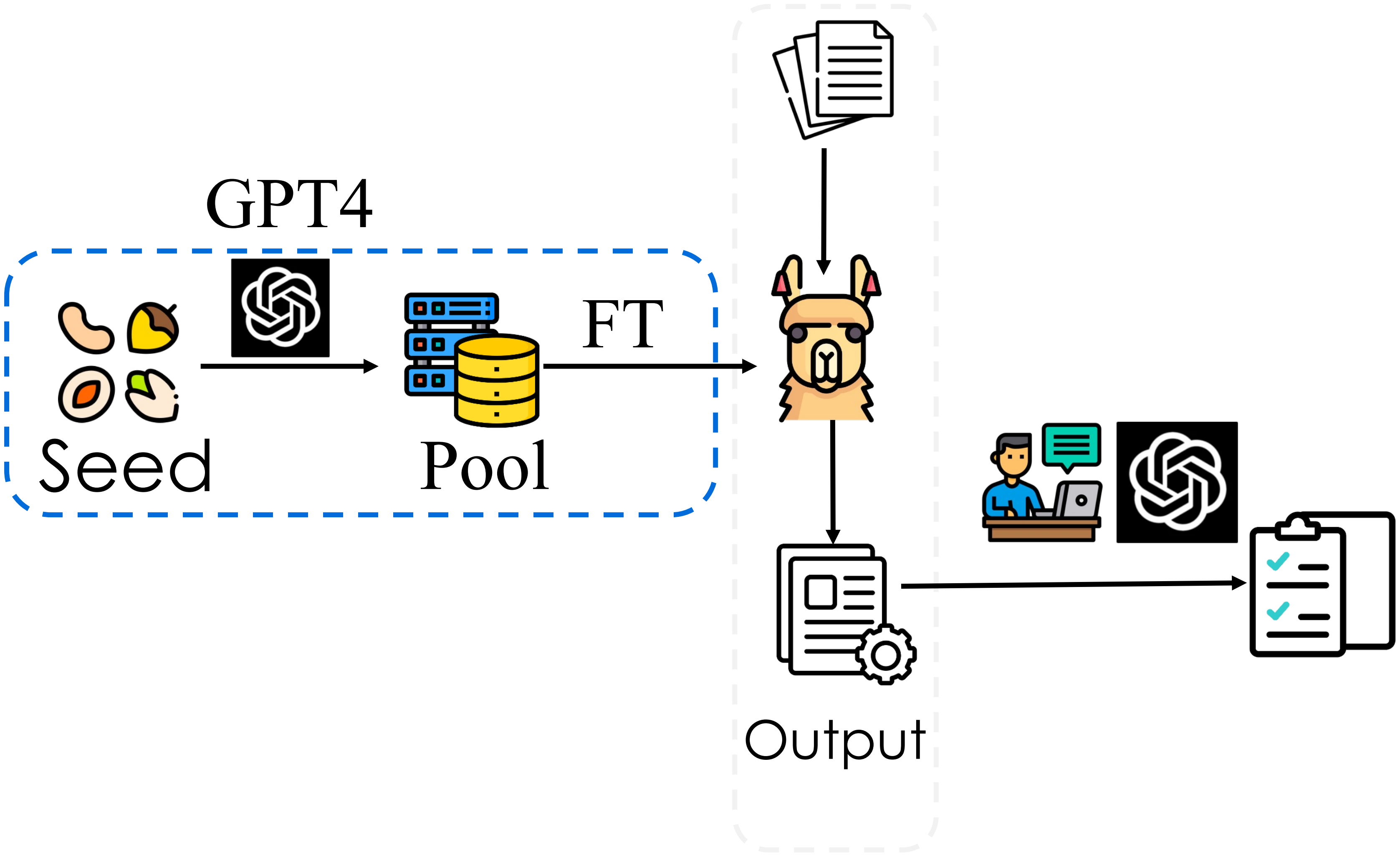
Error type 3: Missing information
Explanation for error 3: The incorrect translation adds the word "annual" to the phrase ...

Failure Mode Categorization

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

Meta-Evaluation of the Explainable Metric

Introducing InstructScore



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

InstructScore: Automatic Feedback

Reference
Candidate

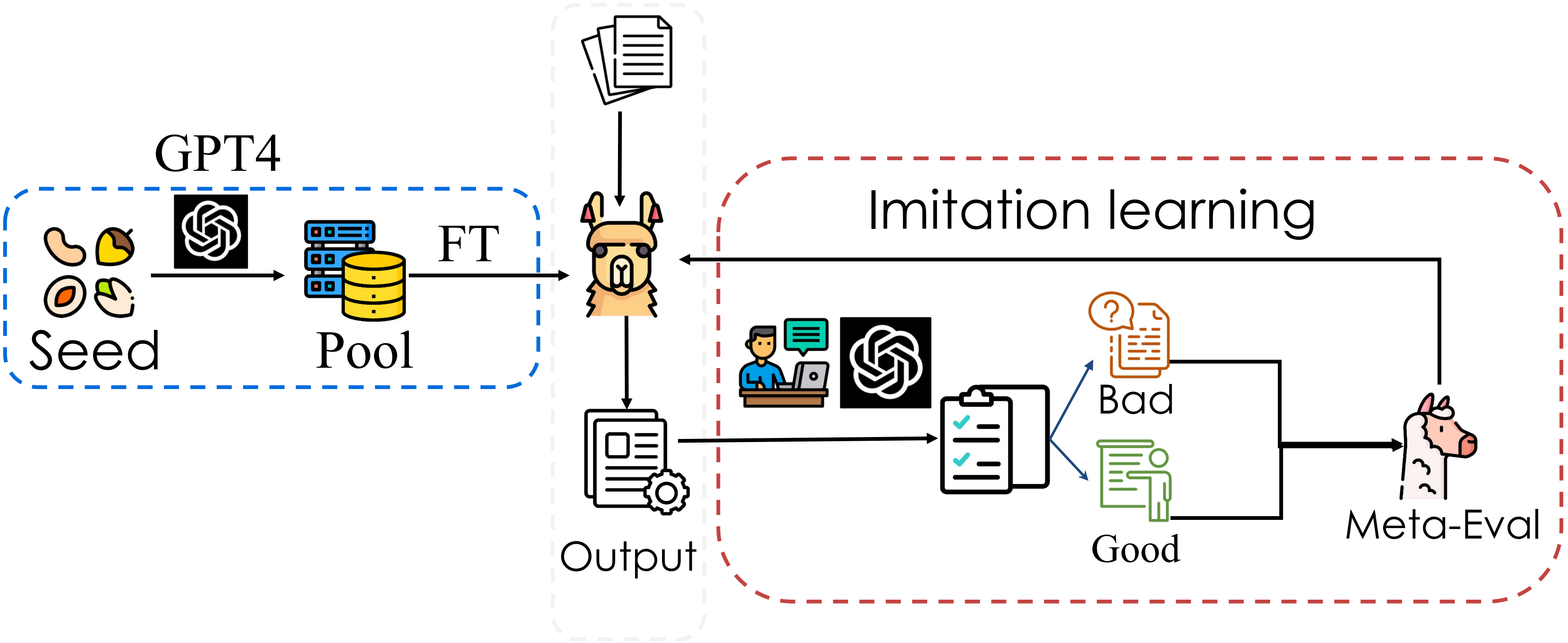
Error location1
Error Type1
Major/Minor
Explanation1

