Toward Explainable, Robust, and Actionable Translation Quality Estimation

Shaomu Tan | 
University of Amsterdam
Language Technology Lab

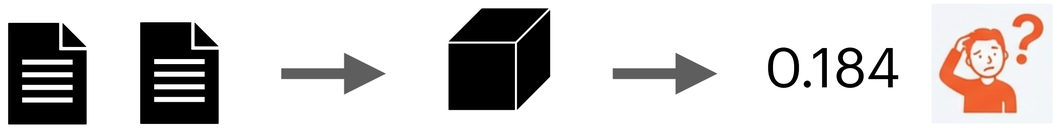
MME workshop


EACL 2026
RABAT • MOROCCO
Mars • March 24-29, 2026 • مارس

Automatic MT Evaluation

Metric	avg rank
METRICX XXL	1.20
COMET-22	1.32
UNITE	1.86
BLEURT-20	1.91
COMET-20	2.36
MATESE	2.57
COMETKIWI*	2.70
MS-COMET-22	2.84
UNITE-SRC*	3.03
YISI-1	3.27
COMET-QE*	3.33
MATESE-QE*	3.85
MEE4	3.87
BERTSCORE	3.88
MS-COMET-QE-22*	4.06
CHRF	4.70
F101SPBLEU	4.97
HWTSC-TEACHER-SIM*	5.17
BLEU	5.31
REUSE*	6.69

BLEU is **poorly correlated** with **Human Preferences**.



source + translation

Black-box metric

Single score, no explanations

Neural **Quality Estimation** (QE) metrics improved correlation, but at the cost of becoming a **black-box** signal.

Freitag, Markus, et al. "Results of WMT22 metrics shared task: Stop using BLEU–neural metrics are better and more robust."

Agenda

In this talk, we discuss:

- When QE becomes a bad optimization signal
- Why one score is not enough: we need multi-dimensional assessment
- Towards Explainable, Robust, and Actionable Quality Estimation

QE is becoming a control signal

not just an evaluator

Quality Estimation is widely used in WMT25 for:

- Data filtering
- Decoding/Test-Time Scaling/MBR
- RL training

WMT25 Human evaluation suggests that MT models may be **biased** when optimizing for these QE Metrics.

If the QE signal is brittle, it can misguide MT in training and inference.

Kocmi, Tom, et al. "Findings of the WMT25 General Machine Translation Shared Task: Time to Stop Evaluating on Easy Test Sets" (2025)

QE Metric Blind Spots

English-Icelandic

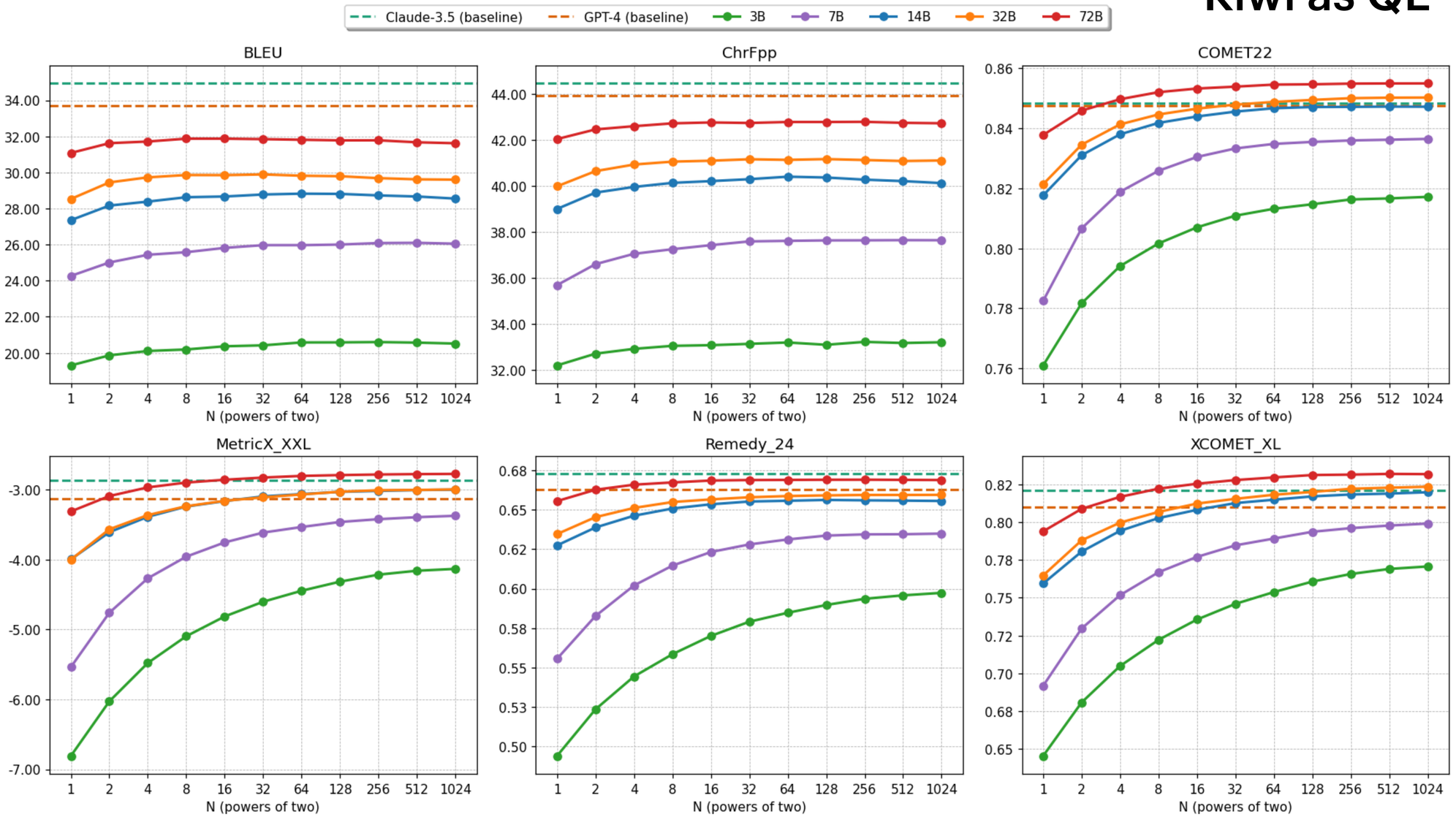
Source	Oversight is not only for the benefit of incarcerated people. Our task force recognized that an oversight office can partner with overburdened jail systems to provide much-needed assistance and recommendations, including smart policies that promote staff safety and reduce recidivism.
Reference	Eftirlit mun ekki einungis bæta líf fólks sem situr í fangageymslum. Vinnuhópurinn okkar komst að þeirri niðurstöðu að slíkur eftirlitsaðili gæti unnið með fangageymslukerfum sem komin eru að þolmörkum og veitt þarfa aðstoð og ráðgjöf, þar á meðal með því að hjálpa til við að móta upplýstar stefnur sem stuðla að öryggi starfsfólks og draga úr endurkomu fanga í fangelsi.

