Unsupervised and Explainable Text Generation Evaluation

Lei Li

Joint work with **Wenda Xu**, Xian Qian, Mingxuan Wang, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Wang

Is Google translate right?

google translate

The correct translation is:

Wife cake != Sweetheart cake

Is there a wife in the Sweetheart cake?

What does InstructScore comment on this mistake?

Output: Is there a wife in the wife cake?

Reference: Is there a wife in the Sweetheart cake?

Your Translation contains 1 error:

Error type 1: Terminology is non-standard or does not fit context. Terminology is used inconsistently

Major/minor: Major

Error location 1: "wife cake"

Explanation for error 1: The term "wife cake" is not the standard term for this food, which is "sweetheart cake".

What if we don’t understand target language?

Reference: Diese Vakuumpumpe mit Mehrfachbetrieb hat einen speziellen
Output: Diese mehrfach angetriebene Vakuumpumpe verfügt über einen speziellen Griff

Your Translation contains 2 errors:
Error type 1: Incorrect translation is missing content from the correct translation
Major/minor: Major
Error location 1: "mit Mehrfachbetrieb" is missing in the incorrect translation
Explanation for error 1: The incorrect translation omits the important information that the vacuum pump is multifunctional, which changes the meaning of the sentence.
……

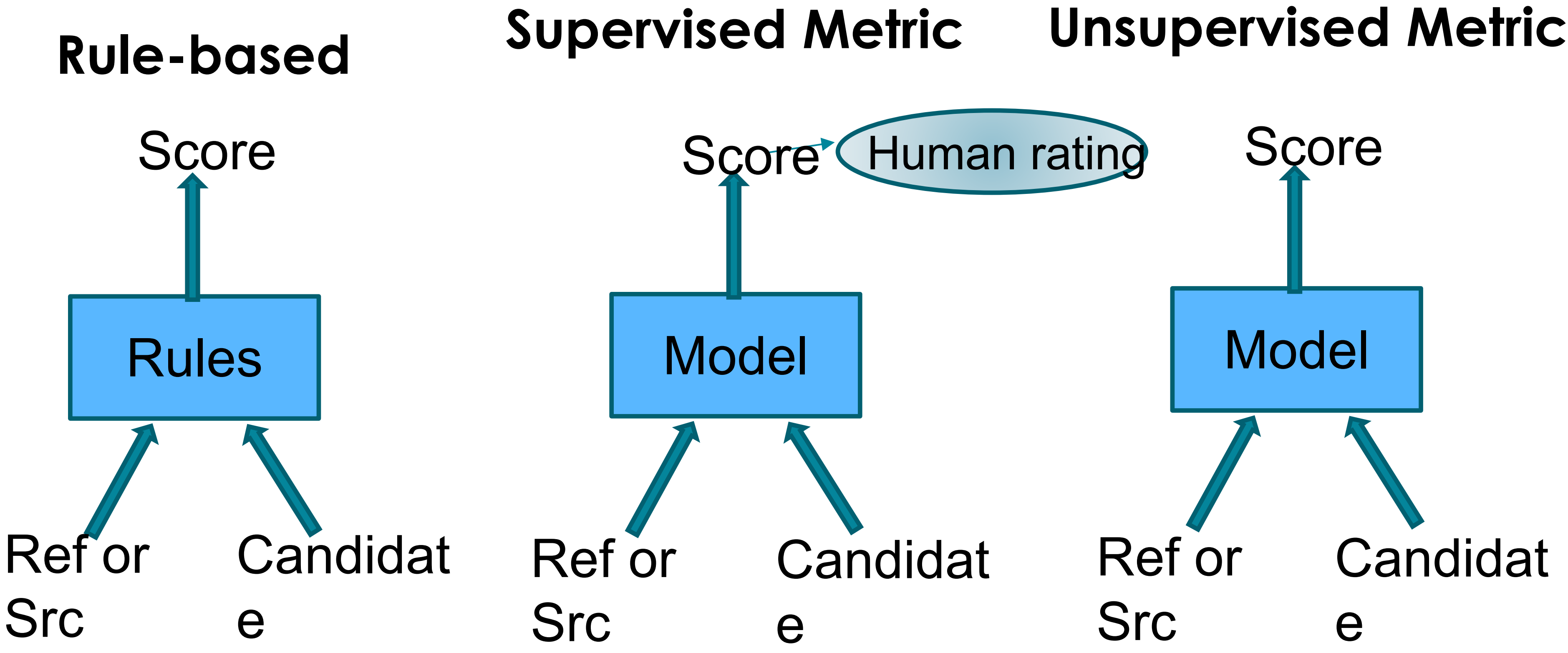
NLG Evaluation: Observations and Challenges

1. In the age of LLMs, no matter what you work on, it boils down to EVALUATION.

1. In the past, even though everyone knows the limitations of BLEU, people still used it for MT for 20 years. It was even used in dialogue, data-to-text generation, and many other tasks.

1. I argue that people can no longer use it anymore: the fundamental advantage of LLMs is the long and diverse output (OOD), and with long/diverse outputs, **BLEU/ROUGE will have significantly decreased correlations with human judgments.**

Rule-based vs Learned Metric



Rule-based

Score

Rules

Ref or Src

Candidat e

Supervised Metric

Score

Human rating

Model

Ref or Src

Candidat e

Unsupervised Metric

Score

Model

Ref or Src

Candidat e

Challenge: Learning Unsupervised Metric

Data Scarcity - lack of large annotated human ratings

<He is a dog person, He is a cat person, -5>

Deviation from expert ratings

Lack of Generalization Capability

Ideal Metric

- Highly Aligned with Expert
- Unsupervised
- Generalizable

WMT rating data – 400K

Rule-based vs Learned Metric

Rule-based

- BLEU
- chrF
- TER
- ROUGE

Surface form difference 💣

Supervised Metric

- BLEURT
- COMET

Overfitting 💣

Unsupervised Metric

- BERTScore
- PRISM
- BARTScore

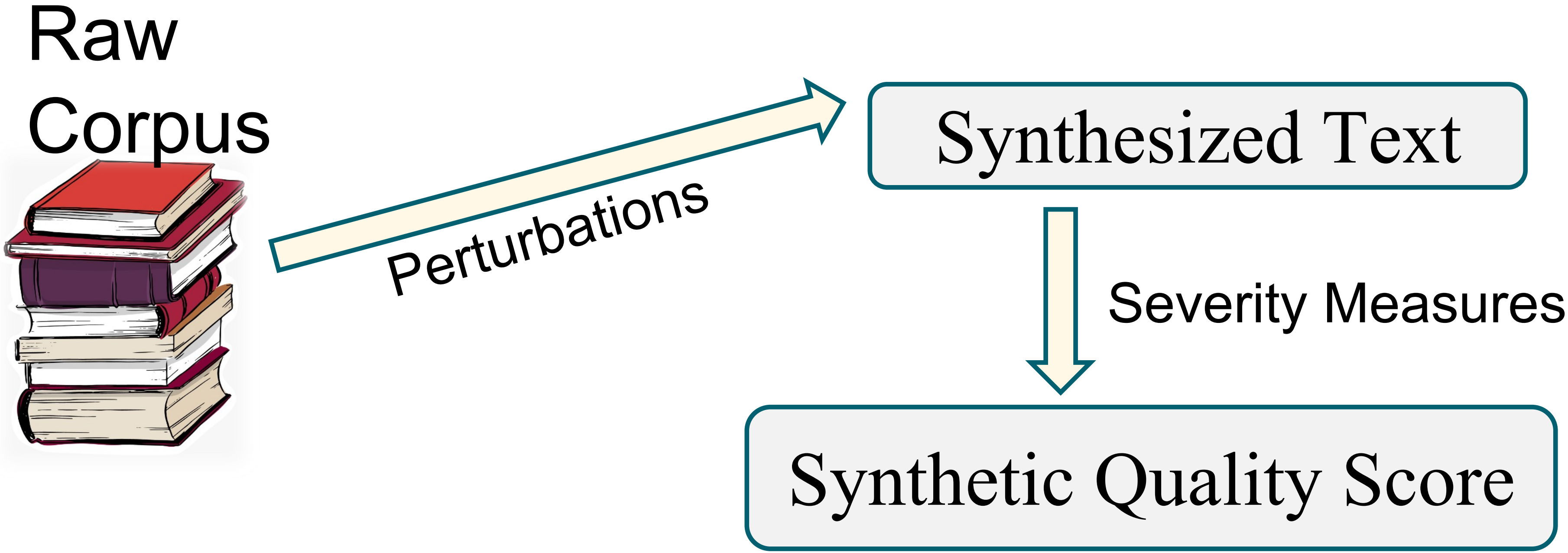
Deviate from human judgements 💣

SEScore1&2: Unsupervised Metric

1. Identify individual errors in the model outputs and Judge severity level of each error
2. Use data without human ratings to train a text generation metric
3. Large-scale synthetic data pretraining

SEScore: Learning Text Generation Metrics using Stratified Error Synthesis (Xu et al., EMNLP 2022)
SEScore 2: Retrieval-Augmented Error Synthesis (Xu et al., ACL 2023)