Error location2
Error Type2
Major/Minor
Explanation2

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

Alignment Score: 7/8

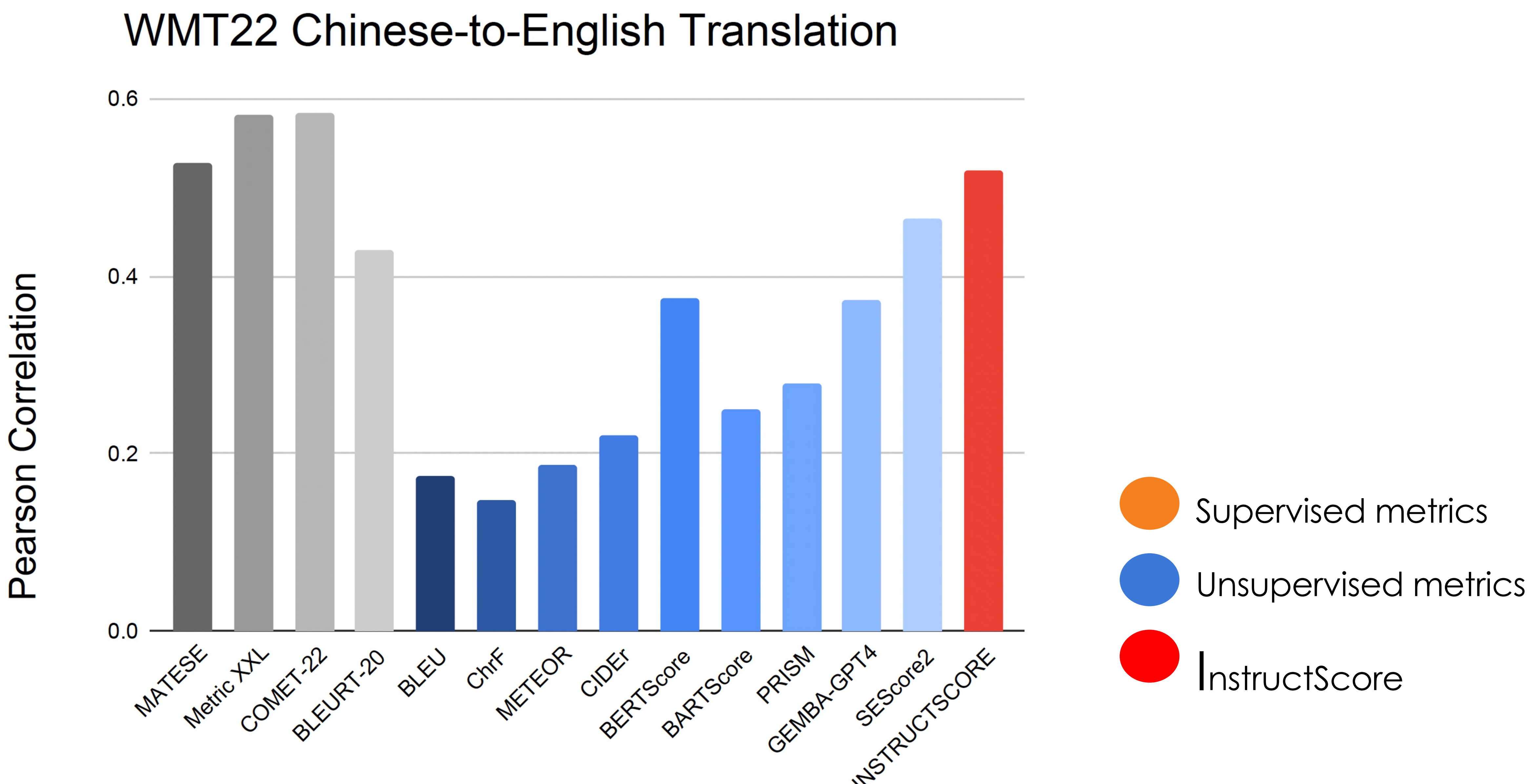
InstructScore: Refinement



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

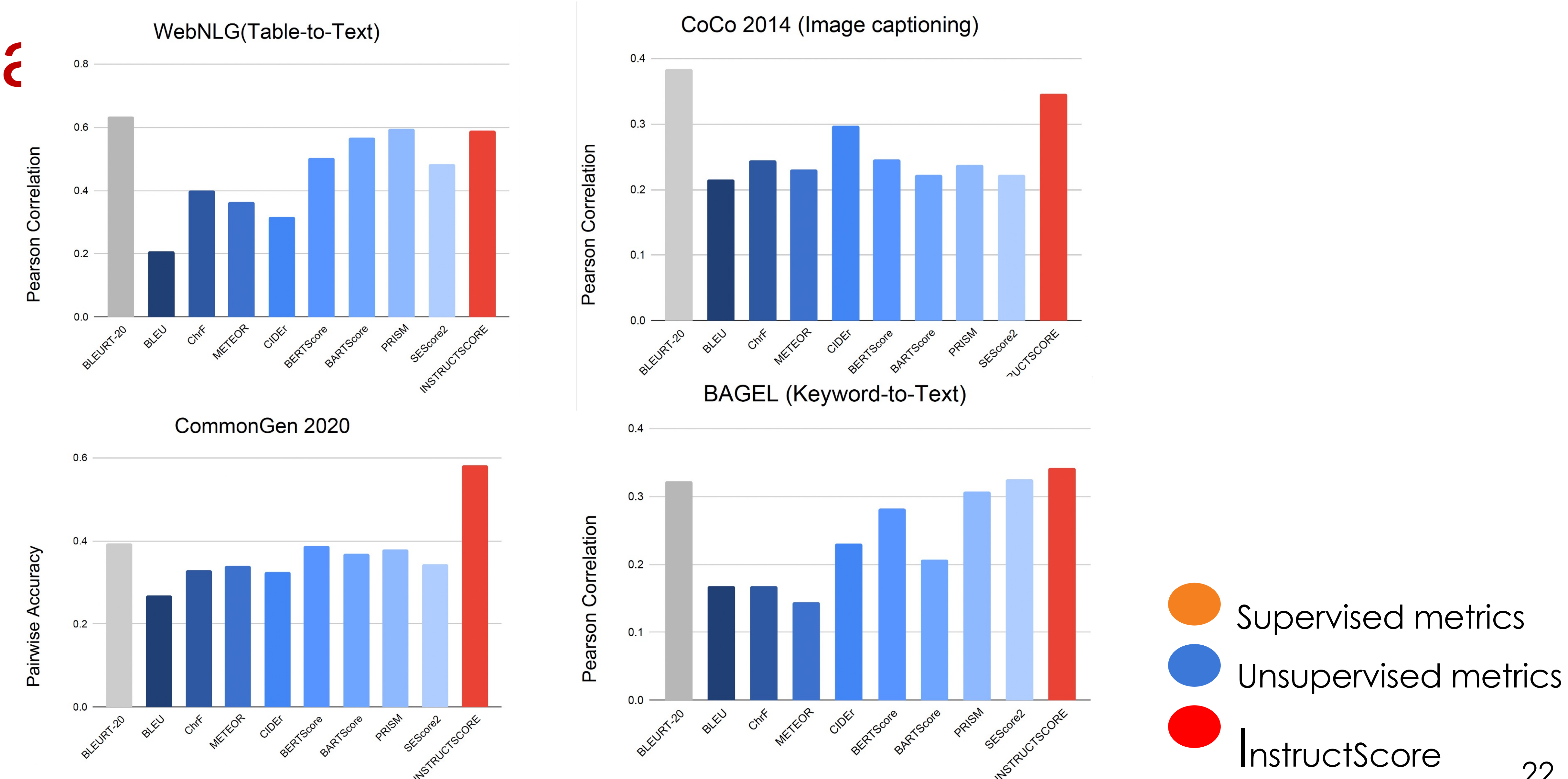
Robust Performance across Tasks (Four seen and one unseen NLG tasks)

InstructScore can judge machine translation!



