Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models

Qingyu Lu[◇], Baopu Qiu[♭], Liang Ding[♯], Kanjian Zhang[◇♠*], Tom Kocmi[♡], Dacheng Tao[♮]

[◇]Southeast University [♭]Nanjing University [♯]The University of Sydney

[♠]Southeast University Shenzhen Research Institute [♡]Microsoft [♮]Nanyang Technological University

luqingyu@seu.edu.cn, qiubaopu@smail.nju.edu.cn,

liangding.liam@gmail.com, tomkocmi@microsoft.com

https://github.com/Coldmist-Lu/ErrorAnalysis_Prompt

TL;DR: We propose a new prompting method – "Error Analysis Prompting" for translation evaluation. By combining Chain-of-Thoughts and Error Analysis, this technique emulates human evaluation framework MQM and produces explainable and reliable MT evaluations.

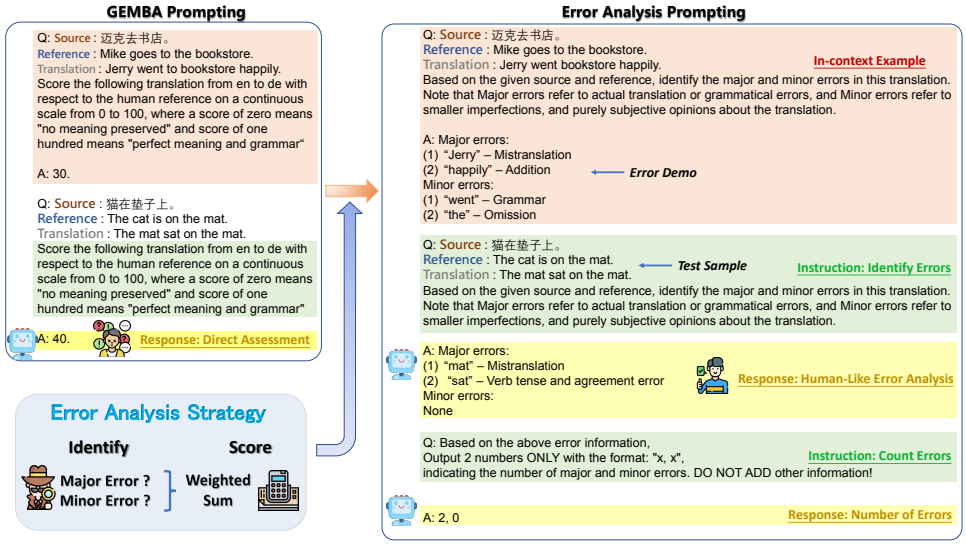

Figure: A comparative overview between GEMBA Prompting and our proposed Error Analysis Prompting in assessing the MT quality with LLMs.

Models	Metrics / Prompts	Ref?	System-Level Acc. All (3 LPs)	Segment-Level Acc* En-De	Segment-Level Acc* En-Ru	Segment-Level Acc* Zh-En
Baselines	MetricsX-XXL	✓	85.0	60.4	60.6	54.4
	BLEURT20	✓	84.7	56.8	54.0	48.9
	COMET22	✓	83.9	59.4	57.7	53.6
	UniTE	✓	82.8	59.8	57.7	51.7
	COMET-QE	✗	78.1	55.5	53.4	48.3
	UniTE-src	✗	75.9	58.2	55.4	50.8
	MaTESe-QE	✗	74.8	57.2	49.9	49.4
Llama2-70b-Chat	GEMBA	✓	74.1	53.7	48.8	45.4
	EAPrompt	✓	85.4 (+11.3)	55.2(+1.5)	51.4†(+2.6)	**50.2†(+4.8)**
	GEMBA	✗	72.6	54.1	47.8	45.0
	EAPrompt	✗	85.8 (+13.2)	55.0†(+0.9)	51.6†(+3.8)	49.3†(+4.3)
Mixtral-8x7b-Instruct	GEMBA	✓	69.7	54.8	48.3	46.7
	EAPrompt	✓	84.0 (+14.3)	53.8 (-1.0)	50.6†(+2.3)	48.2†(+1.5)
	GEMBA	✗	74.1	54.8	47.5	46.2
	EAPrompt	✗	82.5 (+8.4)	54.1 (-0.7)	49.9†(+2.4)	48.3†(+1.1)
GPT-3.5-Turbo	GEMBA	✓	86.5	55.2	49.5	48.2
	EAPrompt	✓	91.2 (+4.7)	**56.7†(+1.5)**	53.3†(+3.8)	50.0†(+1.8)
	GEMBA	✗	86.9	54.7	50.0	47.6
	EAPrompt	✗	89.4 (+2.5)	55.7†(+1.0)	**53.4†**(+3.4)	48.8†(+1.2)

Table: Performance of metrics using pairwise accuracy (%) at the system level And pairwise accuracy with tie calibration (%) at the segment level.

Findings:

- EAPrompt significantly enhances the performance of LLMs at the system level.
- EAPrompt surpasses GEMBA in 8 out of 9 test scenarios at the segment level.
- EAPrompt's strong performance remain consistent even in reference-free settings.

Prompt	Demo of Errors: Detailed	Demo of Errors: Itemized	Type of Queries: 1-step	Type of Queries: 2-step	Mixtral-8x7b-Instruct: All (3 LPs)	Mixtral-8x7b-Instruct: En-De	Mixtral-8x7b-Instruct: En-Ru	Mixtral-8x7b-Instruct: Zh-En	Llama2-70b-Chat: All (3 LPs)	Llama2-70b-Chat: En-De	Llama2-70b-Chat: En-Ru	Llama2-70b-Chat: Zh-En
GEMBA	-	-	-	-	69.7	**54.8**	48.3	46.7	74.1	53.7	48.8	45.4
EAPrompt	✓		✓		75.2	53.4	50.0	45.0	62.0	53.7	47.0	47.8
	✓			✓	75.5	53.4	47.9	45.5	84.7	53.5	46.9	47.5
		✓	✓		60.2	53.4	45.1	45.6	56.9	53.7	48.4	**50.2**
		✓		✓	**84.0**	53.7	**50.6**	**48.2**	**85.4**	**55.2**	**51.4**	**50.2**

Table: Performance comparsion with variants of prompts for EAPrompt.

Findings:

- When designing prompts, we recommend the EAPrompt variant featuring a 2-step separated prompting approach and itemized error demonstrations.

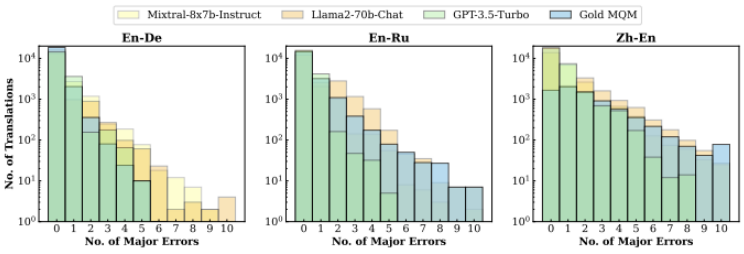

Figure: Distribution of identified error counts across LLMs and human evaluation.

Findings:

- EAPrompt adeptly distinguishes major errors from minor ones, closely aligning its error distribution with MQM.

- Designed by Qingyu Lu