Our proposed method – SEScore1&2



Raw Corpus

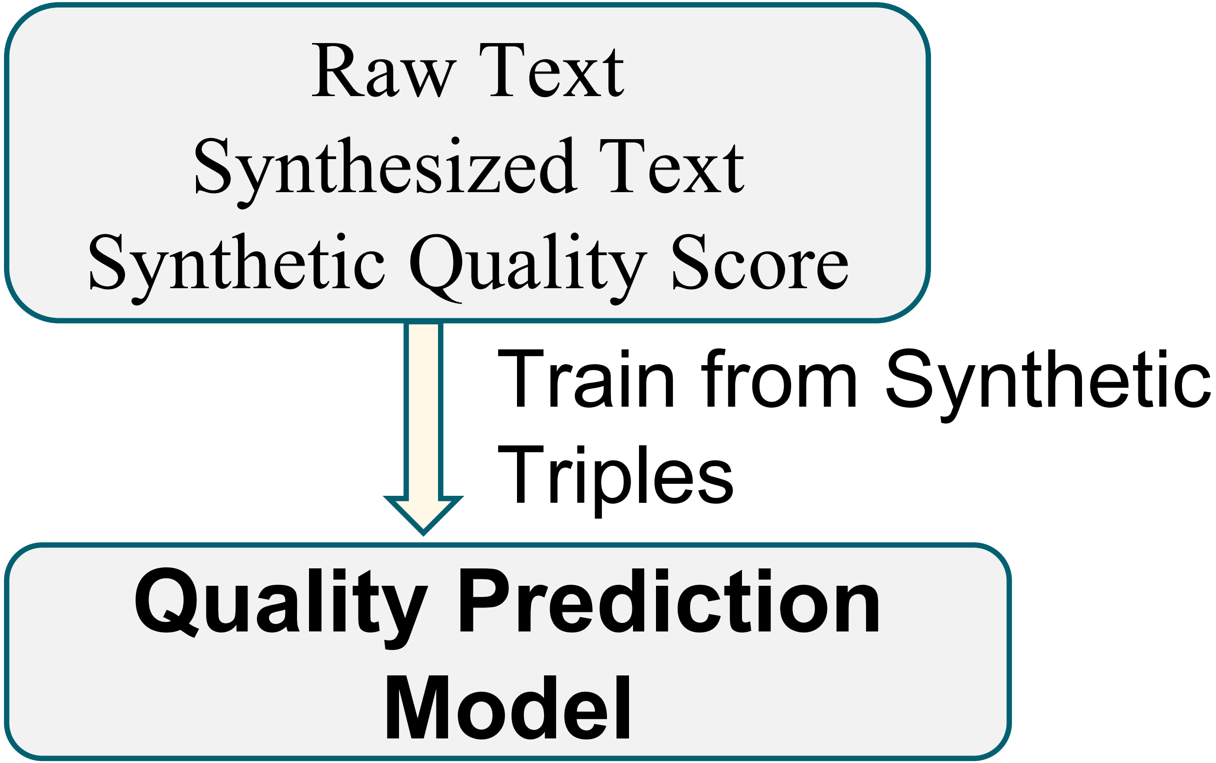
Perturbations

Synthesized Text

Severity Measures

Synthetic Quality Score

Our proposed method - SEScore



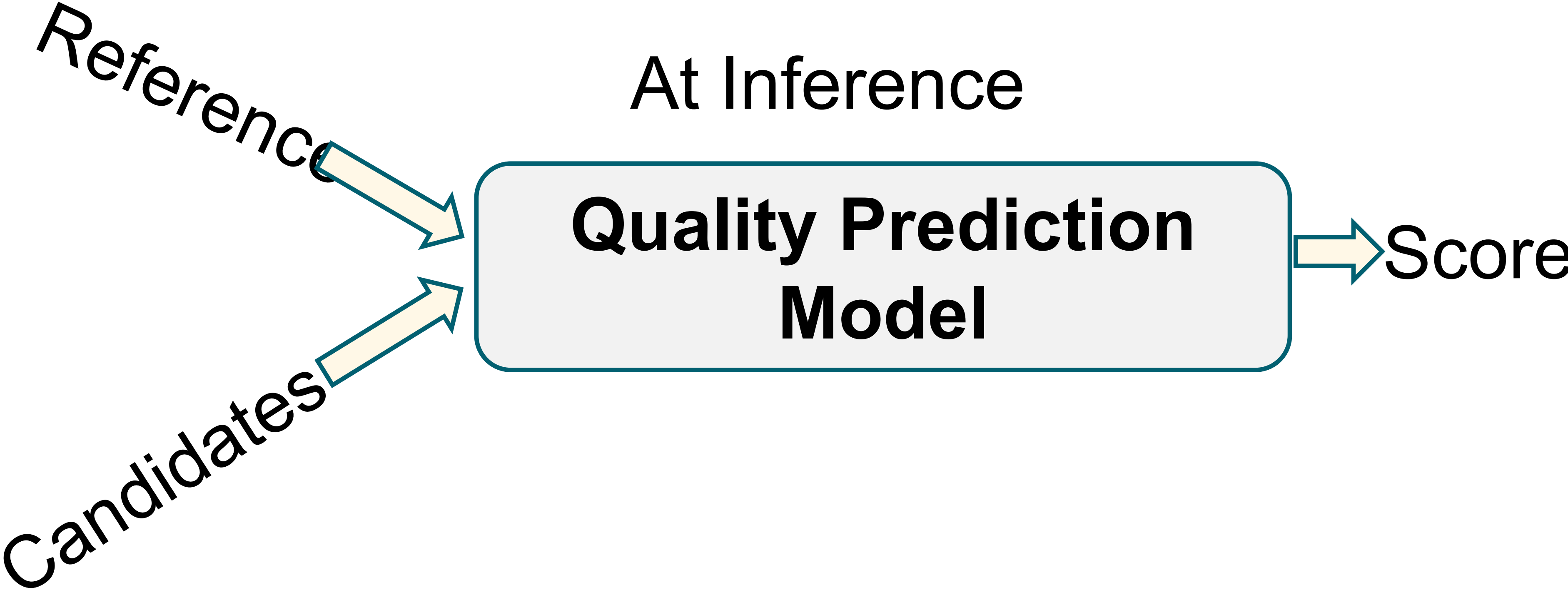
Raw Text
Synthesized Text
Synthetic Quality Score

Train from Synthetic Triples

Quality Prediction Model

Our proposed method - SEScore

At Inference



Reference

Candidates

Quality Prediction Model

Score

Can we mimic those error types and grading?

Reference: He will not accept it because he will not like it

Candidate Text: will He accept it because he hates the plan he will not fancy it

Human Score: -16

Can we synthesize realistic model mistakes?

Stratified Error Synthesis Overview (SEScore1)

Raw text (x_{raw}): He will not accept it because he will not like it

Insert

Step1 Insertion: He will not accept it because **he hates the plan** he will not like it ⇒ Major

Stratified Error Synthesis Overview (SEScore1)

Raw text (x_{raw}): He will not accept it because he will not like it

Step1 Insertion: He will not accept it because he hates the plan he will not like it

Delete

Step2 Deletion: He will accept it because he hates the plan he will not like it → Major

Stratified Error Synthesis Overview (SEScore1)

Raw text (x_{raw}): He will not accept it because he will not like it

Step2 Deletion: He will accept it because he hates the plan he will not like it

Replace

Step3 Replace: He will accept it because he hates the plan he will not fancy it ⟹ Minor

Stratified Error Synthesis Overview (SEScore1)

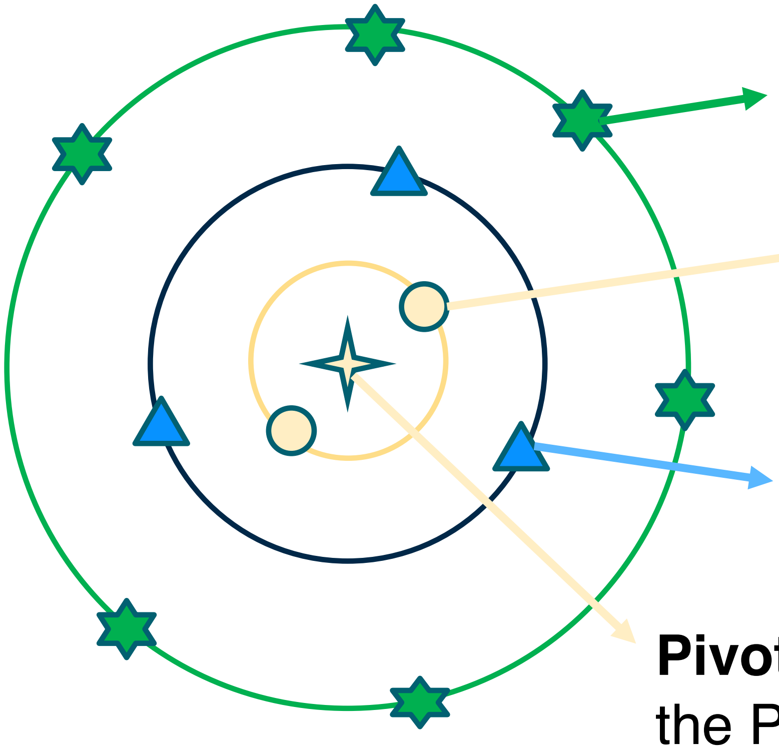
Raw text (x_{raw}): He will not accept it because he will not like it

Step3 Replace: **He will** accept it because he hates the plan he will not fancy it

Swap

Step4 Swap: **will He** accept it because he hates the plan he will not fancy it → Major

SEScore2: synthesize realistic errors with retrieved examples at various severity levels



Random: Rescaling the statement by the **raccoon** of the Council

Minor: **Rescales** the statement by the President of the Council

Major: Rescaling the statement by the President of the **security** Council

Pivot: Rescaling the statement by the President of the Council

Stratified Error Synthesis Overview (SEScore1)

Raw text (x_{raw}): He will not accept it because he will not like it

Step1 Insertion: He will not accept it because **he hates the plan** he will not like it ⟹ Major

Step2 Deletion: He will accept it because **he hates the plan** he will not like it ⟹ Major

Step3 Replace: He will accept it because **he hates the plan** he will not **fancy** it ⟹ Minor

Step4 Swap: **will He** accept it because **he hates the plan** he will not **fancy** it ⟹ Minor

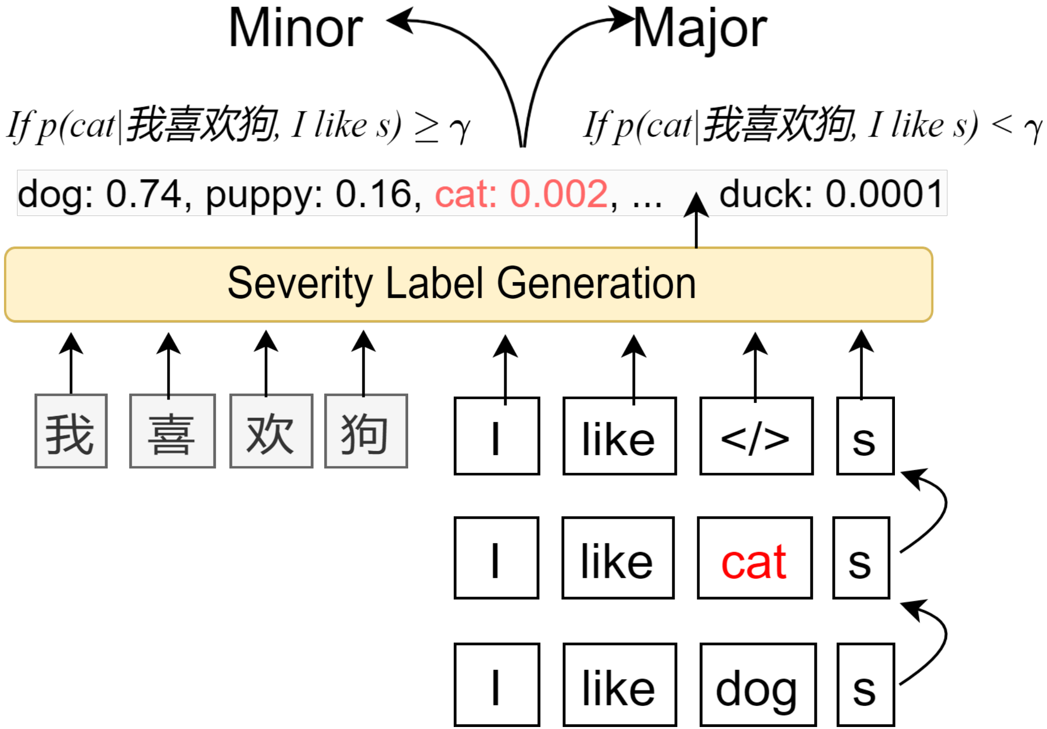
Stratified Error Synthesis Overview

Raw Text Reference: He will not accept it because he will not like it

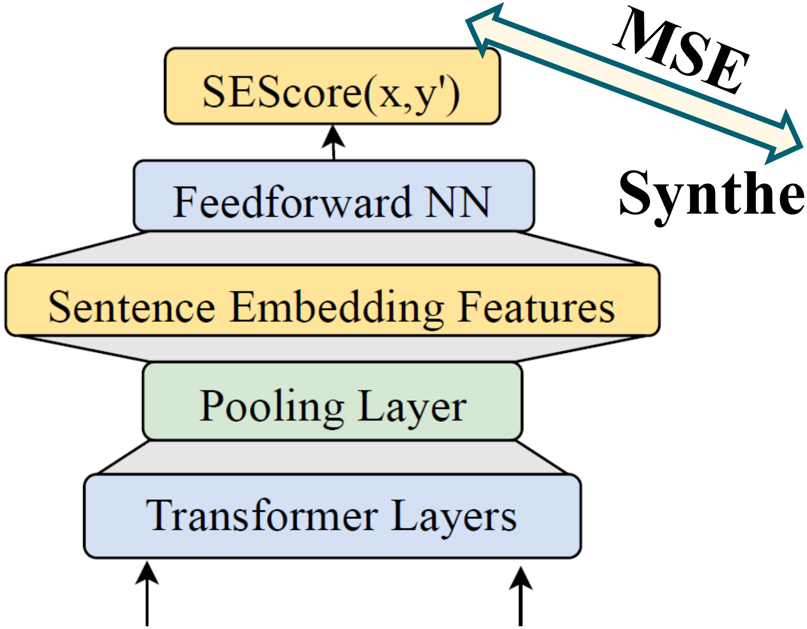
Synthesized Text: will He accept it because he hates the plan he will not fancy it

Synthetic Quality Score: -16

Severity Measure pipeline


Minor
Major
If $p(cat|$我喜欢狗$, I\ like\ s) \geq \gamma$
If $p(cat|$我喜欢狗$, I\ like\ s) < \gamma$
dog: 0.74, puppy: 0.16, cat: 0.002, ... duck: 0.0001
Severity Label Generation
我
喜
欢
狗
I
like
</>
s
I
like
cat
s
I
like
dog
s