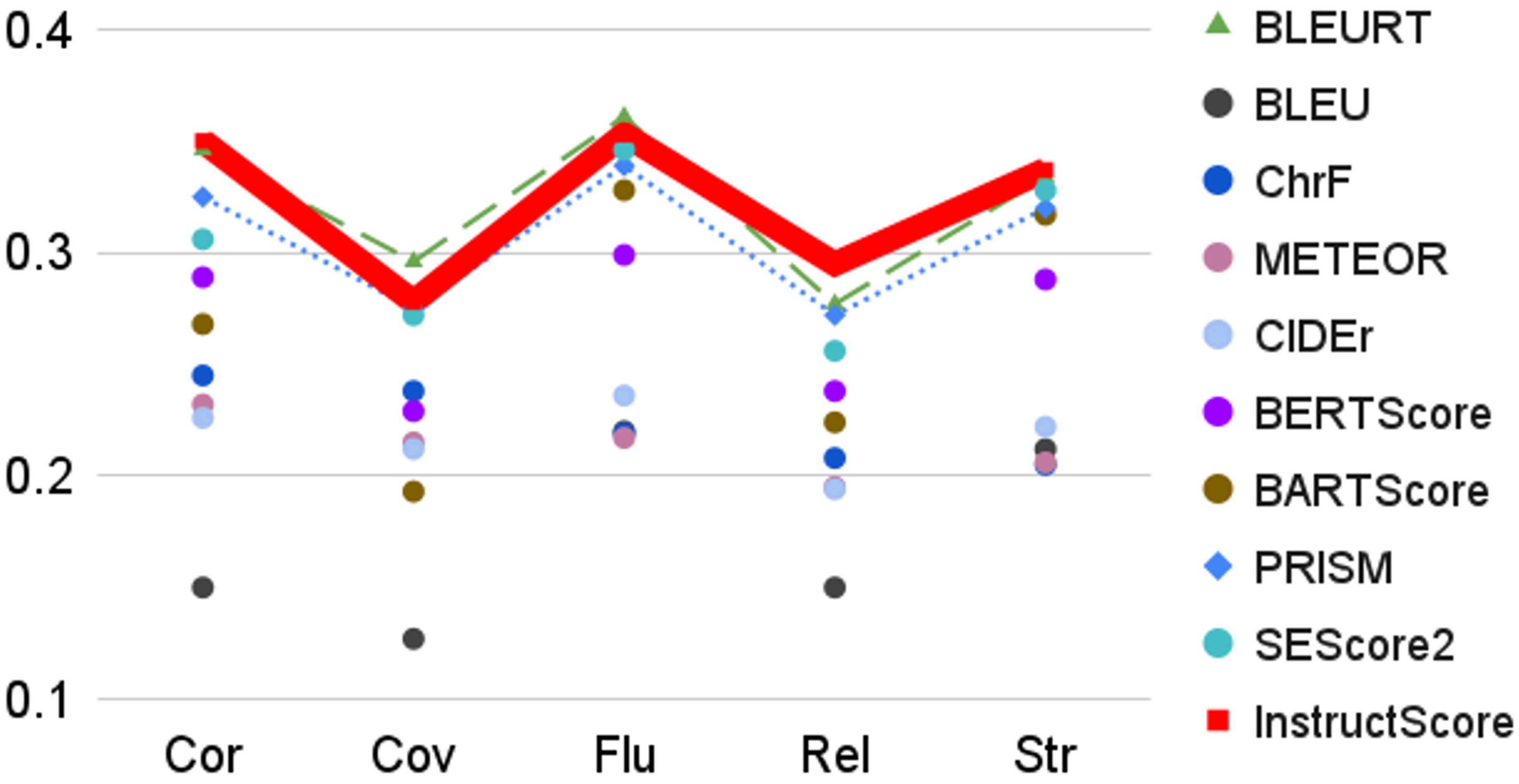
Xu, Wang, Pan, Song, Freitag, Wang, **Li**. INSTRUCTSCORE: Explainable Text Generation Evaluation with Finegrained Feedback. EMNLP 2023.

InstructScore can do well in other tasks



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

Robust Performance across Dimensions (WebNLG20)



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

1. **Error Type:** Lexical Selection
   - **Major/Minor:** Major
   - **Error Location:** "Or" instead of "And"
2. **Error Type:** Omission
   - **Major/Minor:** Major
   - **Error Location:** "how much longer does it take for" is omitted
3. **Error Type:** Lexical Selection
   - **Major/Minor:** Minor
   - **Error Location:** "the restaurant is ready" instead of "the restaurant to prepare it"
4. **Error Type:** Lexical Selection/Omission
   - **Major/Minor:** Major
   - **Error Location:** "Can you ask for me?" instead of "Could you help me ask them?"
5. **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: "Or how long before the restaurant is ready?"

Error type 2: Problems with grammar, other than orthography
Major/minor: Minor
Error location 2: "could you help me ask them?"

Key Advantages of InstructScore

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

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

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

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. ✗

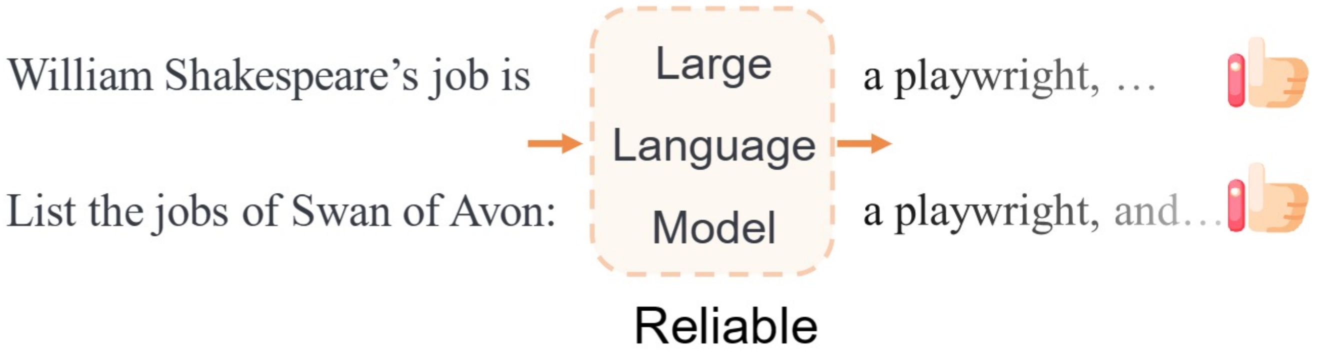
- Example2: ChatGPT

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

Is William Shakespeare a teacher?
: None. ✗

Knowledge Assessment for LLMs

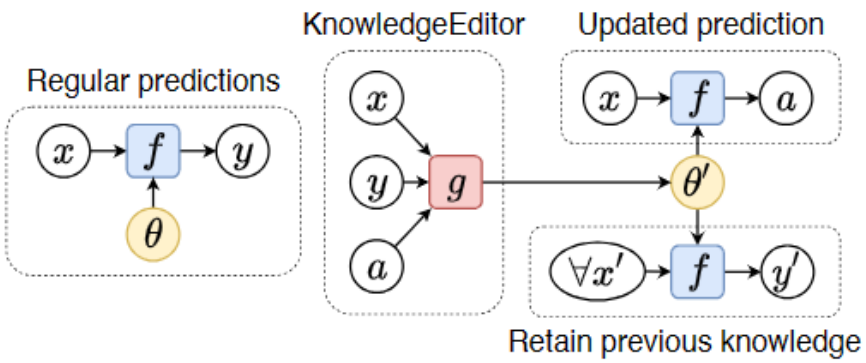
- Given varying prompts regarding a factoid question, can a LLM **reliably** generate factually **correct** answers?



Why Do We Need Knowledge Assessment?

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

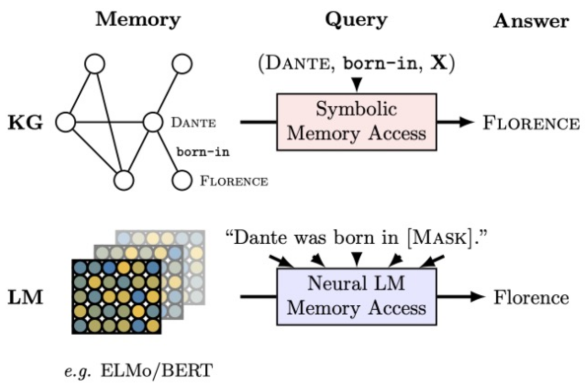
- Once we identify inconsistency of LLM generation, we could potentially correct such knowledge in LLMs.



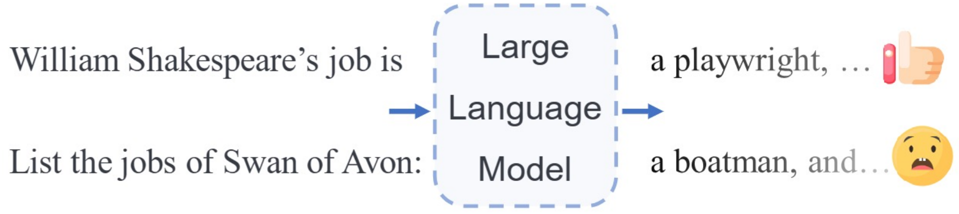
Knowledge Editing Method[1]

[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



Probing method for MLM[1]



Accurate but Unreliable

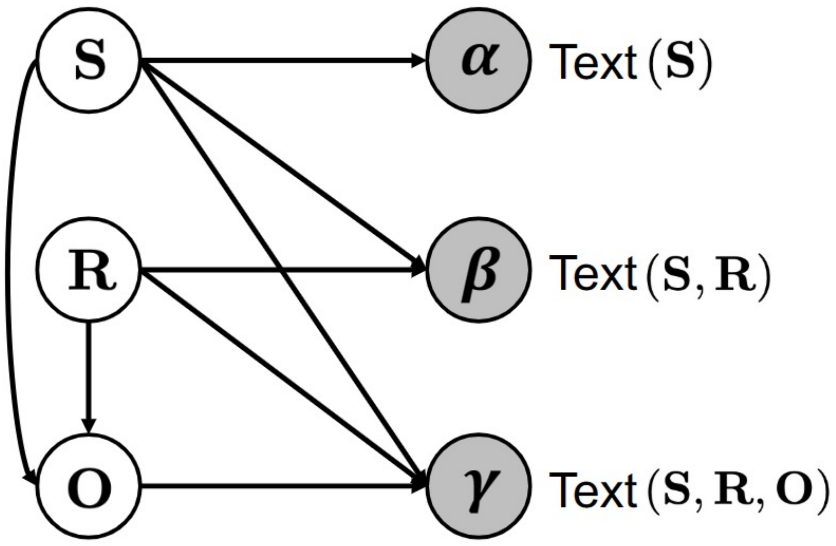
- **Accuracy v.s. Reliability:** Previous studies primarily assess accuracy, not reliability.