Kiwi=0.3110	Hingadir eru ekki minni framan sviðiins svífuglar. Þingreiðilistanu fer út frá þat at hingadeinsborg hvernig er möguleg að myndast með upphaflegu drápað og tvörfuða joðavellir, til að gera tilgvenda spurningar, for því að velja flýgrask heiltengingar og abladlega einingu.
Kiwi=0.4144	Önuspurnur fer ekki fyrir rétt um óskalum. Þarf lýstakistarfgröndunni til að slíta við millðkréfur狱院系统以提供所需的支持和建议，包括促进工作人员安全和降低再犯率的明智政策。Það var óferðum umframlegt að
Kiwi=0.8037	Af监督既不是囚犯的专属利益。我们的任务小组认识到，监督办公室能够与负担过重的监狱系统合作，提供所需的援助和建议，包括促进员工安全和减少再犯率的智能政策。

Tan, Shaomu, et al. "Investigating Test-Time Scaling with Reranking for Machine Translation." (2025)

TTS works when QE is reliable

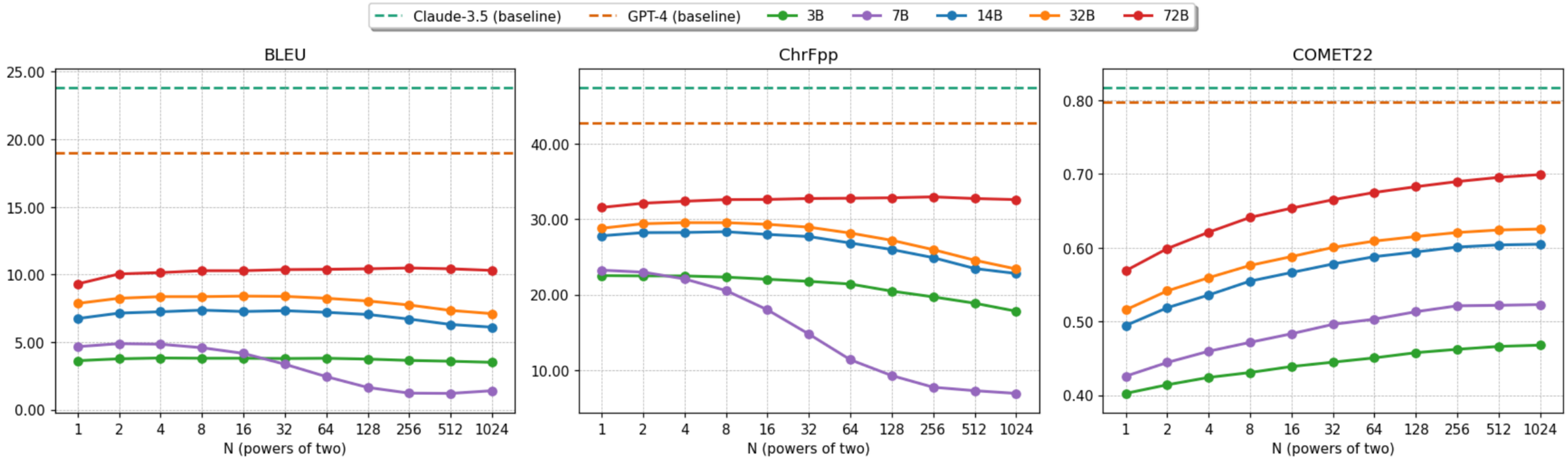
Kiwi as QE



Tan, Shaomu, et al. "Investigating Test-Time Scaling with Reranking for Machine Translation." (2025)

TTS collapses in low-resource settings

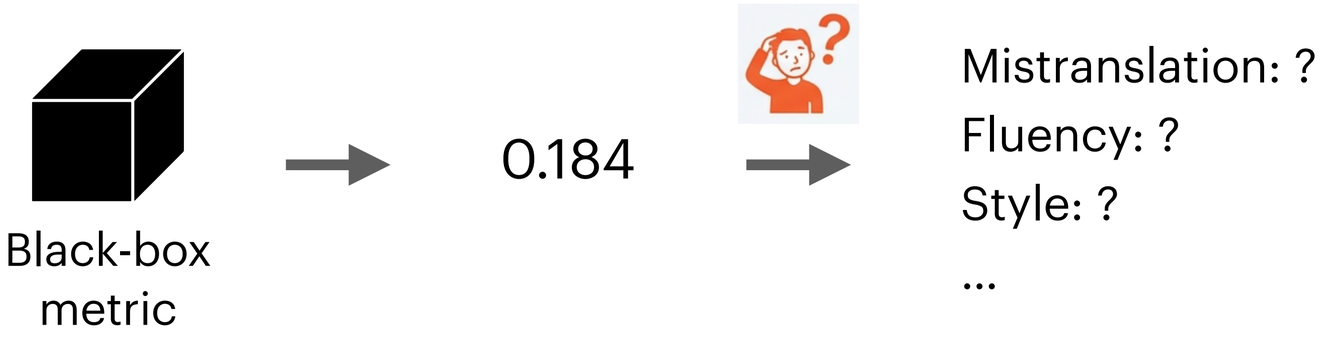
WMT24 English-Icelandic



TTS collapse when using unreliable QE

Tan, Shaomu, et al. "Investigating Test-Time Scaling with Reranking for Machine Translation." (2025)