Quality Prediction Model (Training)



SEScore(x,y')

MSE

Feedforward NN

Sentence Embedding Features

Pooling Layer

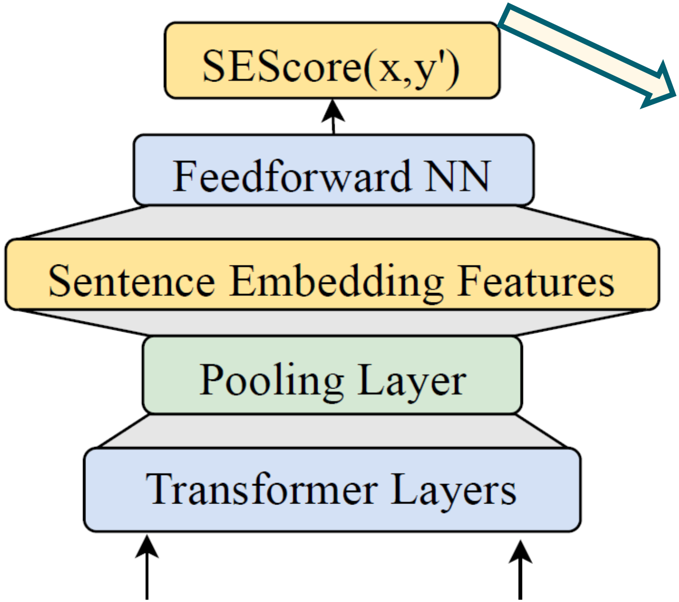
Transformer Layers

Synthetic Quality Score: -5

He is a dog person
(Raw Text)

He is a cat person
(Synthesized Text)

Quality Prediction Model (Inference)



Score: -0.5
(Very positive Score!)

He is a dog person
(Reference)

He likes dogs
(Candidate Text)

Experimental Setup

- Testing Datasets
  - WMT21(MQM) **Machine Translation** En-De, Zh-En, De-En
  - WebNLG20 **Data-to-Text**
  - IWSLT22 **Speech Translation** En-J
  - BAGEL **Dialogue Generation**
- Correlation to Humans
  - Segment-level Kendall Correlation
- Kendall Formulation

$$t_b = \frac{P-Q}{\sqrt{(P+Q+X_0)(P+Q+Y_0}}$$

Benchmark	# sys	# per sys
WMT21 En-De News	15	527
WMT21 Zh-En News	17	650
WMT21 De-En News	10	100
WMT21 En-De TED	15	529
WMT21 Zh-En TED	14	529
IWSLT22 En-Ja	4	118
BAGEL Dialogue	-	202
WebNLG Data-to-Text	17	177

Experimental Details

1) It takes 10 mins to generate 5M sentences (64 CPUs)

2) use RemBERT as backbone

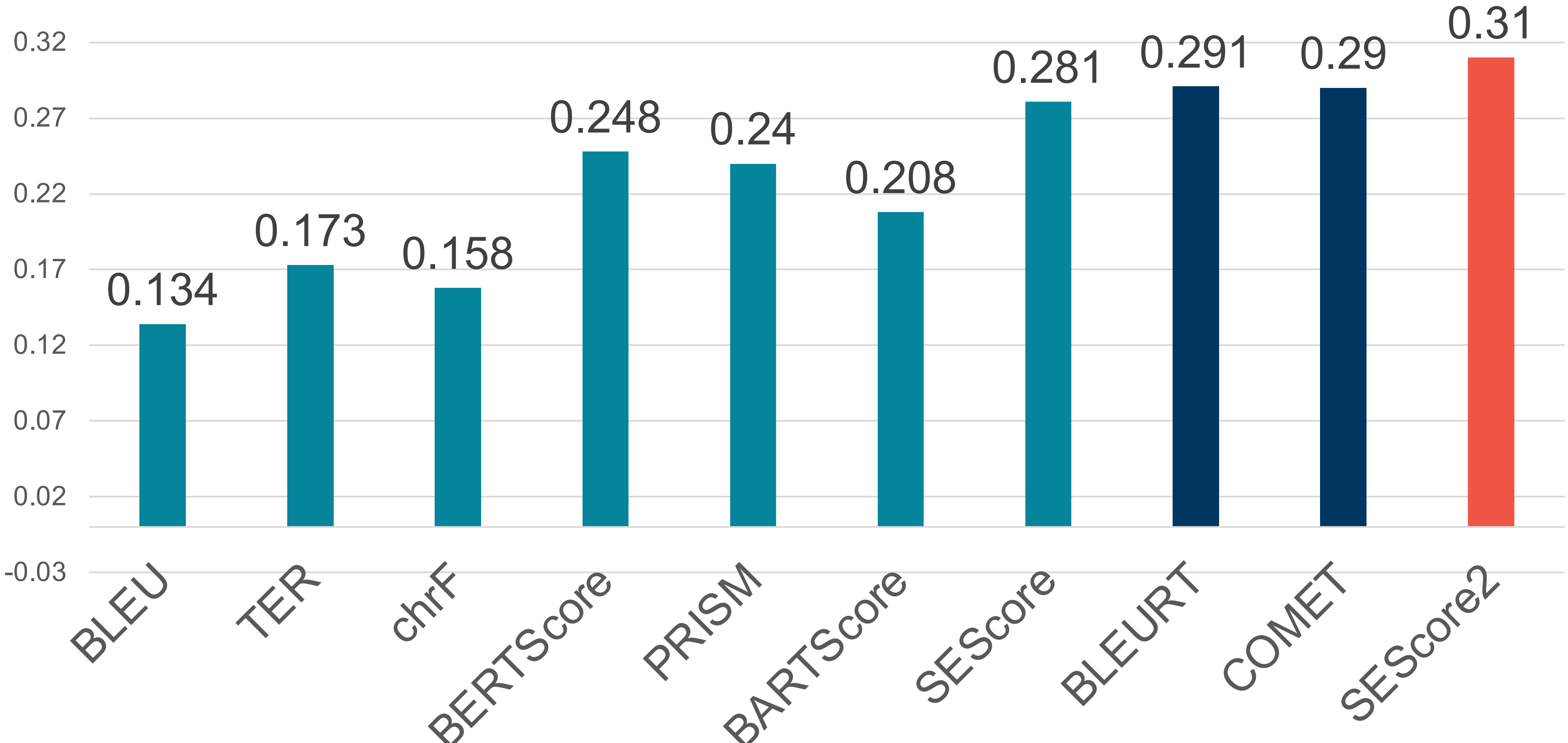
3) batch size: 256, learning rate: 3e-5 and dropout rate: 0.15

	Index Table		Pretraining Data	
Language	News	Wikipedia	# Anchor	# Retrieved
English	20M	20M	5M	13.5M
German	4.5M	16M	4.5M	13.2M
Japanese	18M	12M	5M	13.3M

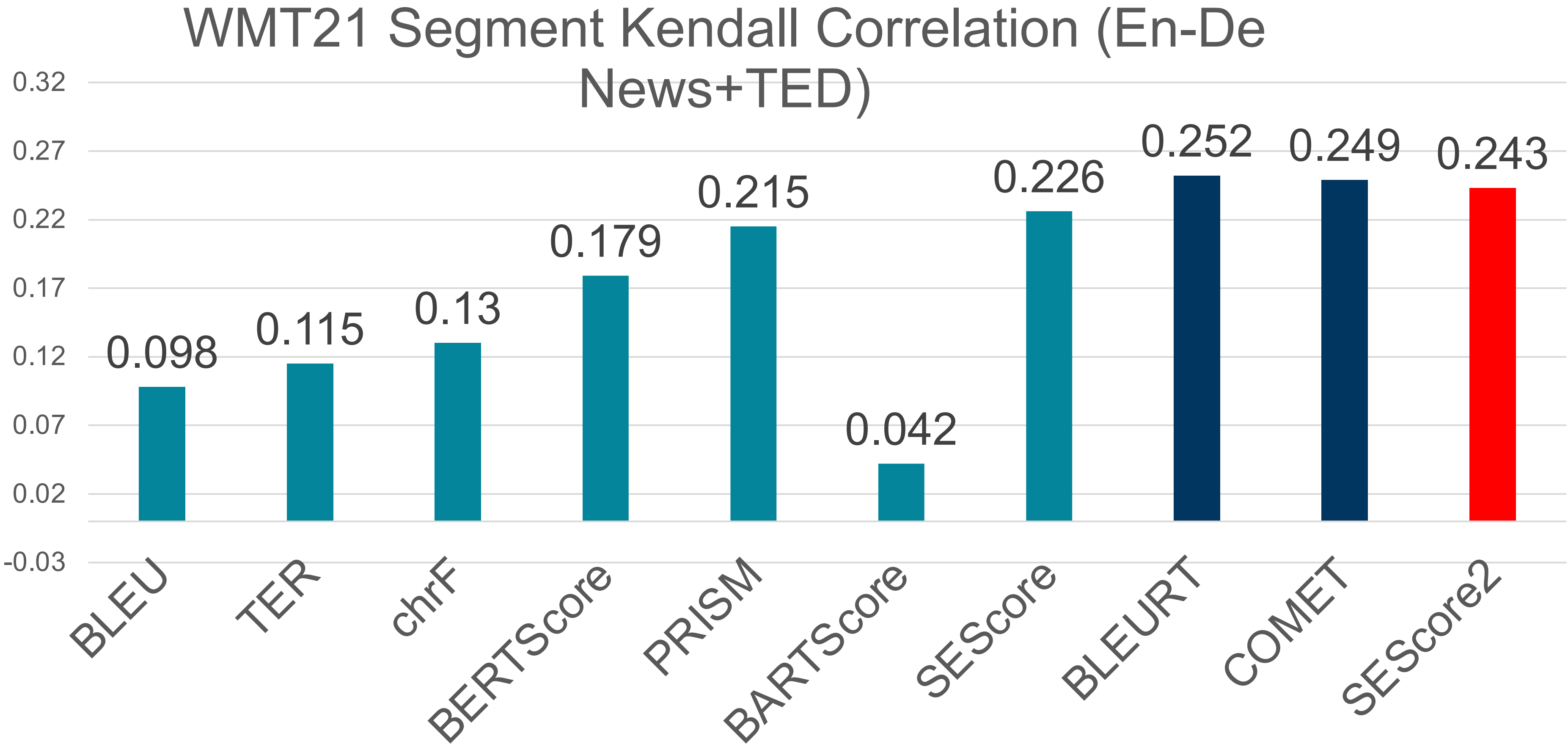
Do we need separate metrics for each text generation task?