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

[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.

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.

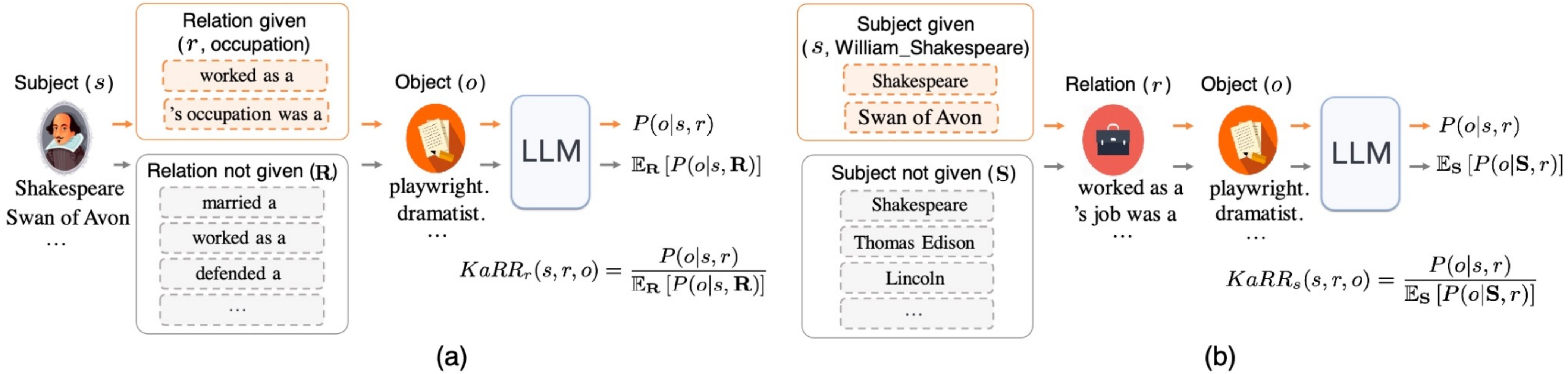
Goal: estimate the model knowledge on **symbols** through the observable model probability across diverse corresponding **textual forms**.

Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

Knowledge Assessment Risk Ratio

- Based on the graphical model, we propose Knowledge Assessment Risk Ratio (KaRR).

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



Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

Knowledge Assessment with Wide Coverage

- 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%

- Accurate

- Less Variance and Spurious Correlation

Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

Knowledge Assessment with High Human Correlation

- Good coverage

- Accurate -- strong correlation with human assessment

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

- Less Variance and Spurious Correlation

Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

Knowledge Assessment with Less Bias

- Good coverage
- Accurate
- Less Variance and Spurious Correlation

Method	Var (↓)	Std (↓)
LAMA@1	1.90	1.37
LAMA@10	5.14	2.27
ParaRel	0.77	0.94
K-Prompts	2.34	5.47
KaRR	**0.67**	**0.82**

(a) Evaluation variance towards varied prompts.

Method	SP (↓)	ΔP (↓)
LAMA@1	3.81	0.00
LAMA@10	64.29	47.31
ParaRel	2.66	-0.51
K-Prompts	**0.00**	-7.54
KaRR	1.94	**-14.94**

(b) Spurious correlation of knowledge assessment.

Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

KaRR Scores on 20 LLMs

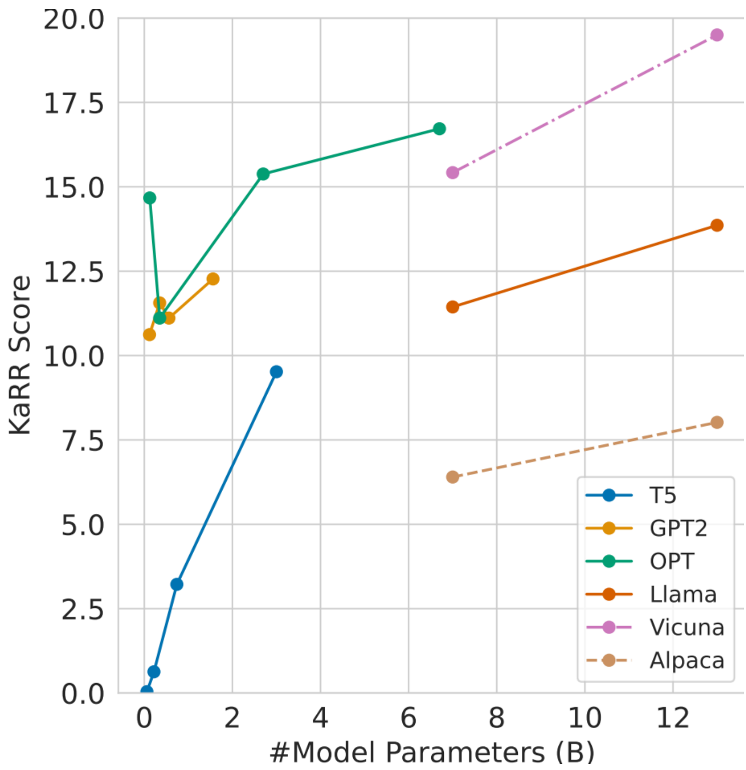
- Most small and medium-sized LLMs struggle with generating correct facts consistently.
- Vicuna's KaRR score – Finetuning LLMs with data from more knowledgeable models can enhance knowledge.

Model	Size	KaRR Score	Model	Size	KaRR Score
GPT	0.12B	9.57	GLM	10B	5.59
XLNet	0.12B	5.86	Dolly	12B	15.60
T5-large	0.74B	3.22	LLaMA	13B	13.86
Phi-1.5	1.3B	10.58	Alpaca	13B	8.24
GPT2-XL	1.56B	12.27	Vicuna	13B	19.50
GPT-NEO	2.65B	13.44	WizardLM	13B	16.90
T5-3B	3B	9.52	Moss	16B	11.20
Falcon	7B	7.97	LLaMA	65B	14.56
BLOOM	7B	7.72	LLaMA2	65B	19.71
LLaMA	7B	12.37	OPT	175B	23.06

Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

Scaling Effect on Knowledge

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

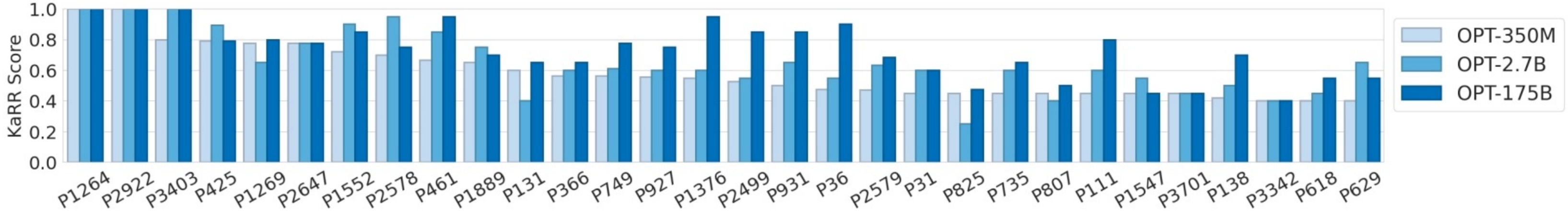


Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

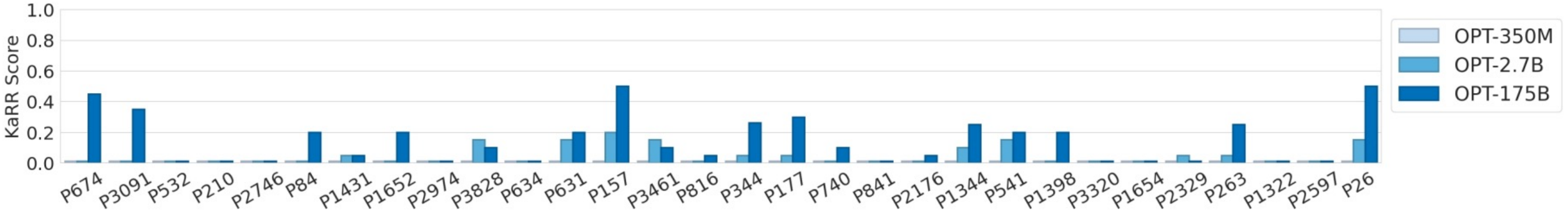
Scaling Effect on Knowledge

- Larger models exhibit better and more consistent knowledge-correct generation ability.