Toward Multi-dimensional Metrics



Even when scalar QE is not wrong, one score is still not enough

- Scalar score tells us: “how good a translation is”
- But it does not tell us:
  - How good is it in different quality dimensions like accuracy, fluency, ...

Toward Multi-dimensional Metrics

What Does LLM Refinement Actually Improve?
A Systematic Study on Document-Level Literary Translation

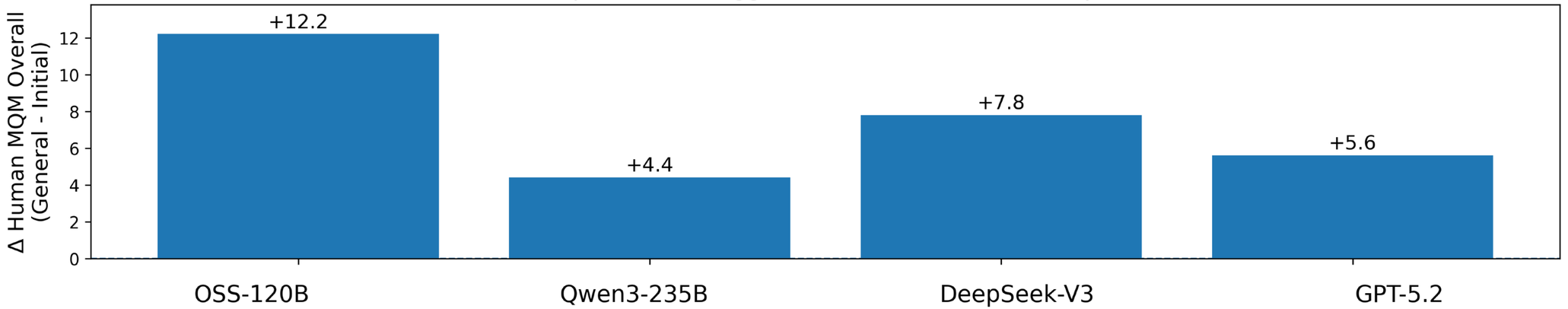
Shaomu Tan[1,2,*] **Dawei Zhu**[2] **Ke Tran**[2] **Michael Denkowski**[2] **Sony Trenous**[2]
Bill Byrne[2] **Leonardo Ribeiro**[2] **Felix Hieber**[2]
[1]University of Amsterdam [2]Amazon AGI

We conducted a comprehensive study investigating doc-level MT WMT24-Literary using seven LLMs:

- 1 million-word human annotation (MQM+Preference DA);
- Seven high-resource languages;
- Several LLM refinement approaches;