SEScore1&2 can be used to evaluate Machine Translation

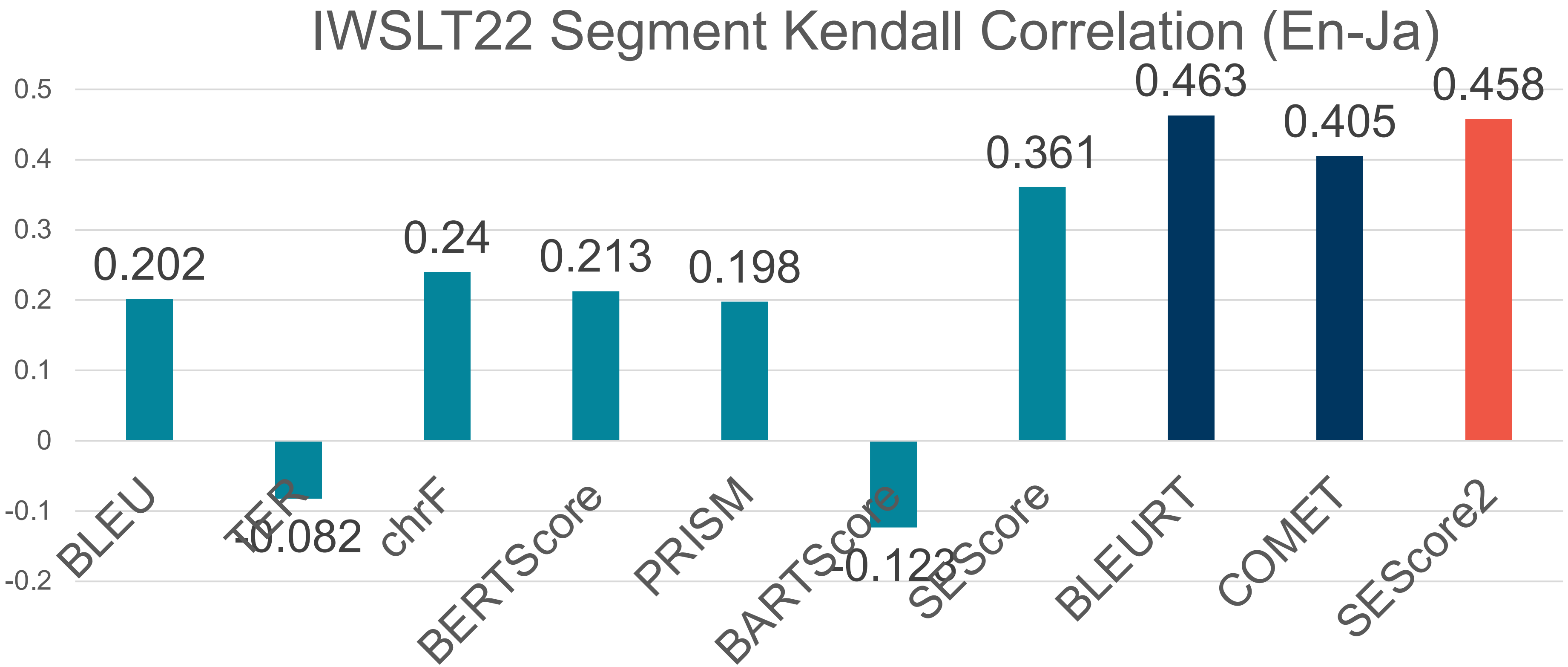
WMT21 Segment Kendall Correlation (Zh-En News+TED)



SEScore1&2 can be used to evaluate Machine Translation



SEScore1&2 can be used to evaluate Speech Translation

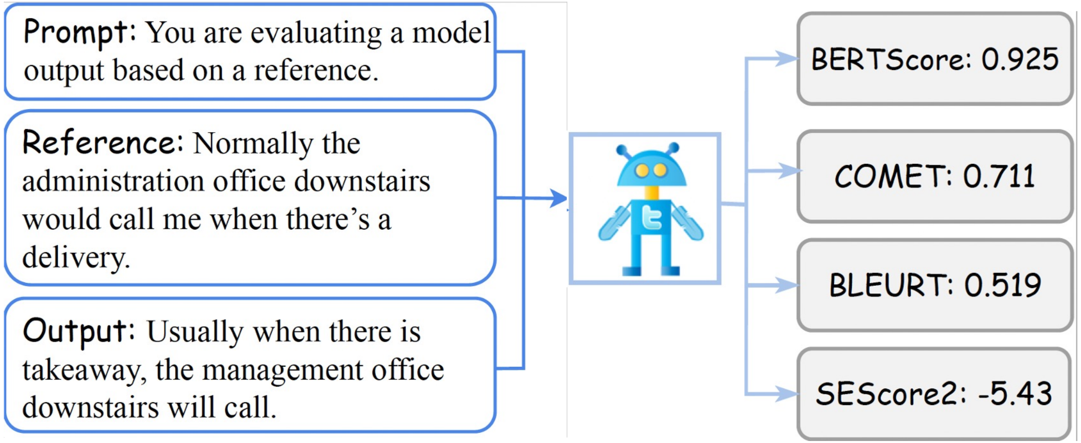


Do we need separate metrics for each domain of the same language and task?

Generating Explanation for Metric

InstructScore (SEScore3), EMNLP 2023

InstructScore (EMNLP 2023)



Prompt: You are evaluating a model output based on a reference.

Reference: Normally the administration office downstairs would call me when there's a delivery.

Output: Usually when there is takeaway, the management office downstairs will call.

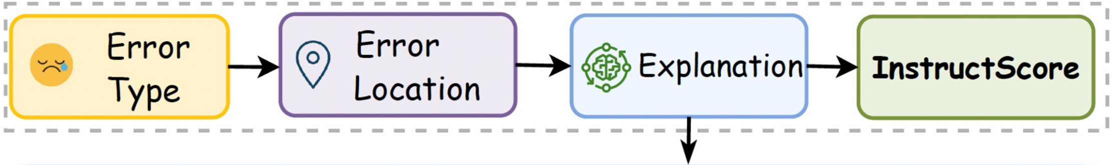
BERTScore: 0.925

COMET: 0.711

BLEURT: 0.519

SEScore2: -5.43

I don't understand those scores mean? Is 0.519 good?



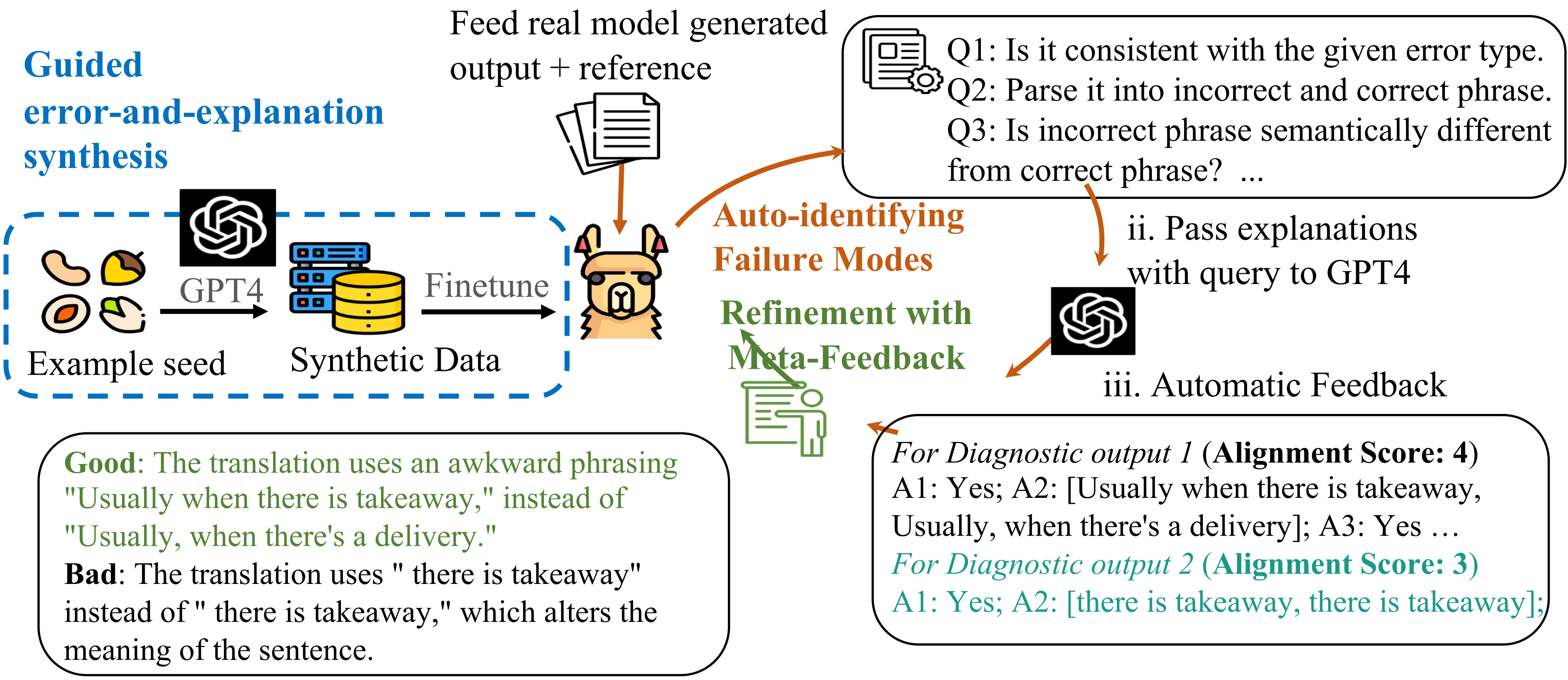
Error Type → Error Location → Explanation → **InstructScore**

Error Type: Incorrect translation has stylistic problems
Severity: Major
Error Location: Usually when there is takeaway,
Explanation: The translation uses an awkward phrasing "Usually when there is takeaway," instead of "Usually, when there's a delivery."
Score: -5

Challenges in Fine-grained Auto Evaluation of NLG

- **Fine-grained Explainability**: Can we build an automated metric that provides natural language explanations, in addition to numerical scores?
- **Compact yet Competitive**: Can we build a 7B model-based evaluator to beat metrics based on 175B LLMs?
- **No Human Annotations on Outputs for Training:** Ideally, we would not want to rely on human annotations of outputs for training, so that we can adapt to different domains and tasks.

InstructScore Pipeline



Guided error-and-explanation synthesis

Example seed

GPT4

Synthetic Data

Finetune

Feed real model generated output + reference

Q1: Is it consistent with the given error type.
Q2: Parse it into incorrect and correct phrase.
Q3: Is incorrect phrase semantically different from correct phrase? ...

Auto-identifying Failure Modes

ii. Pass explanations with query to GPT4

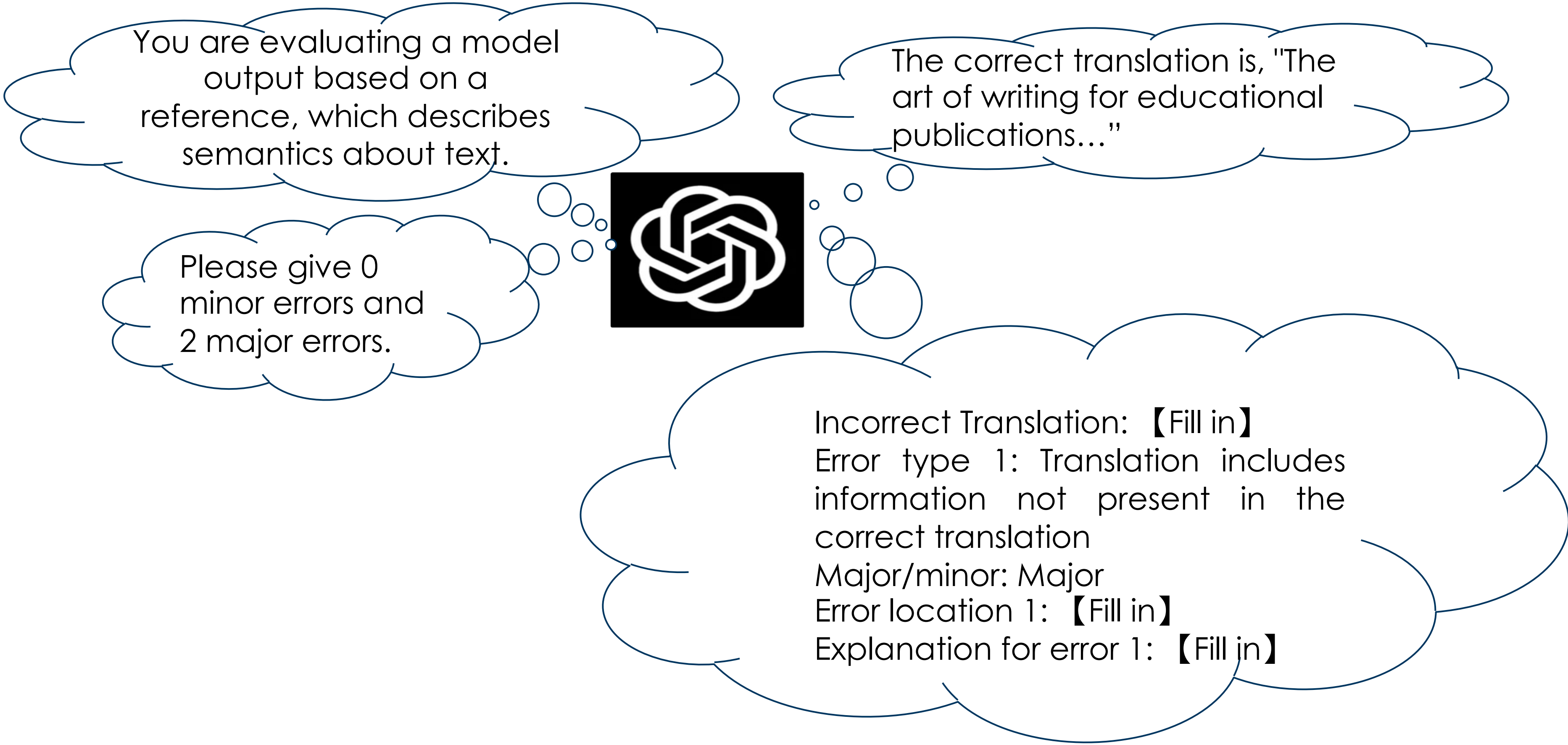
Refinement with Meta-Feedback

iii. Automatic Feedback

For Diagnostic output 1 (**Alignment Score: 4**)
A1: Yes; A2: [Usually when there is takeaway, Usually, when there's a delivery]; A3: Yes …
For Diagnostic output 2 (**Alignment Score: 3**)
A1: Yes; A2: [there is takeaway, there is takeaway];

Good: The translation uses an awkward phrasing "Usually when there is takeaway," instead of "Usually, when there's a delivery."
Bad: The translation uses " there is takeaway" instead of " there is takeaway," which alters the meaning of the sentence.