- Larger models surpass small models in terms of knowledge on a wider range of relations.



KaRR scores on the 30 best known relations of OPT-350M



KaRR scores on the 30 best known relations of OPT-350M

Dong, Xu, Kong, Sui, **Li**. Statistical Knowledge Assessment for Large Language Models. NeurIPS 2023

Summary and takeaway of KaRR

- Distinguishing the knowledge symbols and textual forms helps us build the graphical model for knowledge assessment.

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

- Larger models exhibit better and more consistent knowledge-correct generation ability.

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 "一个餐等了一个半小时。" into English.

LLM's output:

A meal had been waiting for an hour and a half.

What feedback can we give to LLM?

Can we use fine-grained feedback to guide LLM?

Input: Translate "一个餐等了一个半小时。" into English.

LLM's output:

A meal had been waiting for an hour and a half.

Ask LLM to improve?

Source: 一个餐等了一个半小时。
Translation: A meal had been waiting for an hour and a half.
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 "一个餐等了一个半小时。" into English.

LLM's output:
A meal had been waiting for an hour and a half.

Use binary feedback to guide LLM?

Source: 一个餐等了一个半小时。
Translation: A meal had been waiting for an hour and a half.
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 "一个餐等了一个半小时。" into English.

LLM's output:

A meal had been waiting for an hour and a half.

Use scalar feedback to guide LLM?

Source: 一个餐等了一个半小时。
Translation: A meal had been waiting for an hour and a half.
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 "一个餐等了一个半小时。" into English.

LLM's output:

A meal had been waiting for an hour and a half.

Use fine-grained feedback to guide LLM!

Source: 一个餐等了一个半小时。

Translation: A meal had been waiting for an hour and a half.

"A meal has been waiting" is a major mistranslation 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: 一个餐等了一个半小时。
Translation: A meal had been waiting for an hour and a half.
"A meal has been waiting" is a major mistranslation error.
Please improve current translation.

LLM's proposal:
A meal waited an hour and a half.

Reject

resample from LLM

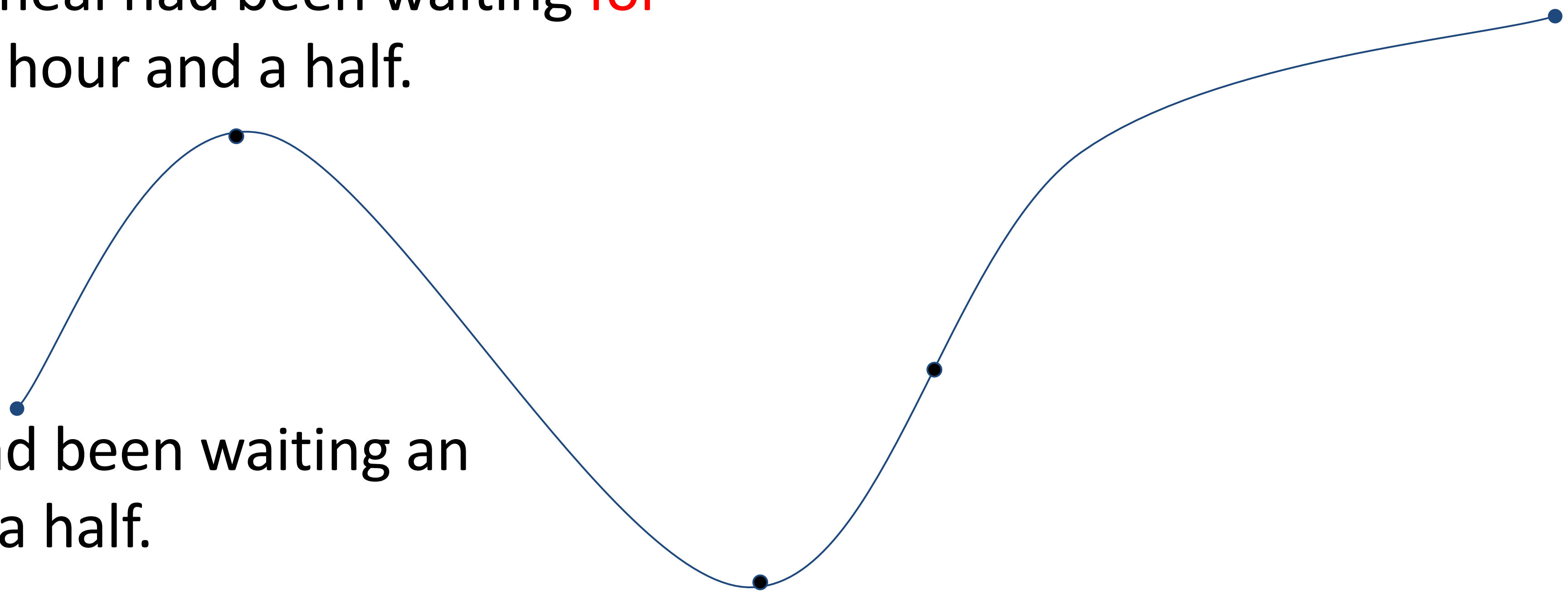
Accept

Repeat above steps for n iterations

LLM's final output:
I've waited one and half hours for one meal.

Source Translation: 一个餐等了一个半小时。

A meal had been waiting for an hour and a half.



A meal had been waiting an hour and a half.

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

Algorithm 1: Simulated Annealing for Iterative Refinement

Input: Input prompt x, Feedback model F, Base model M

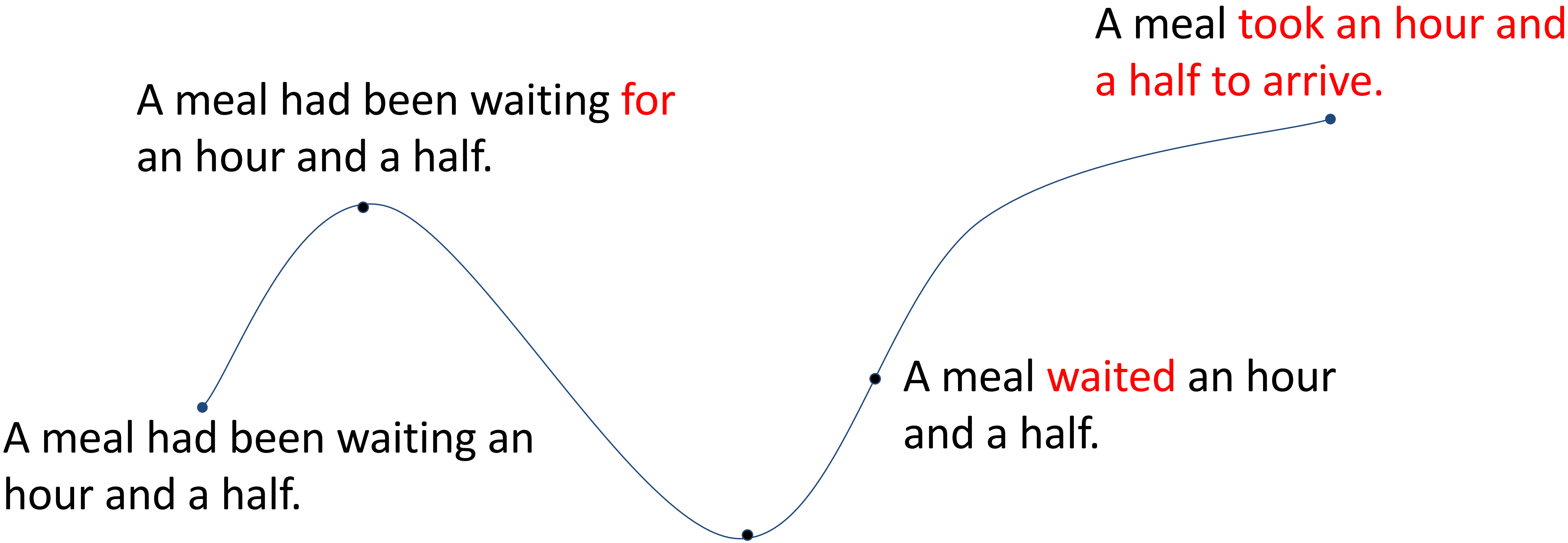
Initialize: $y_0 \leftarrow greedy\_decode(M(x)), T_0, n$ # Initialize candidate, temperature, constant

for $i = 0..n$ **do**

- $f_i \leftarrow F(x, y_i)$ # generate feedback for the current candidate proposal
- $c_i \leftarrow Sampling(M(x, y_i, f_i))$ # Sample next candidate based on prior one and feedback
- $p_{\text{acc}} \leftarrow \min(1, e^{\frac{s(F(c_i)) - s(F(y_i))}{n * T_i}})$
- **if** *Accept* **then**
  - $y_{i+1} \leftarrow c_i$
- **else**
  - $y_{i+1} \leftarrow y_i$
- $T_{i+1} = max(T_i - c * T_i, 0)$ # update temperature for the next iteration