Toward Multi-dimensional Metrics

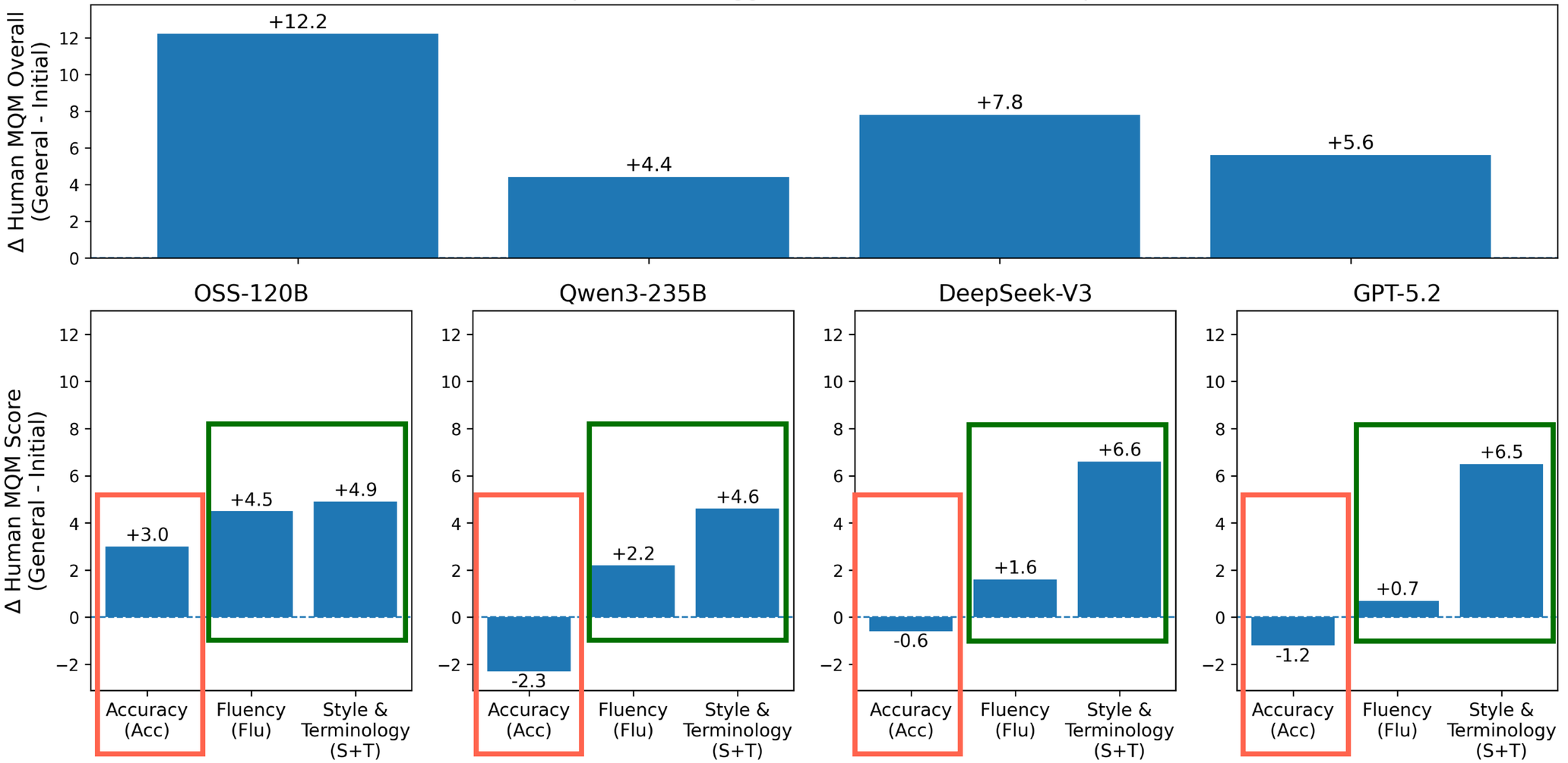
A case study on LLM refinement



LLM Refinement improves translation quality.

Toward Multi-dimensional Metrics

A case study on LLM refinement

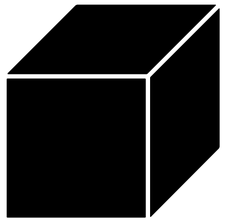


Overall gains hide the fact that refinement mostly improves fluency, not accuracy.

Towards Explainable, Robust, and Actionable Quality Estimation

Existing Metrics

source + translation



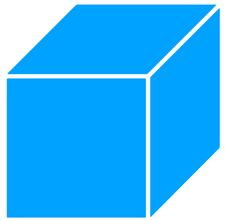
Black-box
Scalar metric

0.184 

Single score, no explanations

Our Goal

source + translation



Interpretable
Generative metric

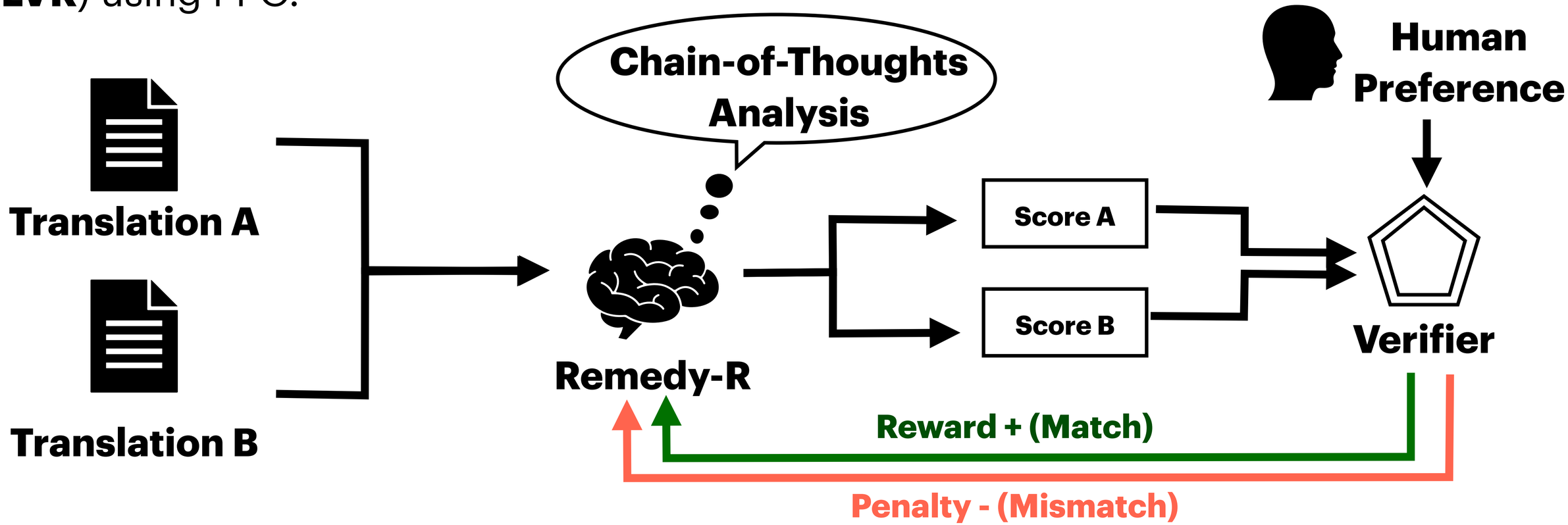
Multi-dimension Analysis

1. Semantic Adequacy: ...
2. Translation Fluency: ...
3. Content Completeness: ...

Final Quality Score: 18.4

Remedy-R: reason first, then score

Remedy-R: a training framework that learns a *generative* MT metric that **reasons about translation quality,** trained with Reinforcement Learning with Verifiable Reward (**RLVR**) using PPO.



Tan, Shaomu, et al. "Remedy-R: Generative Reasoning for Machine Translation Evaluation without Error Annotations." (2025)

Remedy-R: Experimental Setup

- **Training Data**:
  - only 60K pairwise data (WMT20);
  - Only 2 language pairs: en-de, zh-en;
  - no fine-grained error annotations
- **Foundation Models:**
  - Qwen 2.5 series: 7B, 14B, 32B;

Remedy-R Example

Source: Even if I’m not in the income level that would pay more taxes, I can still buy fewer things that are more expensive so that they can be made by people making a living wage and I can be willing to wait a few days to get it so that some gig worker doesn’t have to be worked to the bone...