How do we generate Synthetic Dataset



You are evaluating a model output based on a reference, which describes semantics about text.

The correct translation is, "The art of writing for educational publications…"

Please give 0 minor errors and 2 major errors.

Incorrect Translation: 【Fill in】
Error type 1: Translation includes information not present in the correct translation
Major/minor: Major
Error location 1: 【Fill in】
Explanation for error 1: 【Fill in】

Why not directly use GPT-4 to evaluate?

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

Here is our synthetic data

Correct translation: "The art of writing for educational publications involves striking a delicate balance between providing enough detail to be useful and overwhelming the reader with too much information."

Incorrect Translation: "Waiting for educational publications is about finding a balance between giving enough examples to be useful and making the reader bored with too many details."

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

Major/minor: Major

Error location 1: "...giving enough examples..."

Explanation for error 1: The correct translation talks about providing enough detail, not specific examples, which changes the meaning of the sentence.

Direct Knowledge Distillation can tigger failure modes!

Reference: A series of creative academic achievements were awarded at the opening ceremony of the 2016 Annual Academic Conference of Zhejiang University.

Output: At the opening ceremony of the 2016 academic annual meeting of Zhejiang University, a series of academic achievements were recognized.

Error type 3: Wrong word choice
Major/minor: Minor
Error location 3: "academic annual meeting"
Explanation for error 3: The incorrect translation adds the word "annual" to the phrase "academic meeting," which is not present in the correct translation. However, this does not significantly change the meaning of the sentence.

M1: Error type is inconsistent with explanation

Direct Knowledge Distillation can tigger failure modes!

Reference: A series of creative academic achievements were awarded at the Annual Academic Conference of Zhejiang University.

Output: At the academic annual meeting of Zhejiang University, a series of academic achievements were recognized.

Error type 3: Incorrect translation includes information not present in the correct translation
Major/minor: Minor
Error location 3: "Zhejiang University"
Explanation for error 3: The incorrect translation adds the word "annual" to the phrase "academic meeting," which is not present in the correct translation. However, this does not significantly change the meaning of the sentence.

M2: Error locations are not consistent with the explanation

Direct Knowledge Distillation can tigger failure modes!

Reference: A series of creative academic achievements were awarded at the Annual Academic Conference of Zhejiang University.

Output: At the academic annual meeting of Zhejiang University, a series of academic achievements were recognized.

Error type 3: Incorrect translation includes information not present in the correct translation
Major/minor: Minor
Error location 3: "Annual Academic Conference"
Explanation for error 3: The incorrect translation adds the word "annual" to the phrase "academic meeting," which is not present in the correct translation. However, this does not significantly change the meaning of the sentence.

M3: Error locations are not referred in the output text

Direct Knowledge Distillation can tigger failure modes!

Reference: A series of creative academic achievements were awarded at the Annual Academic Conference of Zhejiang University.

Output: At the academic annual meeting of Zhejiang University, a series of academic achievements were recognized.

Error type 3: Incorrect translation includes information not present in the correct translation
Major/minor: Minor
Error location 3: "academic annual meeting"
Explanation for error 3: The incorrect translation contains "Annual Academic Conference", which is the incorrect translation.

M4: Error locations are not consistent with the explanation

Direct Knowledge Distillation can tigger failure modes!

Reference: A series of creative academic achievements were awarded at the Annual Academic Conference of Zhejiang University.

Output: At the academic annual meeting of Zhejiang University, a series of academic achievements were recognized.

Error type 3: Incorrect translation includes information not present in the correct translation
Major/minor: Major
Error location 3: "academic annual meeting"
Explanation for error 3: The incorrect translation adds the word "annual" to the phrase "academic meeting," which is not present in the correct translation. However, this does not significantly change the meaning of the sentence.

M5: Major and minor labels are not correct

Direct Knowledge Distillation can tigger failure modes!

Reference: A series of creative academic achievements were awarded at the Annual Academic Conference of Zhejiang University.

Output: At the academic annual meeting of Zhejiang University, a series of academic achievements were recognized.

Error type 3: Incorrect translation includes information not present in the correct translation
Major/minor: Minor
Error location 3: "academic annual meeting"
Explanation for error 3: The incorrect translation adds the word "annual" to the phrase "academic meeting," which is not present in the correct translation. However, this does not significantly change the meaning of the sentence.

M6: Explanation is illogical

Direct Knowledge Distillation can tigger failure modes!

Reference: A series of creative academic achievements were awarded at the Annual Academic Conference of Zhejiang University.

Output: At the academic annual meeting of Zhejiang University, a series of academic achievements were recognized.

Error type 3: Word Choice
Major/minor: Minor
Error location 3: "Zhejiang University"
Explanation for error 3: The use of Zhejiang university is wrong. This is not the right entity

G1: Error described in the explanation is not an error

Use GPT-4 as a reward model

Reference: *...... the revolutionary base area of south Jiangxi.*

Output: *......the old revolutionary district of southern Jiangxi.*

Error location 1: "old revolutionary district"

Error type 1: Terminology is non-standard or does not fit the context. Terminology is used inconsistently

Explanation 1: The correct term should be “revolutionary base area" which refers to the area where the revolution started in 1927, not "old revolutionary district" which does not convey the same meaning.

......

revolutionary base area

old revolutionary district

Use GPT-4 as a reward model

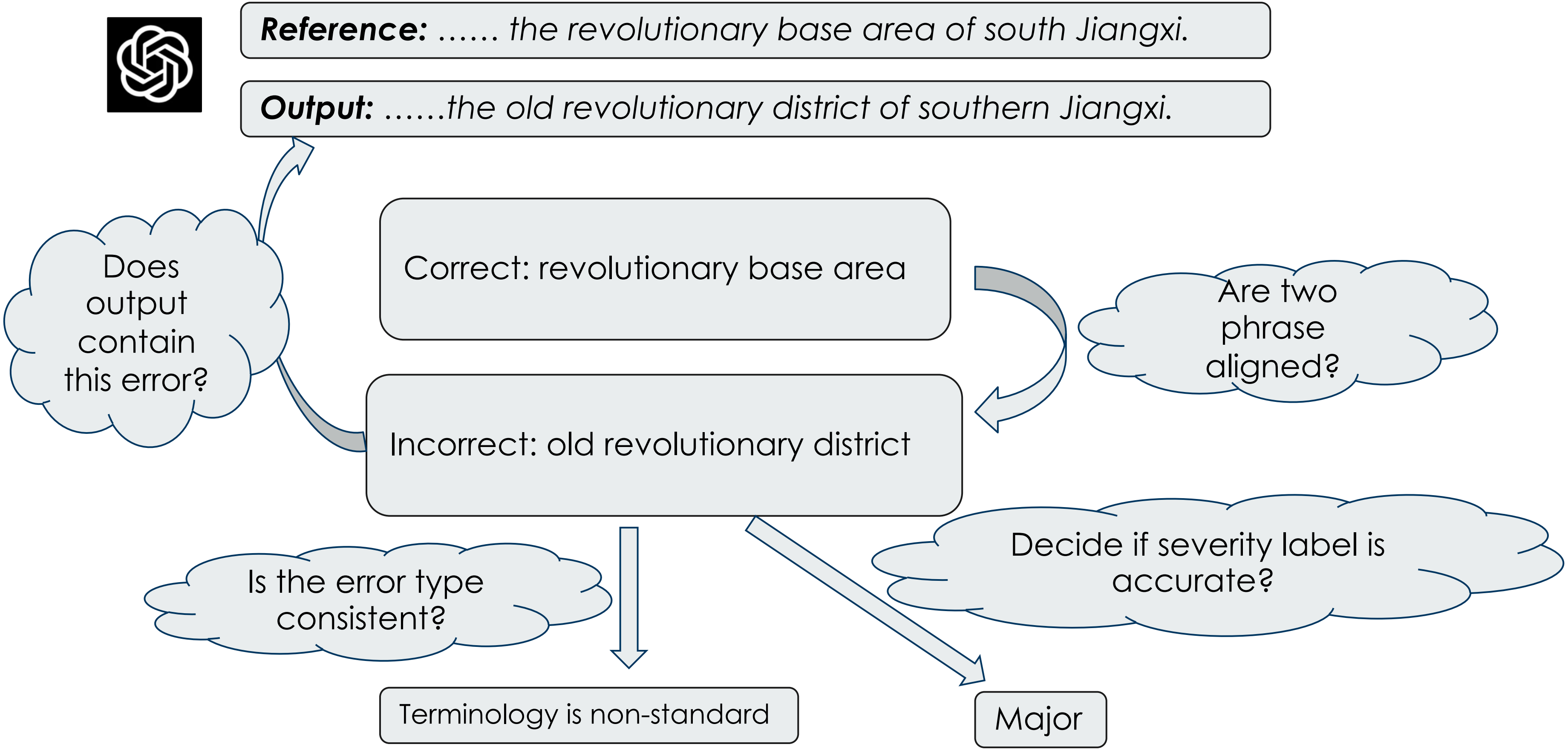
Reference: *…… the revolutionary base area of south Jiangxi.*
Output: *……the old revolutionary district of southern Jiangxi.*

Error location 1: "old revolutionary district"
Error type 1: Terminology is non-standard or does not fit the context. Terminology is used inconsistently
Explanation 1: The correct term should be "revolutionary base area" which refers to the area where the revolution started in 1927, not "old revolutionary district" which does not convey the same meaning.
……

Correct: revolutionary base area

Incorrect: old revolutionary district

Use GPT-4 as a reward model



Reference: *...... the revolutionary base area of south Jiangxi.*

Output: *......the old revolutionary district of southern Jiangxi.*

Does output contain this error?

Correct: revolutionary base area

Are two phrase aligned?

Incorrect: old revolutionary district

Is the error type consistent?

Decide if severity label is accurate?

