Guiding Large Language Models to Post-edit Machine Translation with Error Annotations

Dayeon Ki Marine Carpuat

University of Maryland, College Park
dayeonki@umd.edu marine@umd.edu

NAACL 2024

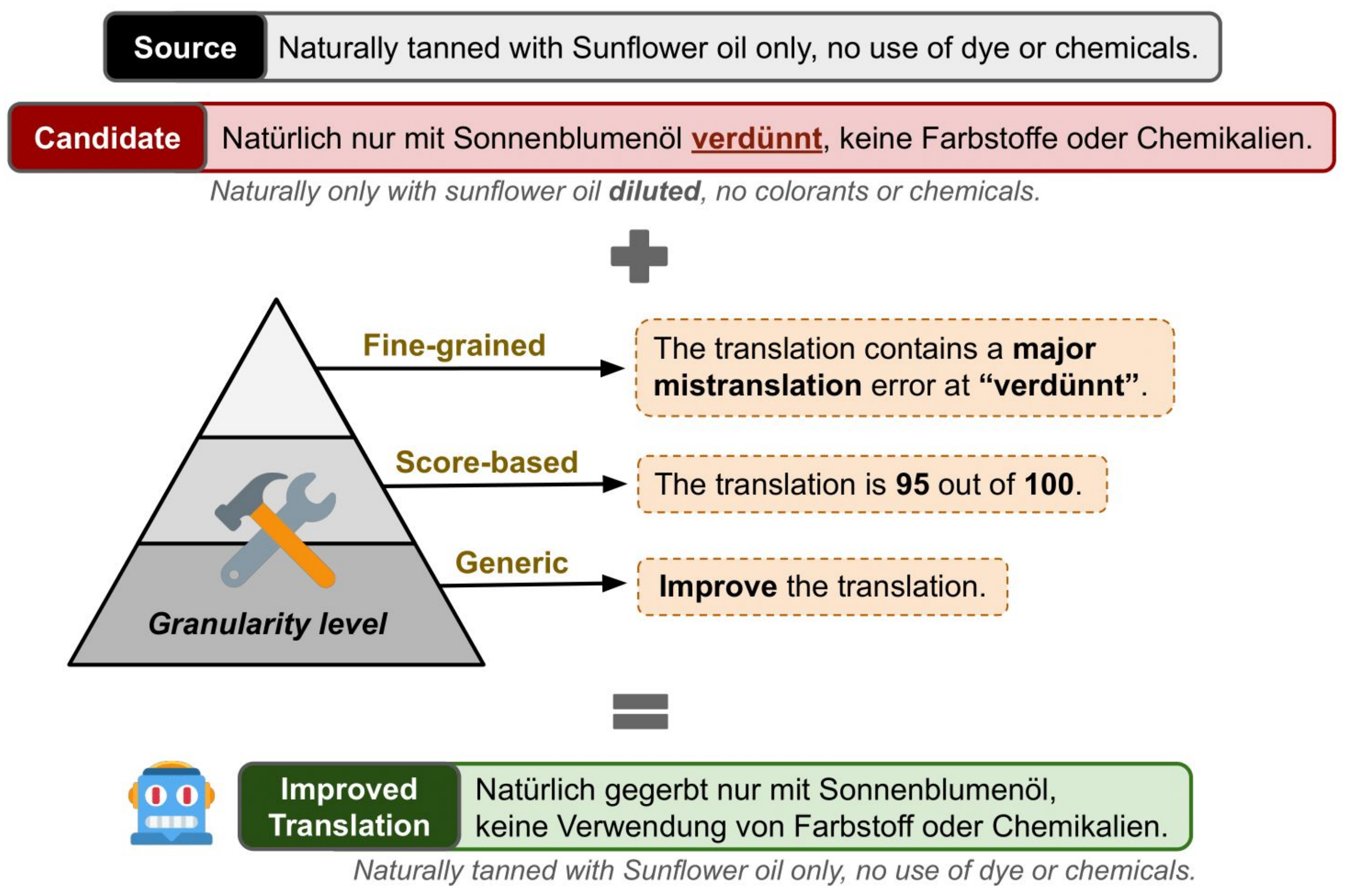
Paper

Twitter

Motivation

- Performance of LLMs remains uneven with variation across models, languages, and translation directions.
- Exploit the complementary strengths of LLMs and supervised Machine Translation system.

Can smaller LLMs use fine-grained feedback to refine their translations?


Source Naturally tanned with Sunflower oil only, no use of dye or chemicals.