Translation: Sogar wenn ich nicht in der Einkommensklasse bin, in der mehr Steuern gezahlt werden, kann ich immer noch weniger teure Dinge kaufen, sodass sie von Menschen hergestellt werden können, die maßgeblich dazu beitragen, und ich bin bereit, ein paar Tage zu warten, bevor der Arbeitslohn abläuft. . .

[Model’s Generation]

Let's evaluate this translation step by step:

1. Accuracy:
- "sodass sie von Menschen hergestellt werden können, die maßgeblich dazu beitragen" is not clear and does not accurately reflect the original meaning. The original text states that the products should be made by people who are paid a living wage, but the translation does not convey this clearly.

- "bevor der Arbeitslohn abläuft" is incorrect and does not make sense in this context. The original text states "so that some gig worker doesn’t have to be worked to the bone" which means the gig worker is not overworked. The translation should reflect this meaning.

2. Fluency:
- The sentence structure is somewhat complex and could be clearer.

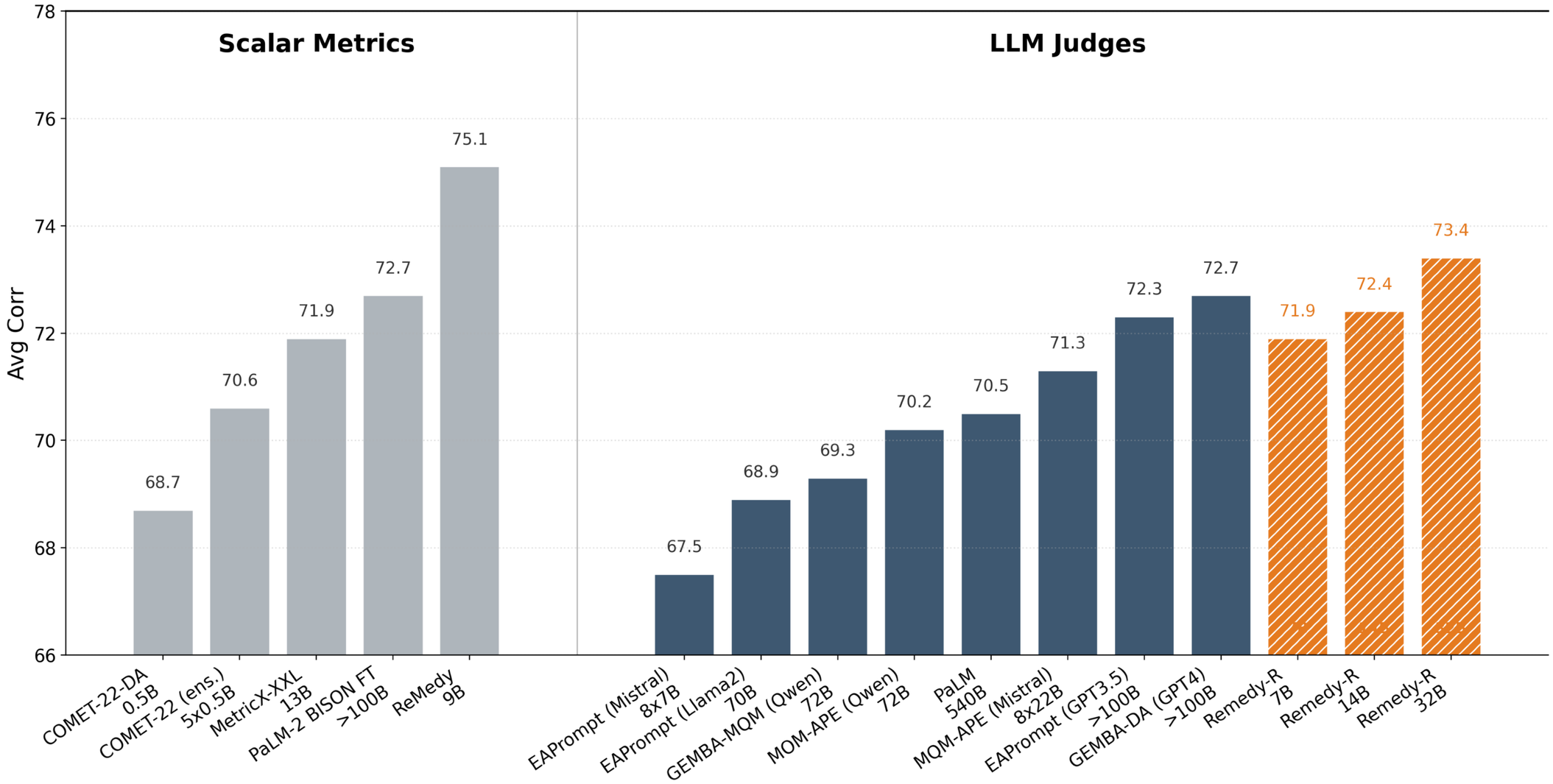
- There are some awkward phrasings, such as "in der Einkommensklasse bin, in der mehr Steuern gezahlt werden," which could be simplified.

- The use of "maßgeblich dazu beitragen" is not clear and does not fit the context.

3. Completeness:
- The original text also states that the speaker is willing to wait a few days to get the product so that a gig worker doesn’t have to be overworked. This part is also not clearly conveyed in the translation.