Terminology is non-standard

Major

GPT-4's Feedback

Reference: *...... the revolutionary base area of south Jiangxi.*
Output: *......the old revolutionary district of southern Jiangxi.*

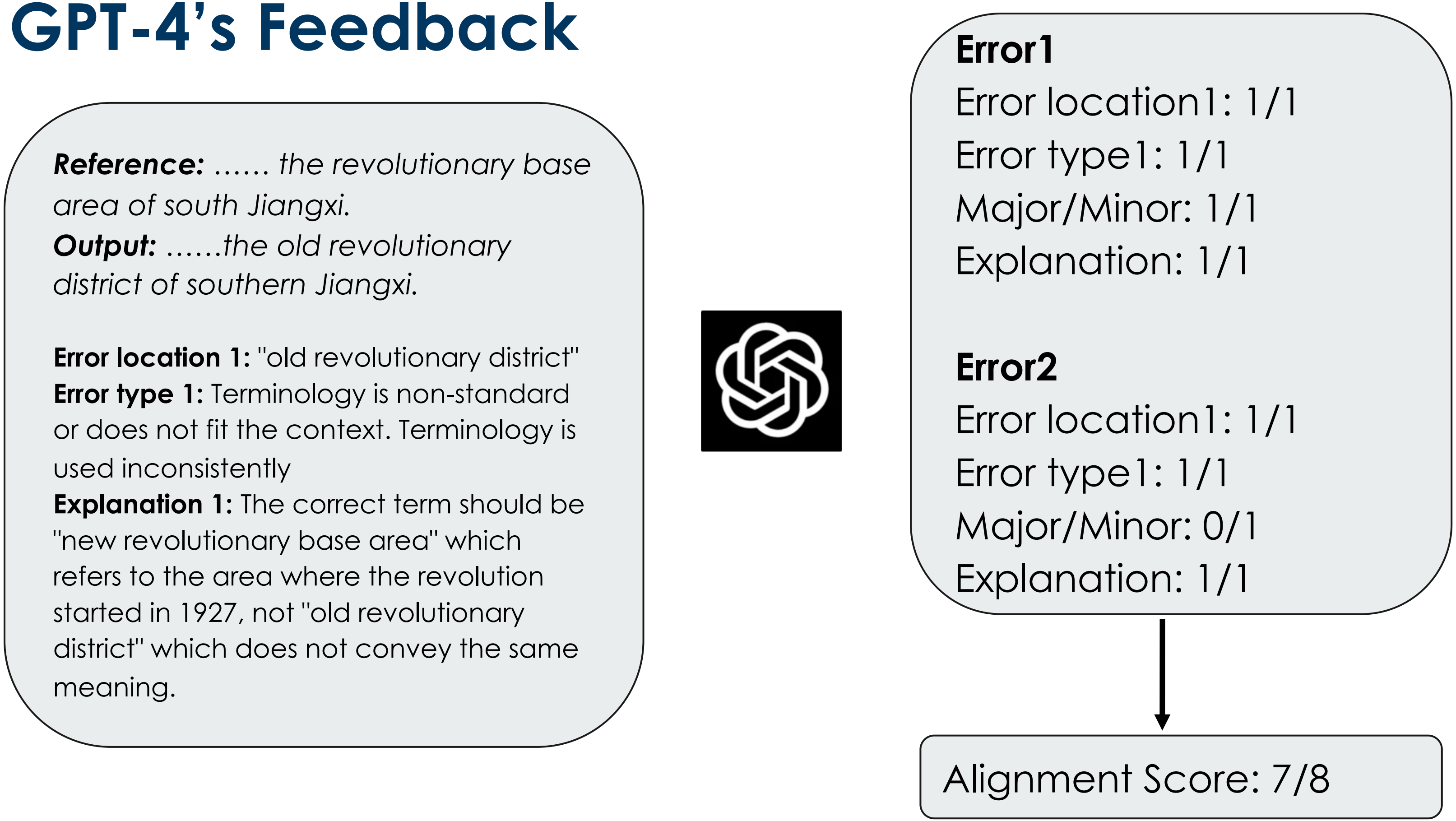
Error location 1: "old revolutionary district"
Error type 1: Terminology is non-standard or does not fit the context. Terminology is used inconsistently
Explanation 1: The correct term should be "new revolutionary base area" which refers to the area where the revolution started in 1927, not "old revolutionary district" which does not convey the same meaning.

Error 1:
A1: "old revolutionary district"
A2: ["old revolutionary district", "revolutionary base area"]
A3: "Yes"
A4: "major-error"
A5: "Yes"
A6: "Yes",

Error 2:
A1: "dominant"
A2: ["dominant","privileged"]
A3: "Yes"
A4: "minor-error"
A5: "Yes"
A6: "Yes"

A7: "No, 0"

GPT-4’s Feedback



Reference: *...... the revolutionary base area of south Jiangxi.*
Output: *......the old revolutionary district of southern Jiangxi.*

Error location 1: "old revolutionary district"
Error type 1: Terminology is non-standard or does not fit the context. Terminology is used inconsistently
Explanation 1: The correct term should be "new revolutionary base area" which refers to the area where the revolution started in 1927, not "old revolutionary district" which does not convey the same meaning.

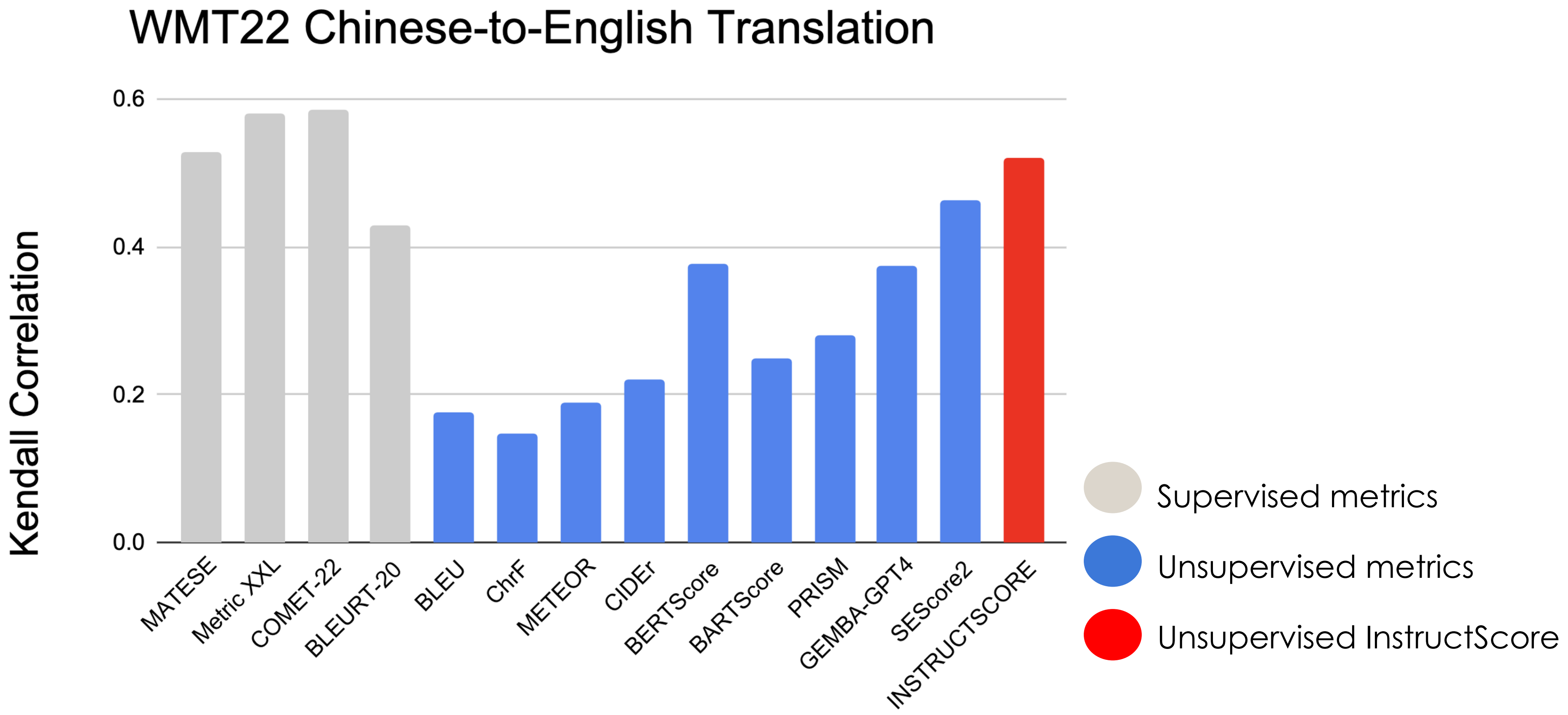
Error1
Error location1: 1/1
Error type1: 1/1
Major/Minor: 1/1
Explanation: 1/1

Error2
Error location1: 1/1
Error type1: 1/1
Major/Minor: 0/1
Explanation: 1/1

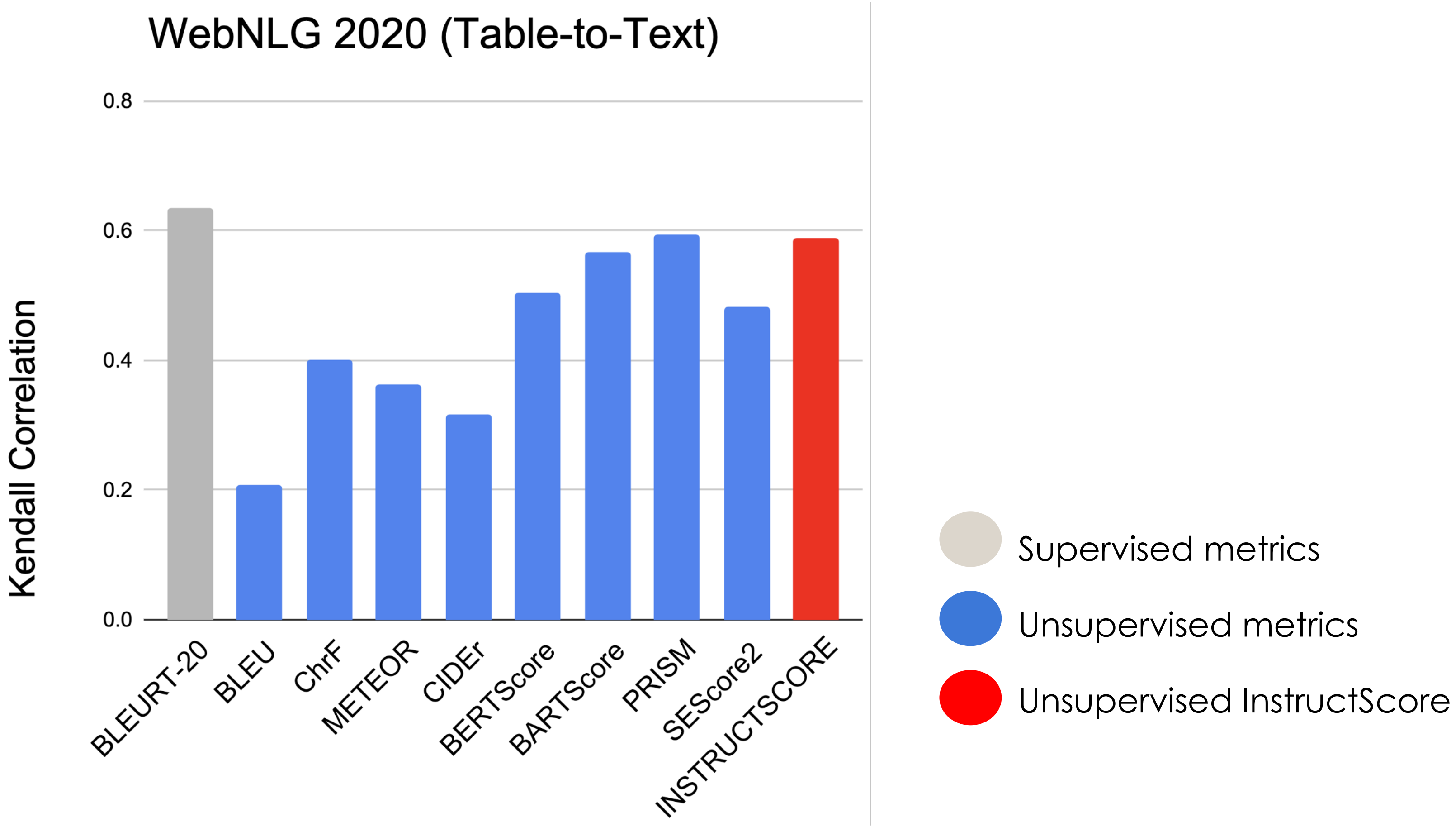
Alignment Score: 7/8

Robust Performance across Tasks (Five NLG tasks)