Output: Sampled sequence y_n with n iterations

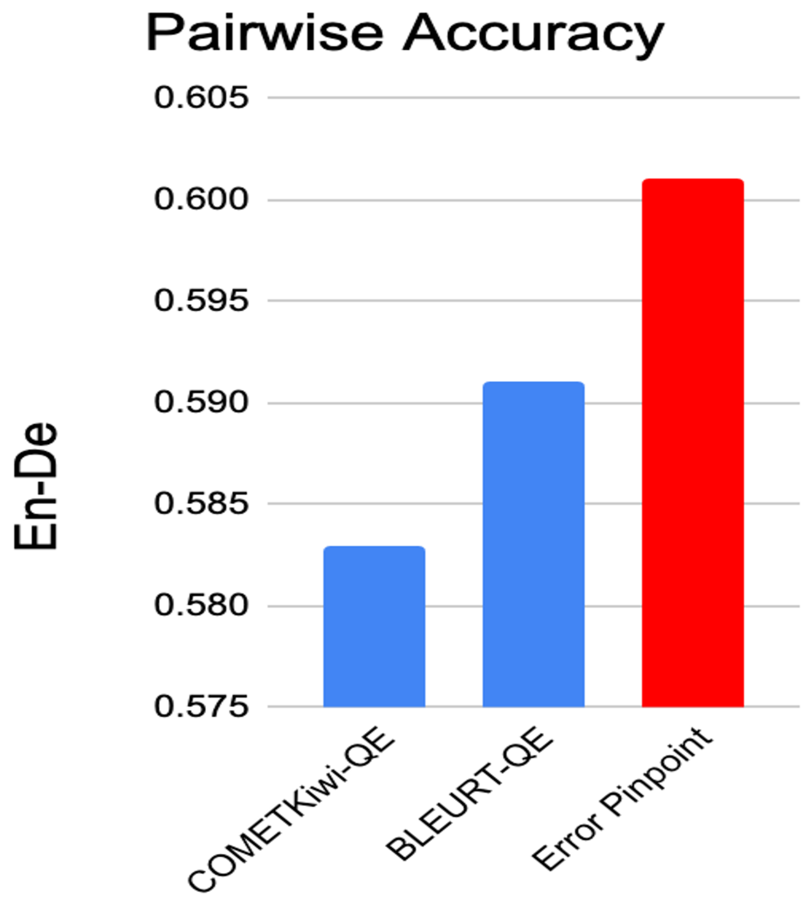
Source Translation: 一个餐等了一个半小时。

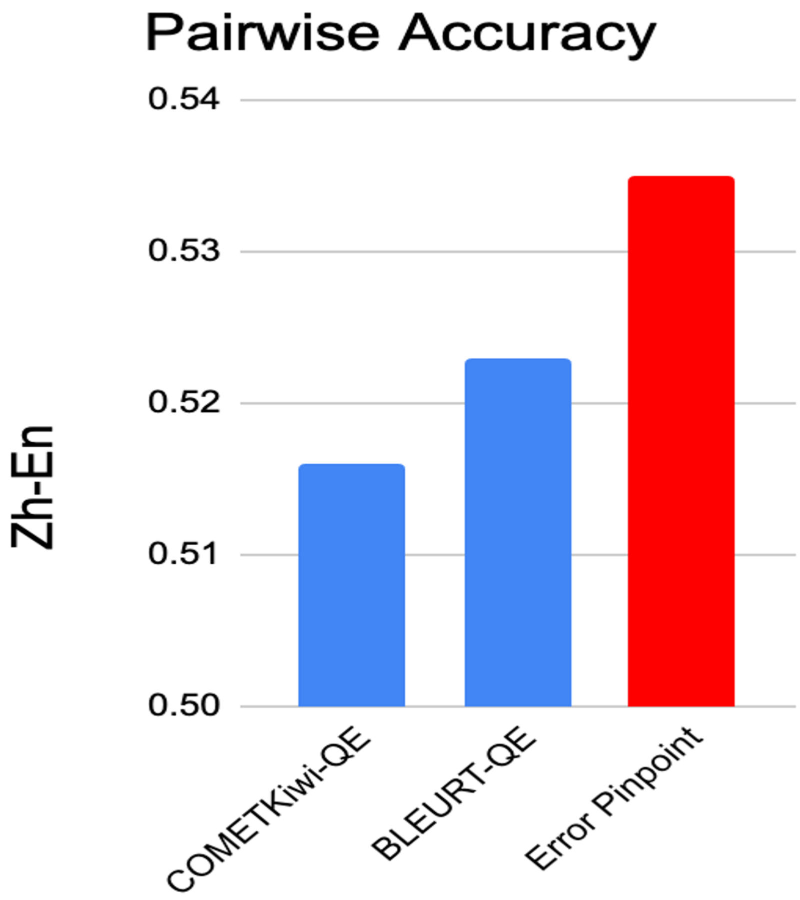


COT: "A meal had been waiting for an hour and a half." is a major mistranslation error.The correct translation should be: "A meal waited an hour and a half."

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

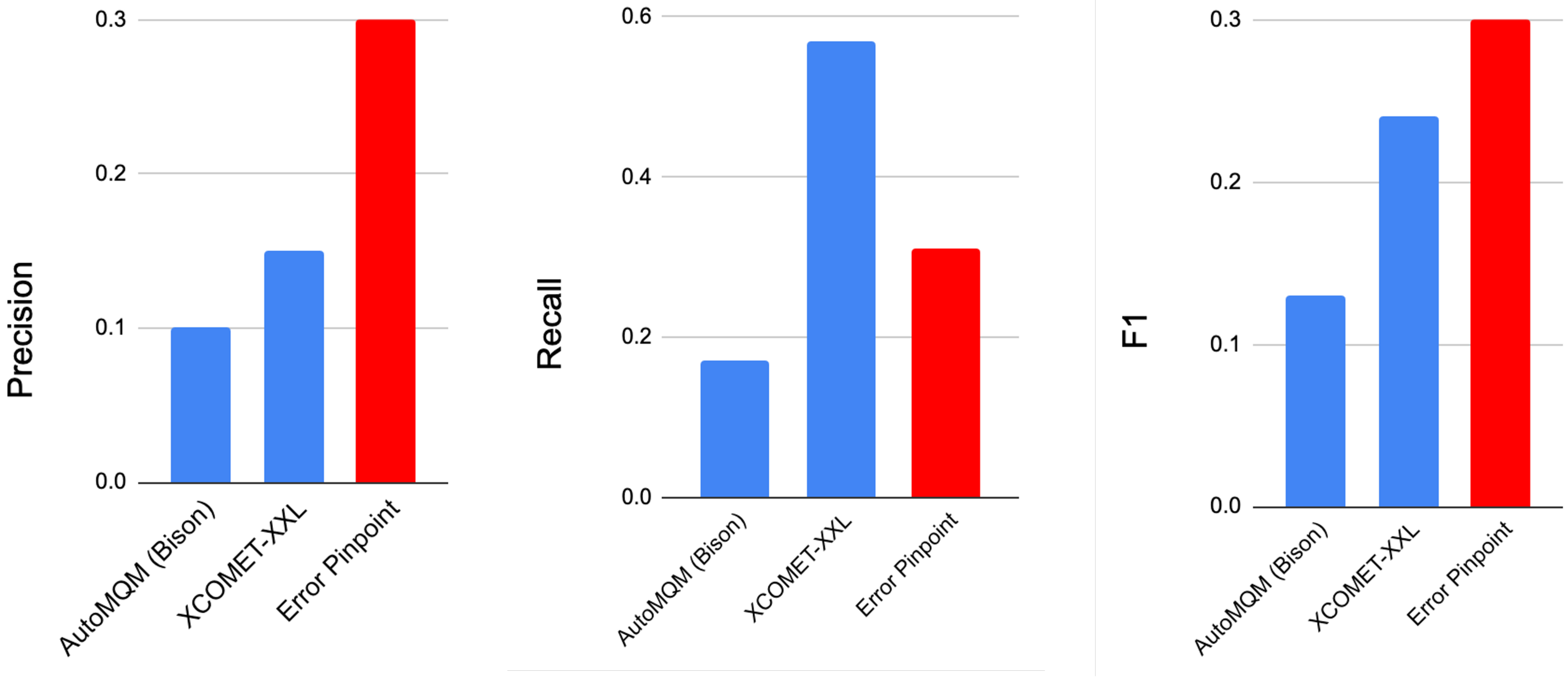
Our correlation to human judgements are high!