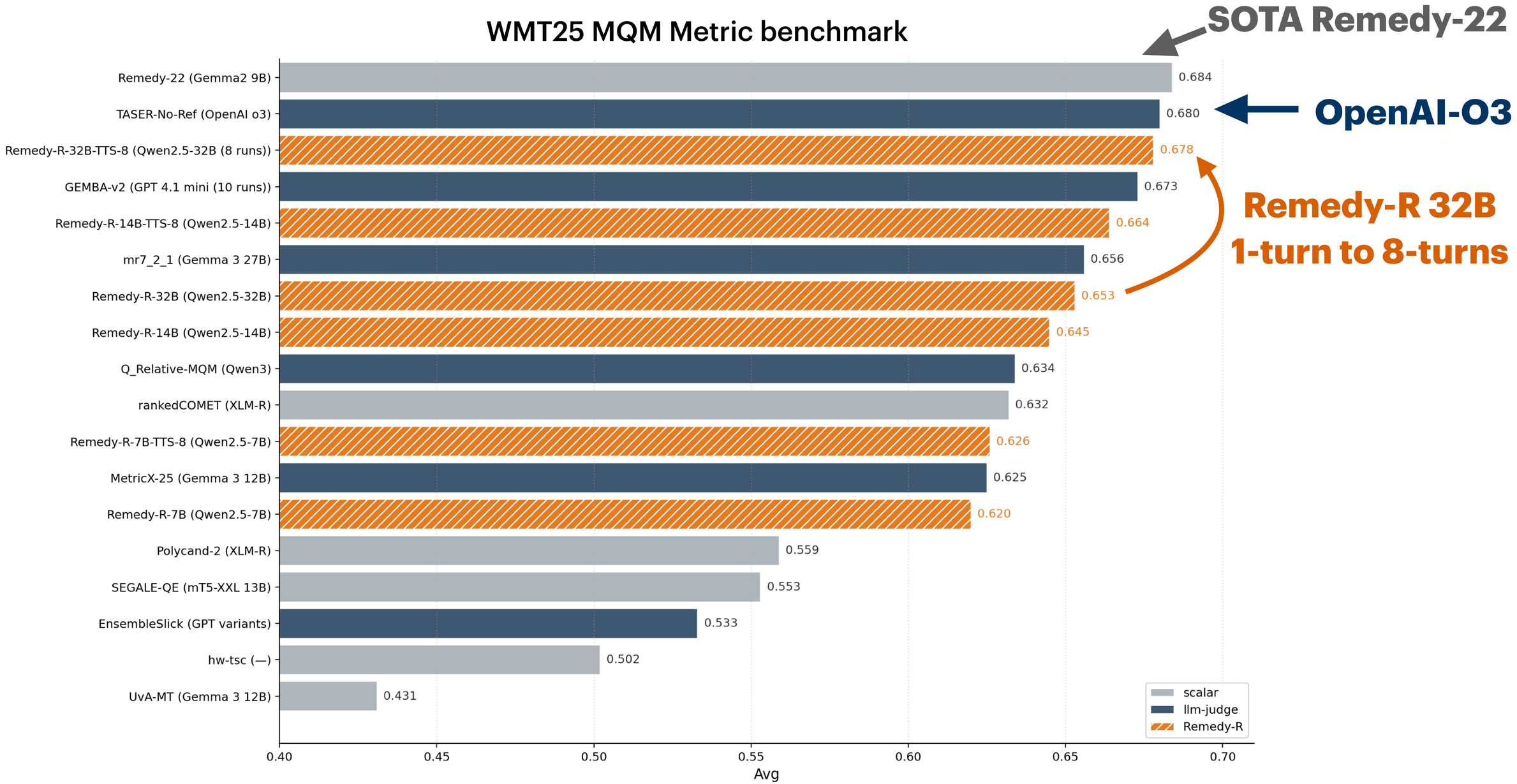
Score: 65

Remedy-R is competitive with strong metrics

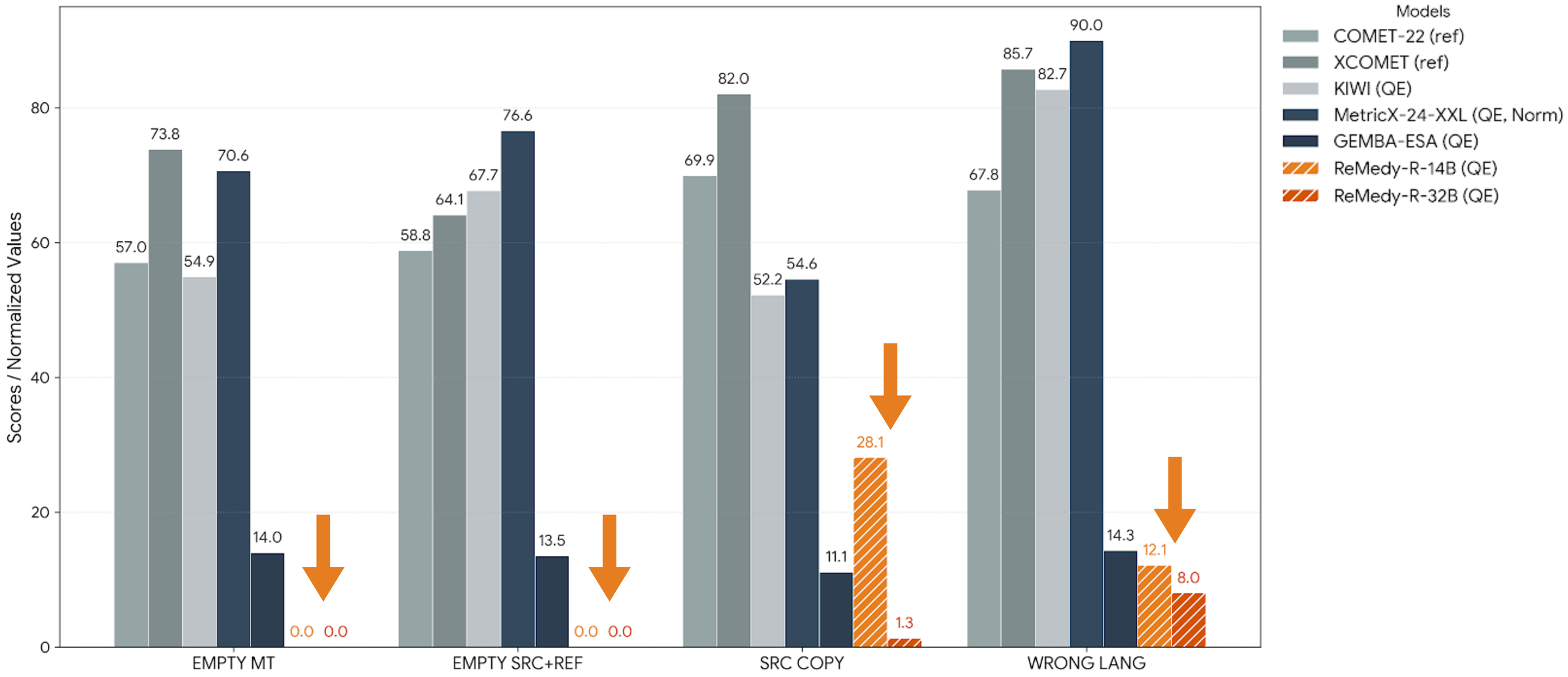


Remedy-R achieves high performance in estimating translation quality

Remedy-R TTS can further help



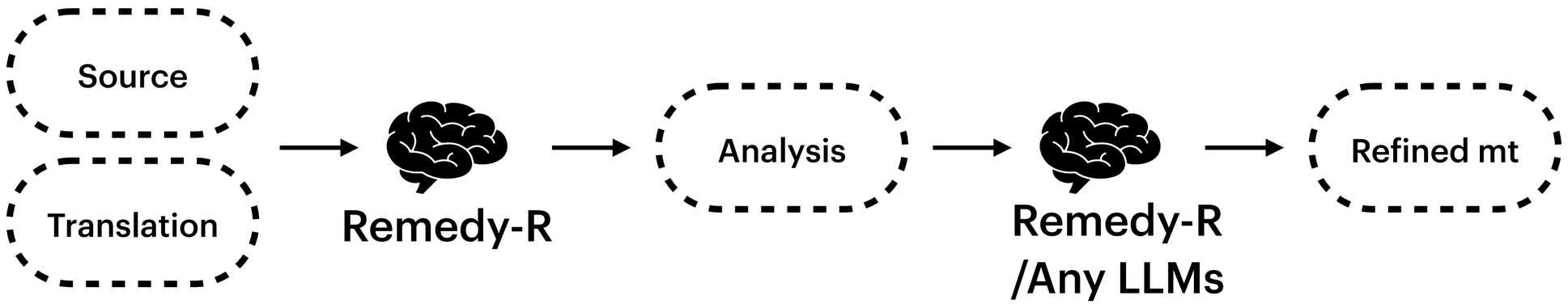
Remedy-R remains robust on OOD outputs



OOD robustness testset: Lower is Better

Remedy-R Agent

- Using Remedy-R in an **agentic way** (**Feedback-Refinement**):
  - Step 1: Prepare a source and translation.