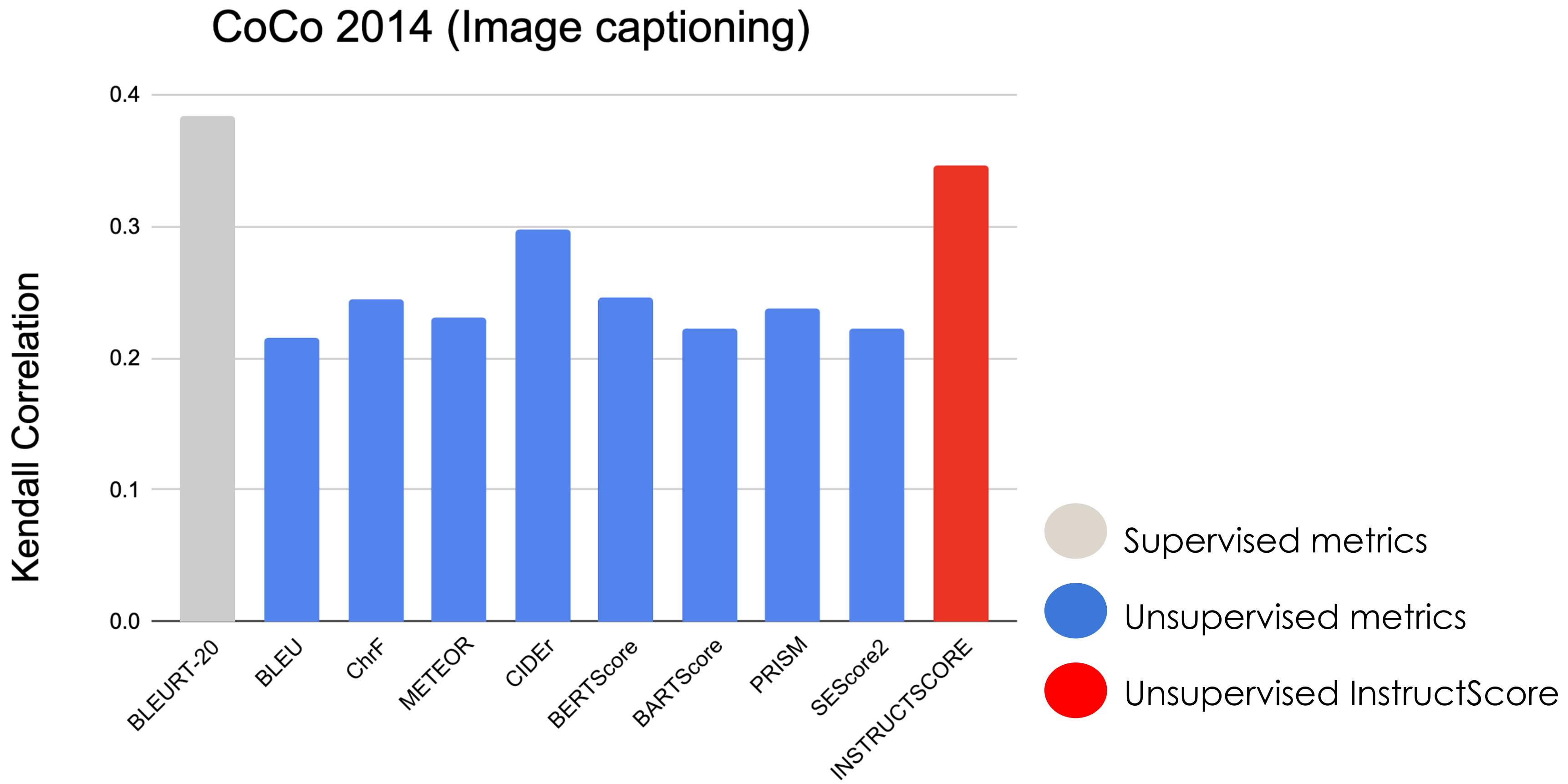
InstructScore can judge machine translation!



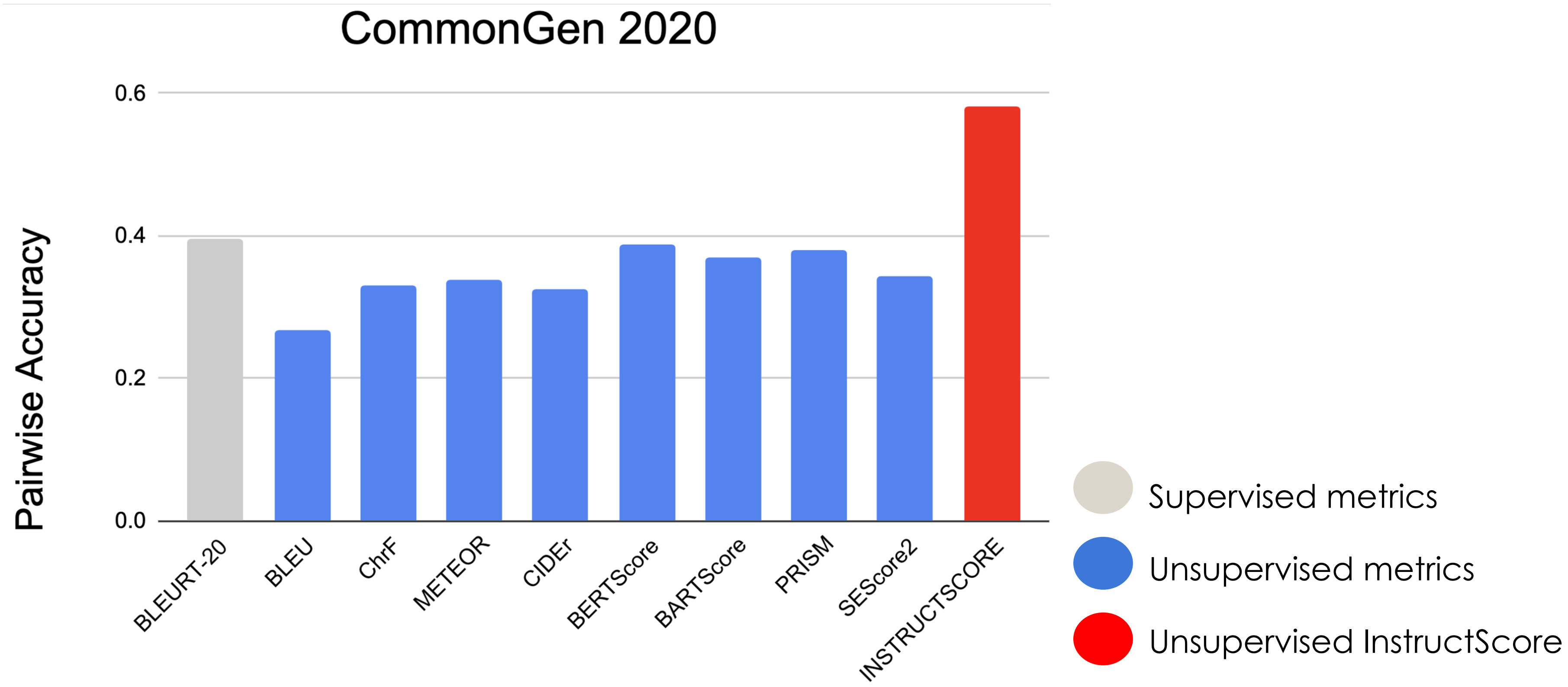
InstructScore can judge structure data-to-text!



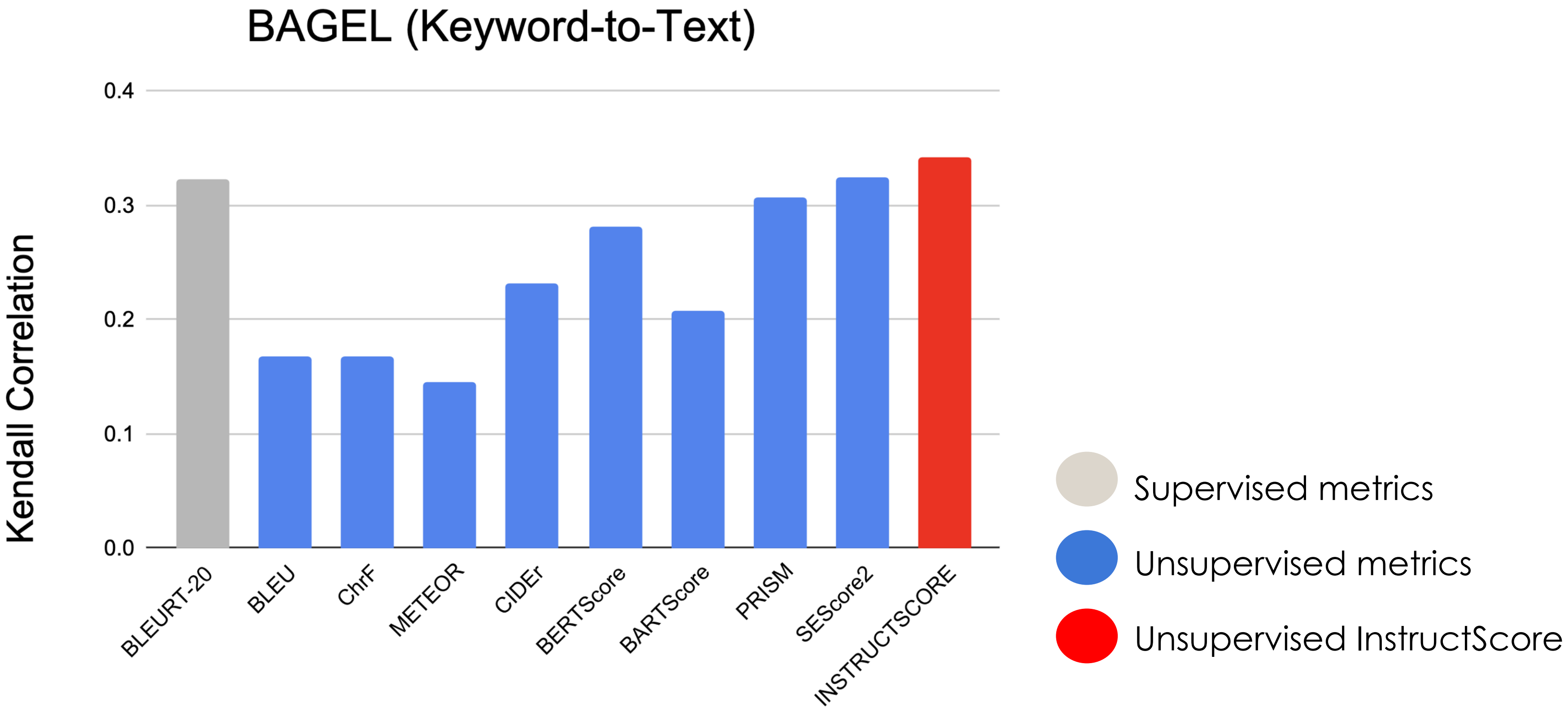
InstructScore can judge image captioning!



InstructScore can judge commonsense generation!