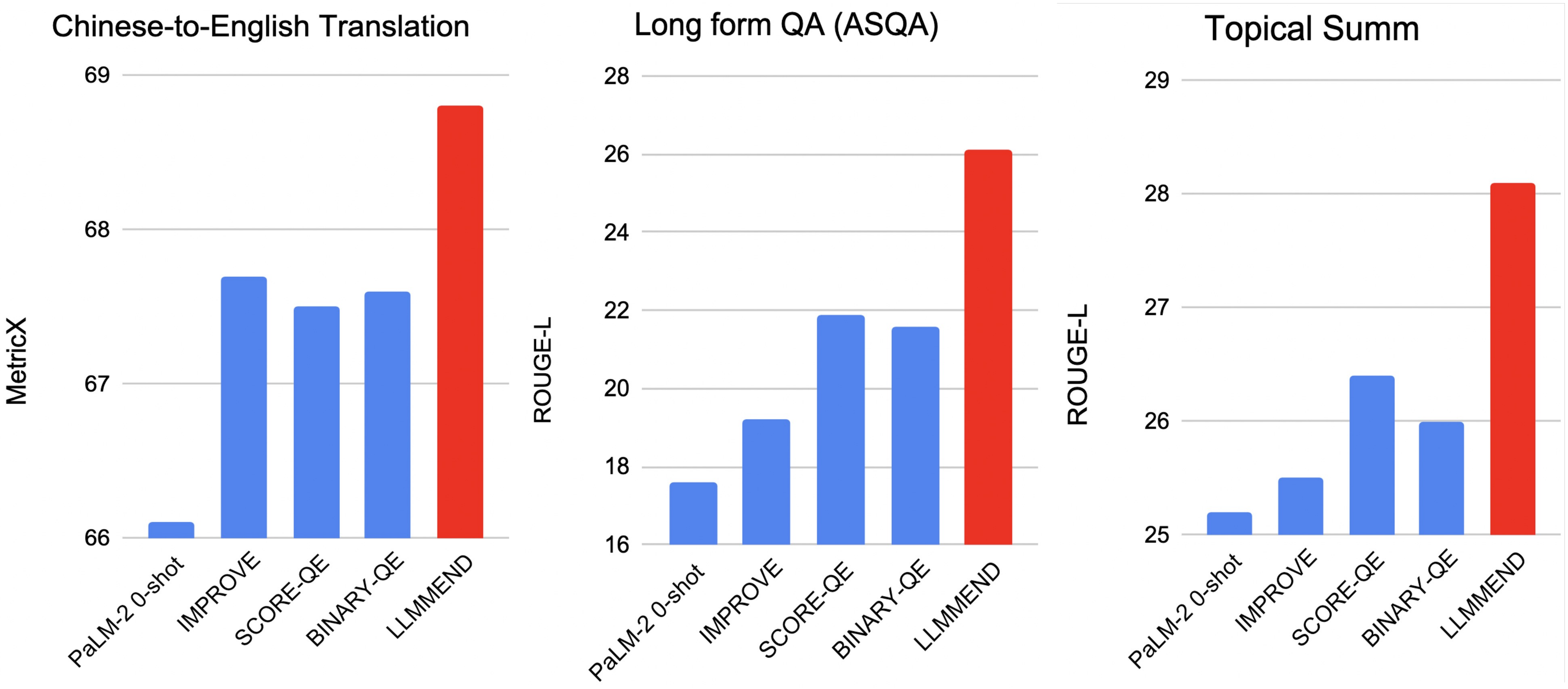


RQ1: How well does our error pinpoint model align with human annotations of translation quality?

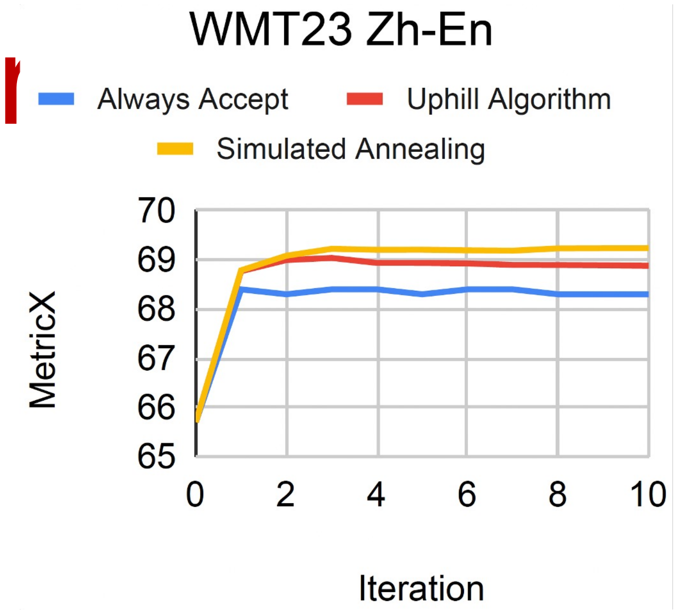
Our span-level precision and F1 are high at Chinese-to-English

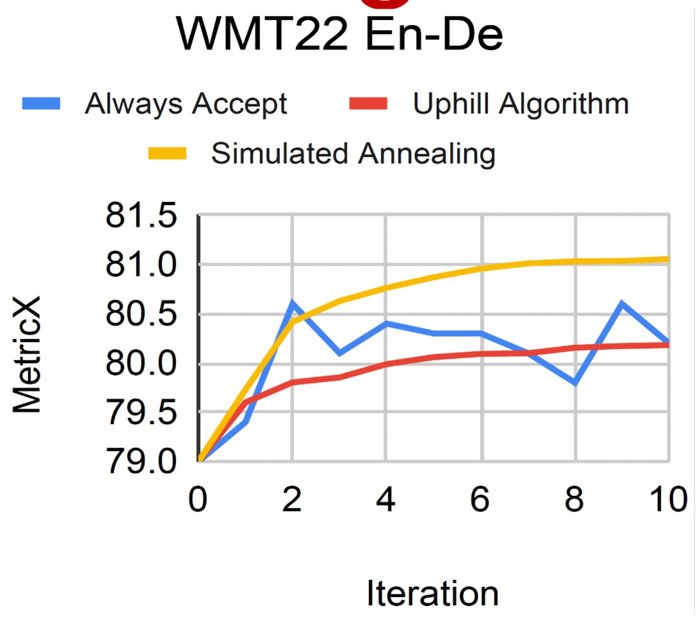


RQ2: Does fine-grained feedback result in better downstream translations than more coarse feedback?