Candidate Natürlich nur mit Sonnenblumenöl verdünnt, keine Farbstoffe oder Chemikalien.

Naturally only with sunflower oil ***diluted****, no colorants or chemicals.*

+

Fine-grained: The translation contains a **major mistranslation** error at **"verdünnt"**.

Score-based: The translation is **95** out of **100**.

Generic: **Improve** the translation.

Granularity level

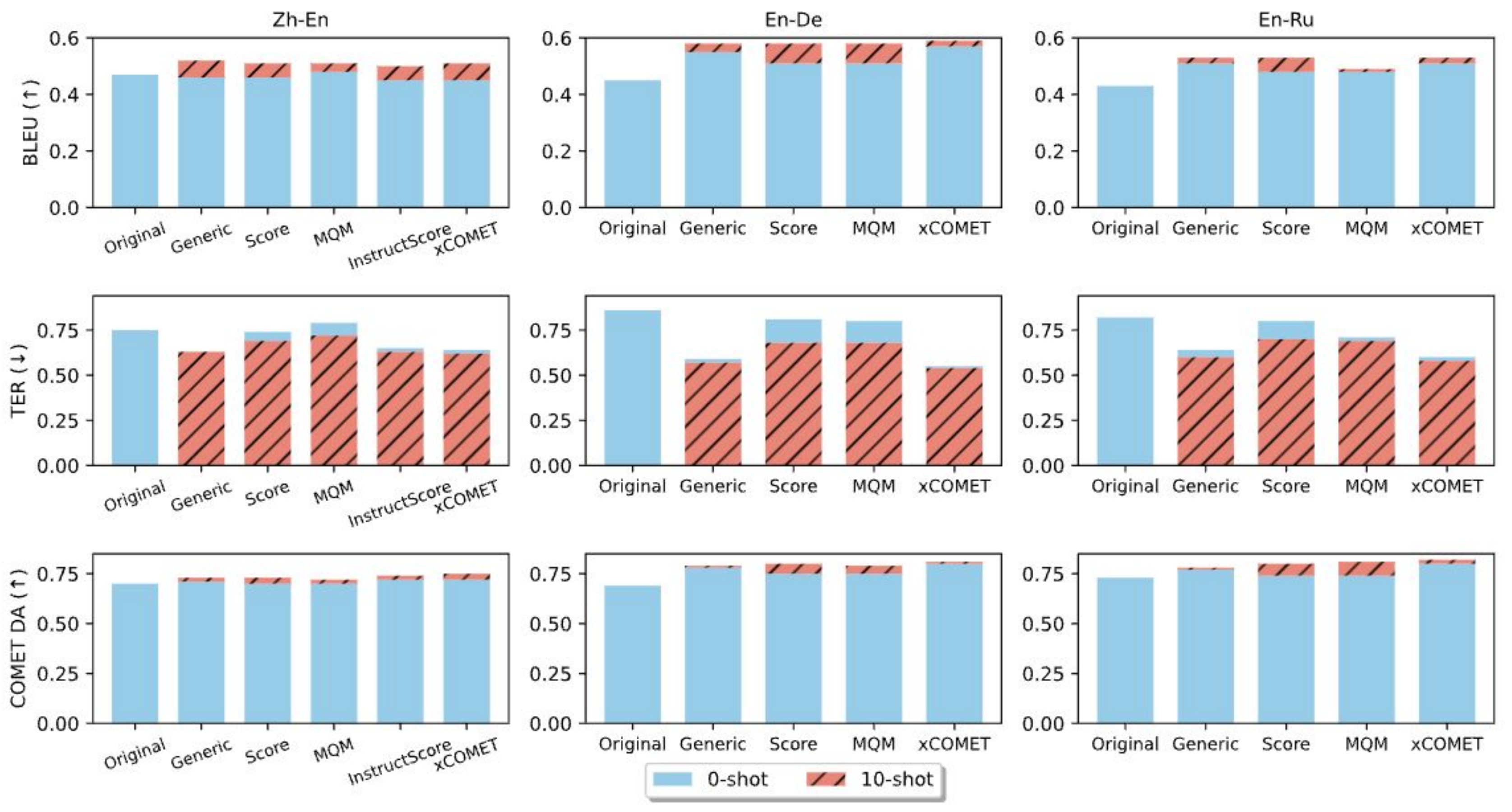
=

Improved Translation Natürlich gegerbt nur mit Sonnenblumenöl, keine Verwendung von Farbstoff oder Chemikalien.

Naturally tanned with Sunflower oil only, no use of dye or chemicals.