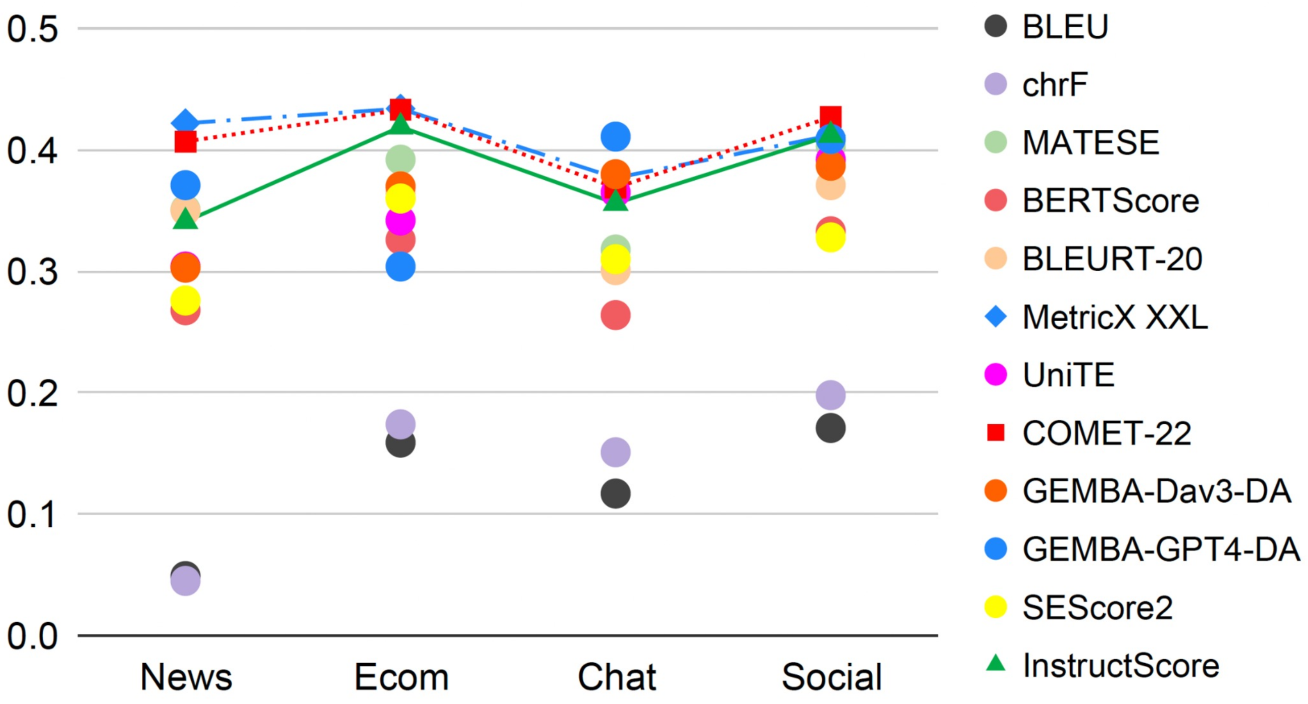


InstructScore can judge unseen keyword-to-text generation!

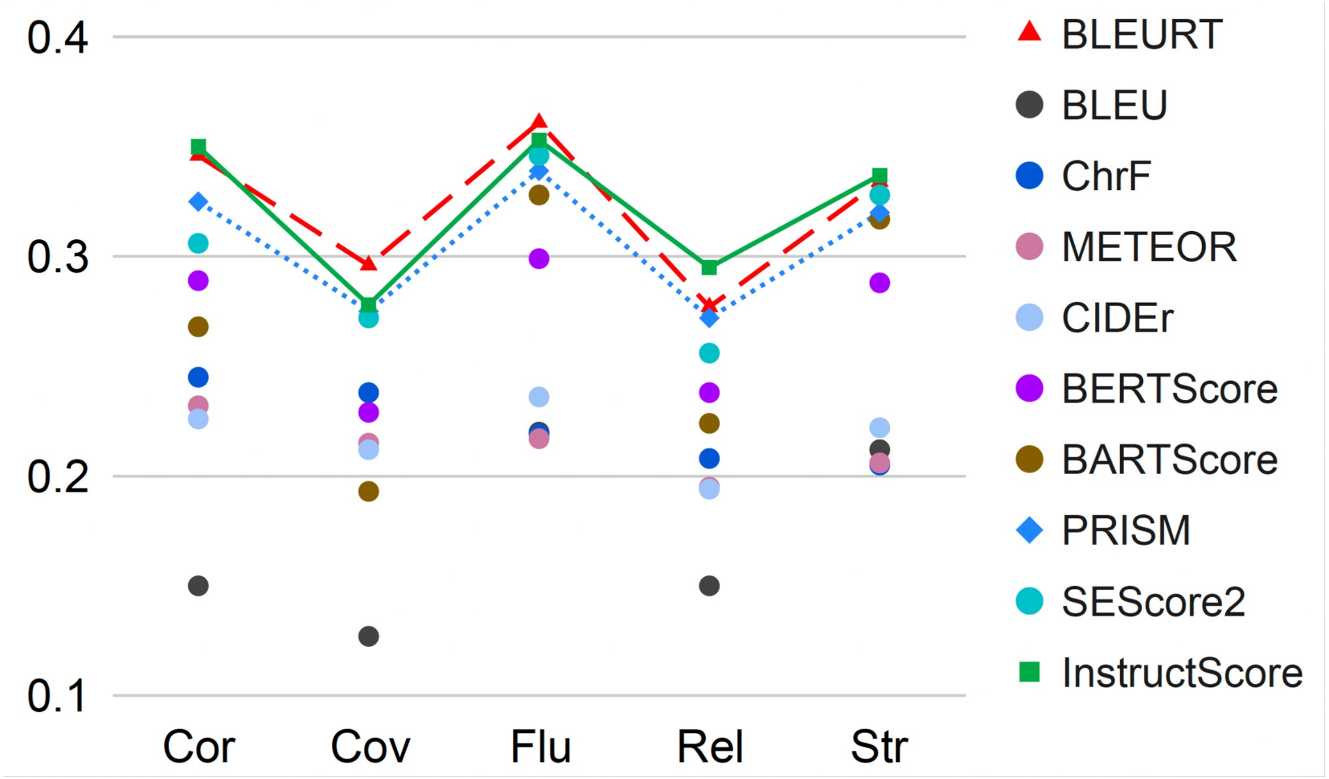


BAGEL (Keyword-to-Text)

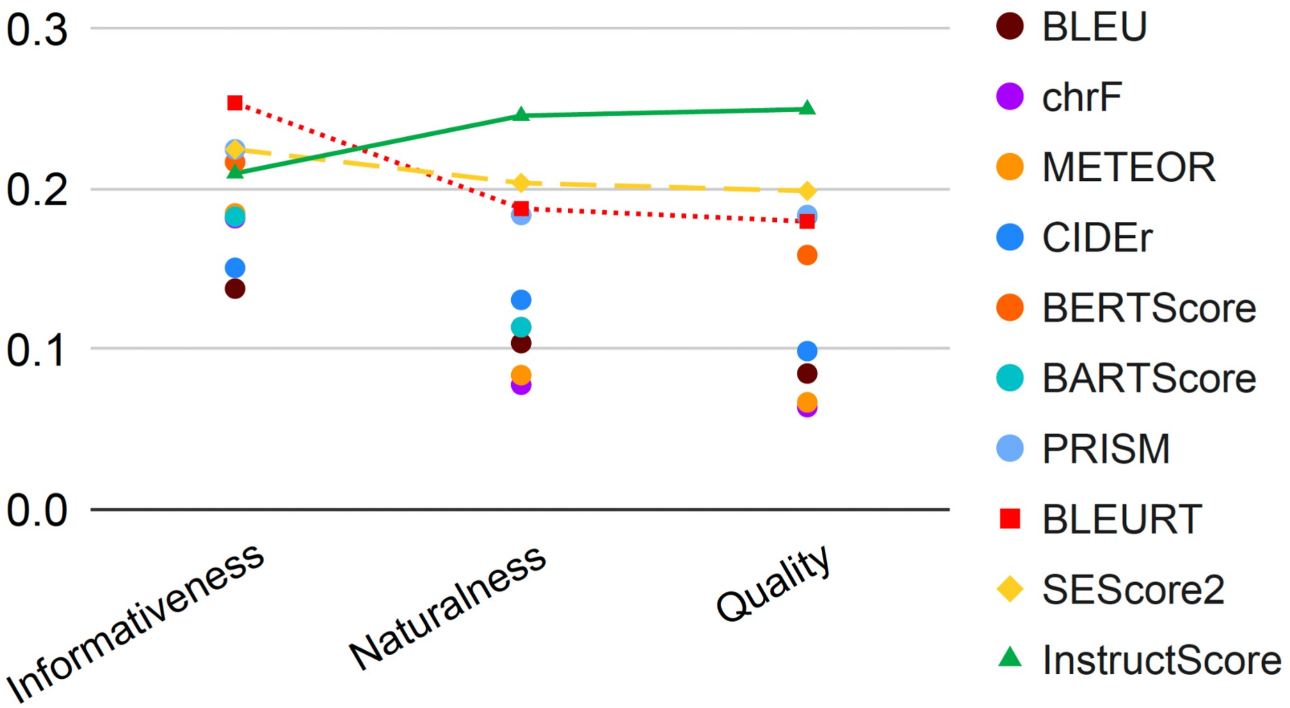
Robust Performance across Domains (WMT22 Zh-En)



Robust Performance across Dimensions (WebNLG20)



Generalization under Unseen Task (BAGEL D2T)



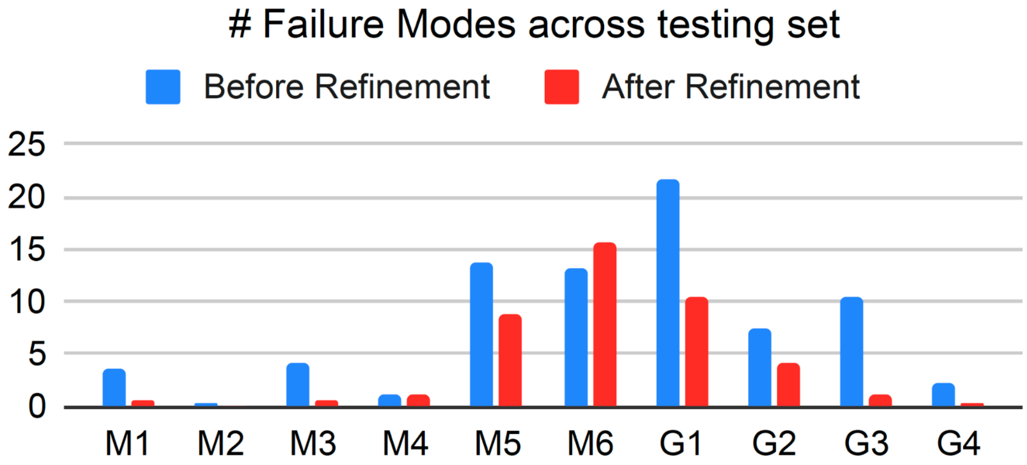
Automatic critique and Self-training can improve human alignments

Precision = # correct fields / total # annotated fields by InstructScore

Recall = # correct fields / total # annotated fields by human

We reduce G1 from **22** to **11** across 100 examples

InstructScore	**Precision**	**Recall**	**Human**
Finetune	0.778	0.824	0.773
Finetune+Refinement	**0.894**	**0.856**	**0.879**



Our Refinement can improve InstructScore Feedback

Reference: Washing hands with air, diagnosing disease through breath, and accurately seeking for source of smog; last night…

Output: Wash hands with air, judge diseases, accurately search for the source of fog Yesterday evening…

Error type 3: Incorrect translation includes information not present in the correct translation

Major/minor: Minor

Error location 3: "academic annual meeting"

Explanation for error 3: The incorrect translation adds the word "annual" to the phrase "academic meeting," which is not present in the correct translation.

Error type 3: Incorrect translation does not accurately represent the correct translation

Major/minor: Minor

Error location 3: "academic annual meeting"

Explanation for error 3: The incorrect translation uses "academic annual meeting" instead of "academic annual conference,"

Key Advantages of InstructScore

1. Compact yet competitive
2. Explainability
3. Easy to use
4. 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**!

Future Direction

1. Use fine-grained feedback to guide text generation
2. Better incorporate human rating data and synthetic data
3. Extend InstructScore to source-based setting and multilingual setting

SEScore/InstructScore

Arxiv: https://arxiv.org/abs/2305.14282

Github: https://github.com/xu1998hz/SEScore3

HuggingFace: https://huggingface.co/xu1998hz/InstructScore