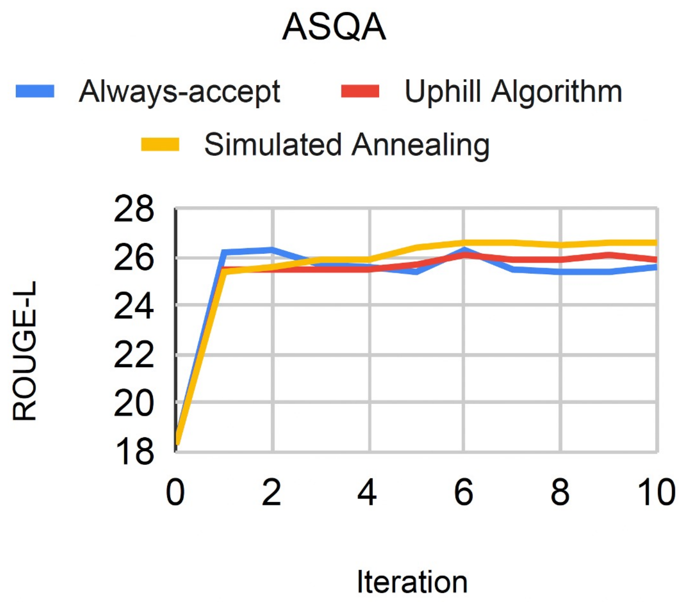


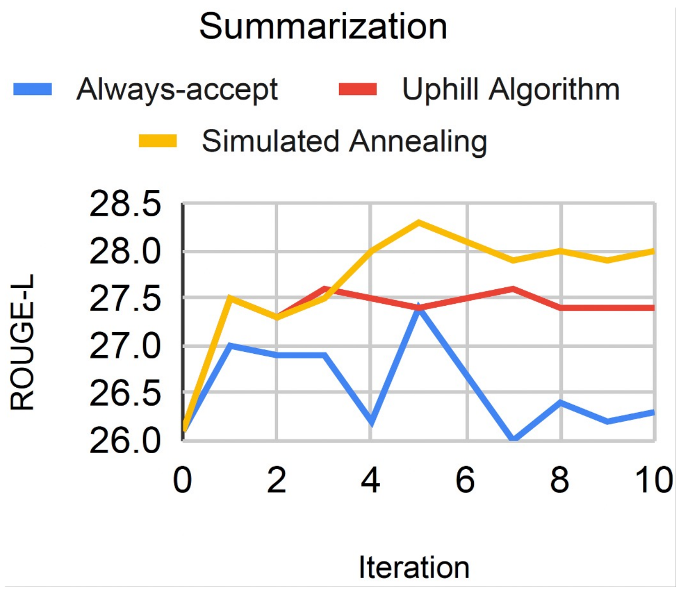
Simulated Annealing can boost iterative



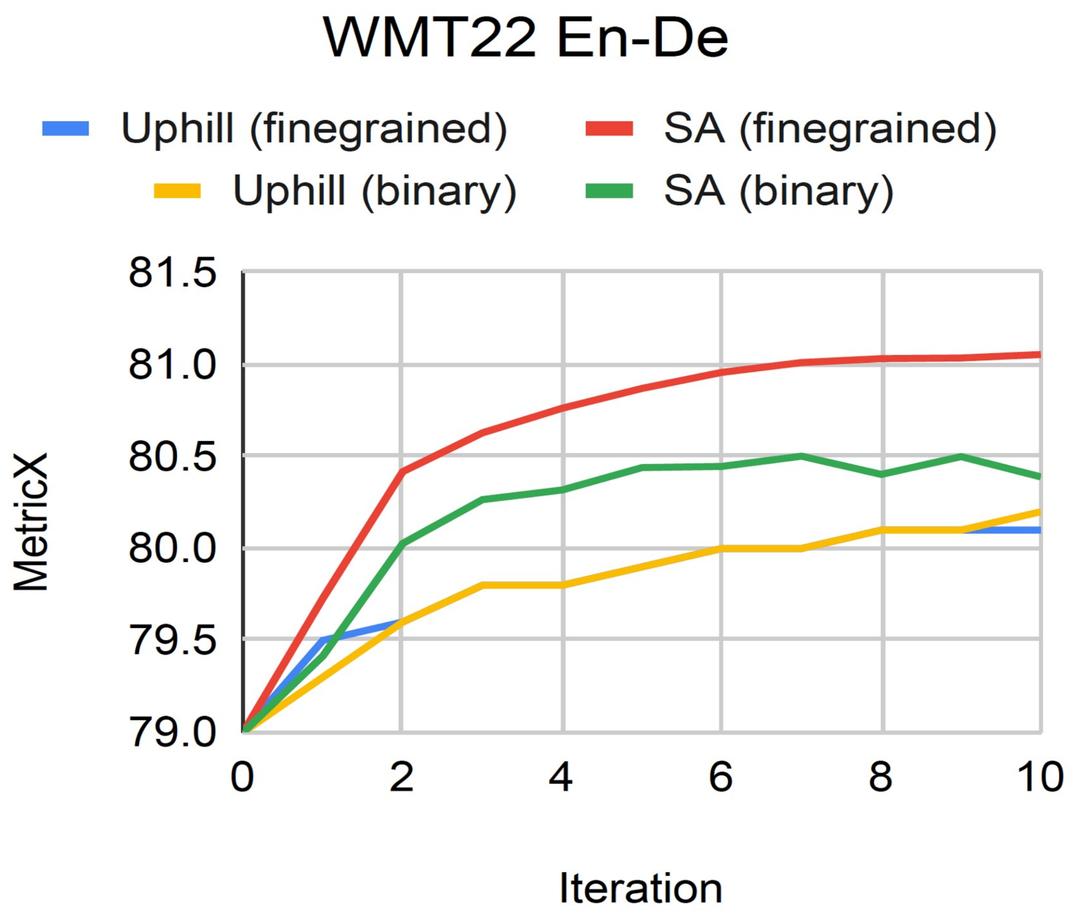


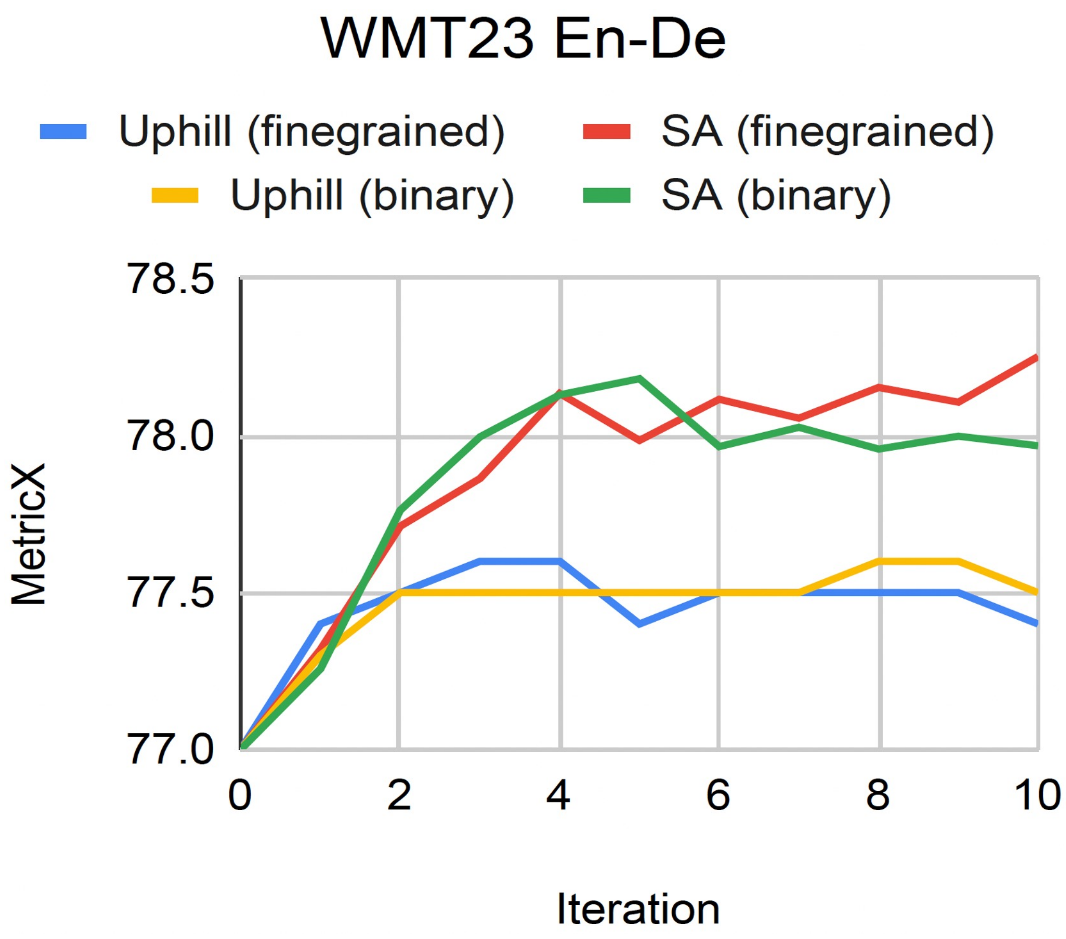
Simulated annealing outperforms always-accept and uphill algorithm significantly across MT, Summ and QA





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