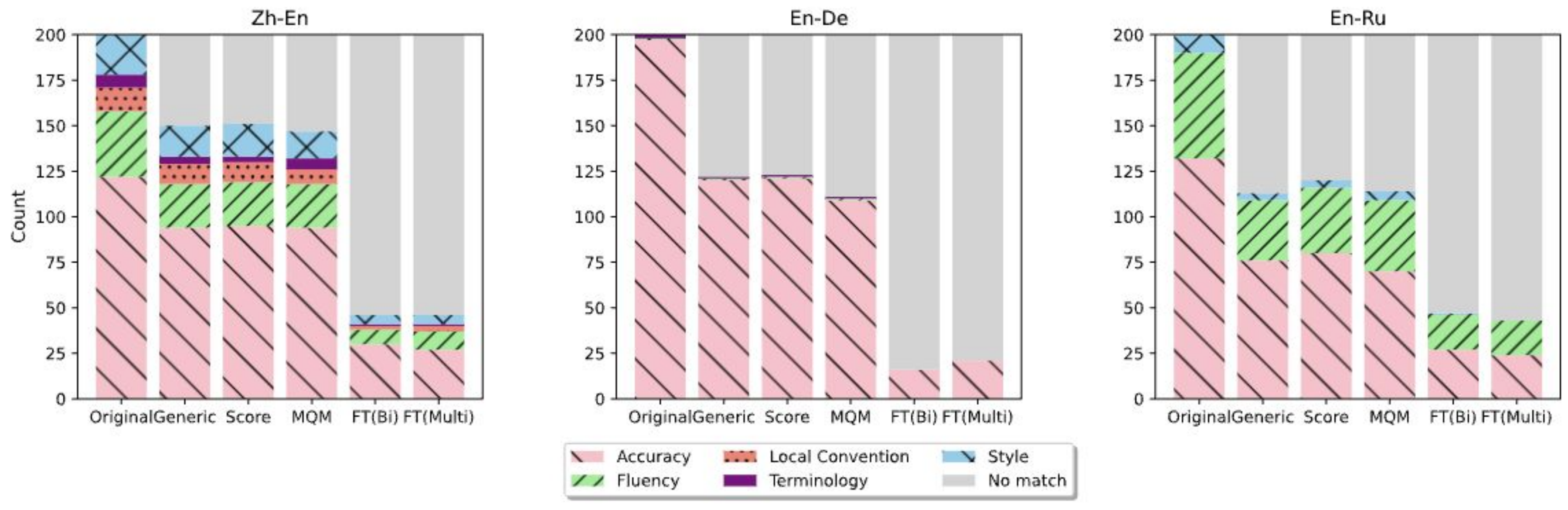
Prompting Results

- Prompting LLaMA-2 to post-edit with any forms of feedback consistently improve MT performance.
- Most forms of our tested feedback, **regardless of granularity, converge to similar performance.**



Error Analysis

- Fine-tuning with error annotations help **align LLM behavior** with the provided feedback.



Types of Feedback

Category	Prompt
Generic	Improve the translation from English to German without any explanation. English: *The newer items are bagged only.* German: *Neue Gegenstände werden nur mit Gepäck versehen.* Improved German:
Score	Improve the translation from English to German without any explanation. This translation is scored 85 out of 100. English: *The newer items are bagged only.* German: *Neue Gegenstände werden nur mit Gepäck versehen.* Improved German:
Fine-grained	Improve the translation from English to German based on the identified errors without any explanation. (1) There is a major mistranslation error at "mit Gepäck versehen". English: *The newer items are bagged only.* German: *Neue Gegenstände werden nur mit Gepäck versehen.* Improved German:

Fine-tuning Results

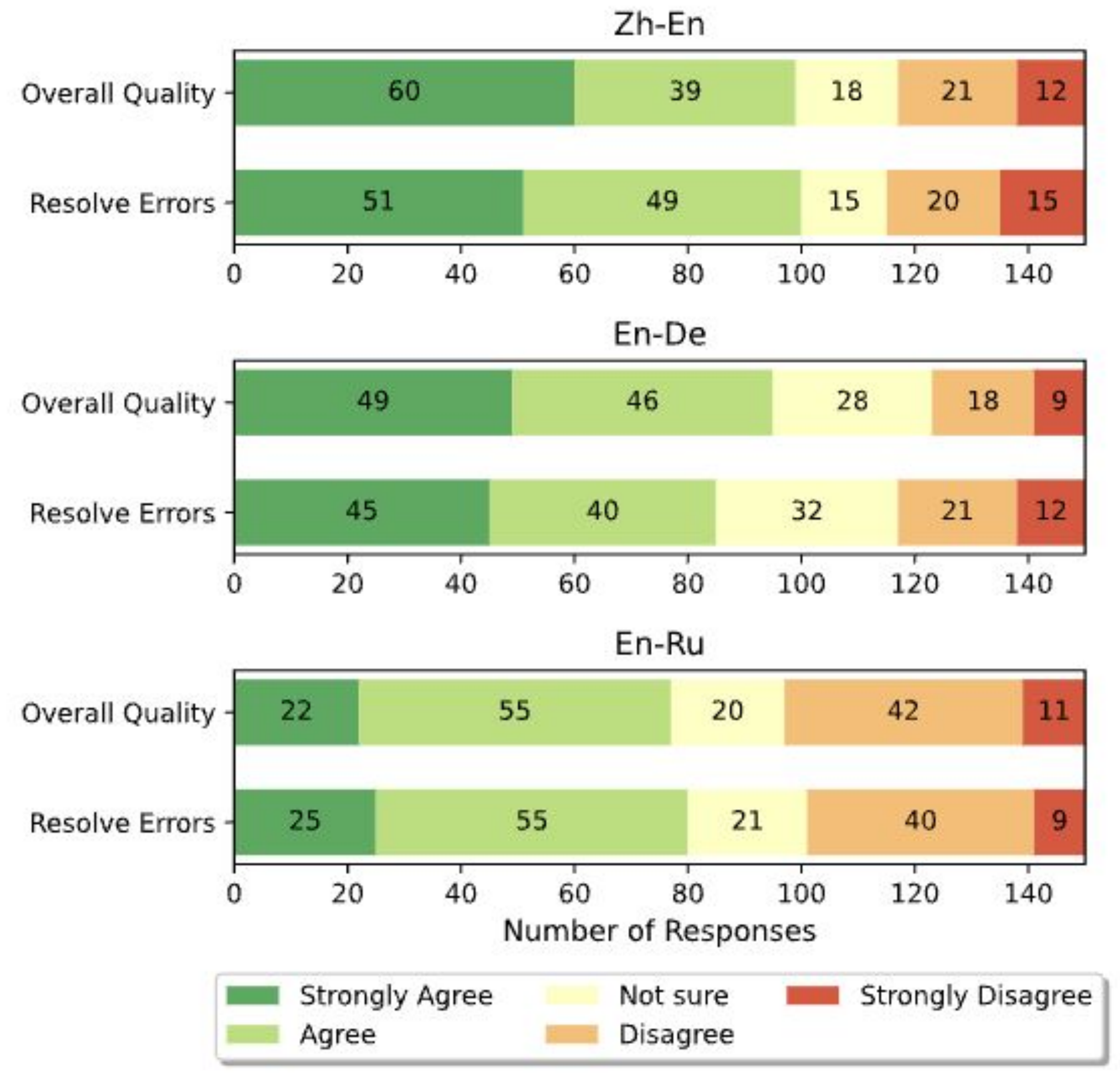
Automatic Evaluation

- Fine-tuning LLaMA-2 with error annotated translations gives an **extra boost** on top of prompting results.

Language	Type	BLEU (↑)	TER (↓)	COMET$_{DA}$ (↑)
Zh-En	Original	0.47	0.75	0.70
	prompt (k=0)	0.48†	0.72†	0.70†
	prompt (k=10)	0.51†	0.65†	0.72†
	FT (Bi)	**0.53**†	0.63†	**0.76**†
	FT (Multi)	**0.53**†	**0.61**†	**0.76**†
En-De	Original	0.45	0.86	0.69
	prompt (k=0)	0.51†	0.68†	0.75†
	prompt (k=10)	0.58†	0.56†	0.79†
	FT (Bi)	0.56†	0.58†	0.79†
	FT (Multi)	**0.59**†	**0.55**†	0.79†
En-Ru	Original	0.43	0.82	0.73
	prompt (k=0)	0.48†	0.69†	0.74†
	prompt (k=10)	0.49†	0.67†	0.79†
	FT (Bi)	0.51†	0.65†	**0.80**†
	FT (Multi)	**0.52**†	**0.63**†	**0.80**†

Human Evaluation

- Fine-tuning **rewrites for naturalness** in the target language. (*better explains the context, flows better in target language)*



Discussions

- Post-editing MT **does not require the largest proprietary LLMs** and can be done with **smaller open-source models.**
- Building on LLMs fine-tuned for many translation related tasks is a promising direction for encouraging transfer learning from limited amounts of annotation.