  - Step 2: Feed to Remedy-R for quality analysis.
  - Step 3: Using the explanations from Remedy-R to improve the translation.
    - This can be using the Remedy-R itself for refinement or any LLMs.



Remedy-R Agent Example

Source: A final push for female equality. | **Reference**: Not Provided.

Translation: 女性の平等への最後の push。

[QE Evaluation/Feedback (no ref)]

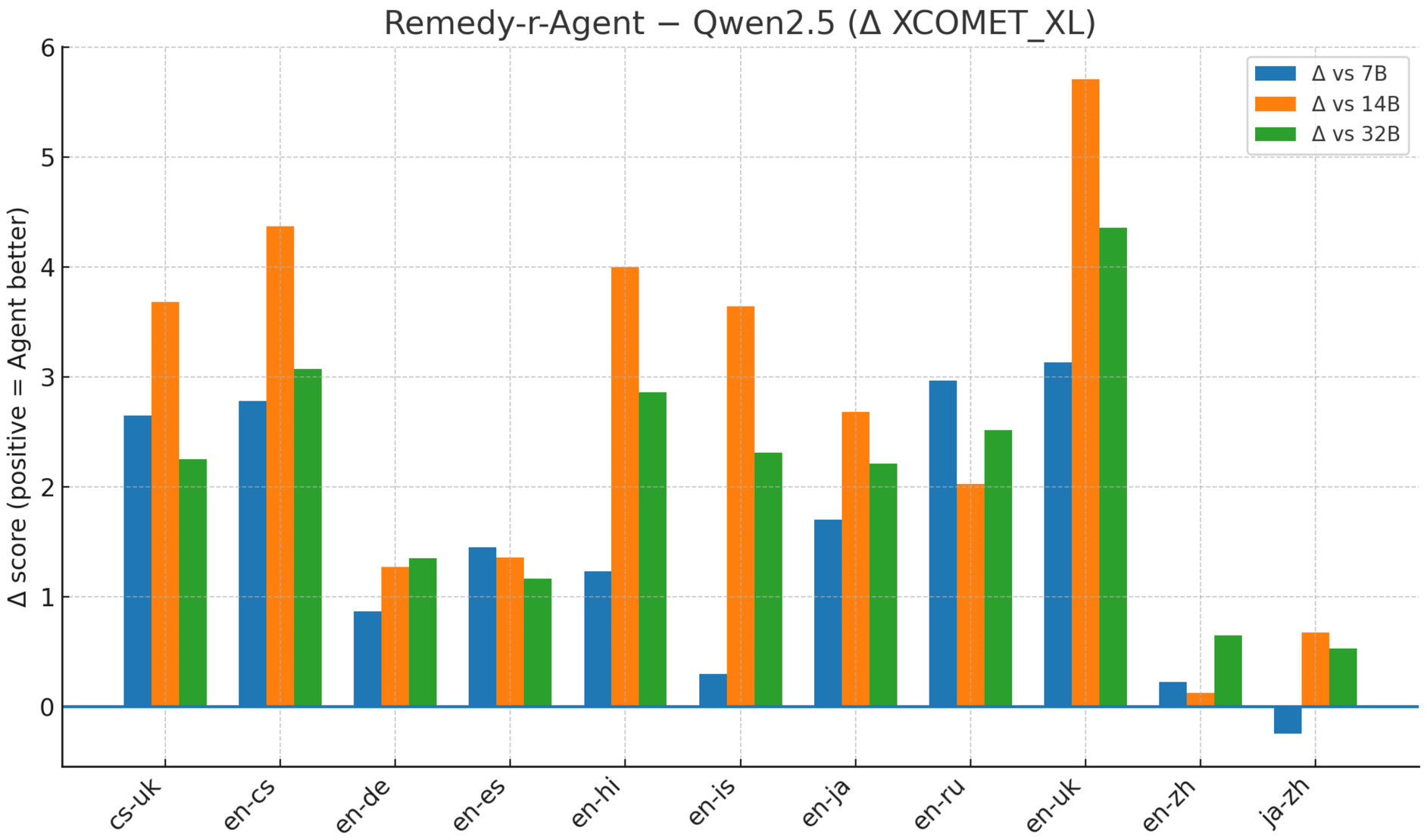
Let's evaluate the translation "女性の平等への最後の push" step by step based on the given criteria:

1. **Accuracy**:
   - The translation attempts to convey the meaning of "A final push for female equality."
   - "女性の平等" correctly translates "female equality."
   - "最後の" correctly translates "final."
   - **However, "push" is left in English** rather than being translated to Japanese.
2. Fluency:
   - The translation is mostly fluent but the inclusion of "push" in English breaks the flow.
   - **"push" is typically translated as "推進力" (suishōryoku) or "押し" (oshi) in Japanese.**
3. Completeness:
   - The translation is mostly complete but "push" is not translated, which affects the completeness.
   - A more complete translation would fully convey the concept of "push" in Japanese.

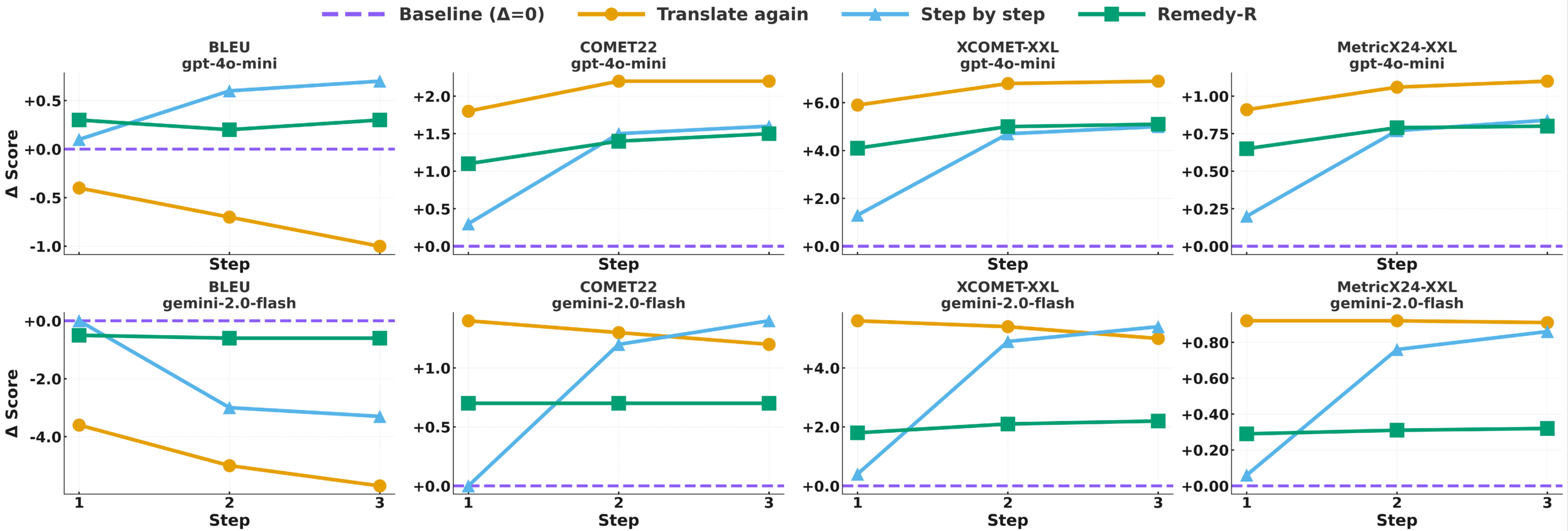
Given these points, the translation is partially accurate and mostly fluent but lacks completeness and language correctness due to the inclusion of "push" in English.

Score: 75 (0-100)

Remedy-R Agent - WMT24



Remedy-R Agent on commercial LLMs



Remedy-R Agent can refine translations from strong commercial LLMs, achieving half performance gains of self-refinement (GPT4/Gemini-2 refines itself)

Conclusions

Takeaways:

- QE is becoming an optimization signal.
- We need multi-dimensional assessments.
- Reasoning enables explainable, robust, and actionable QE.

Open challenges:

- Efficiency vs utility.
- Faithful reasoning.
- Score calibration and integer bias.
- Dimension-wise validation remains difficult.