Why Do Metrics Think That? Towards Understanding Large Language Models as Machine Translation Evaluators

Runzhe Zhan, Xinyi Yang, Junchao Wu, Lidia S. Chao, Derek F. Wong†

University of Macau

ICML2025 @ Workshop on Actionable Interpretability

Poster Link

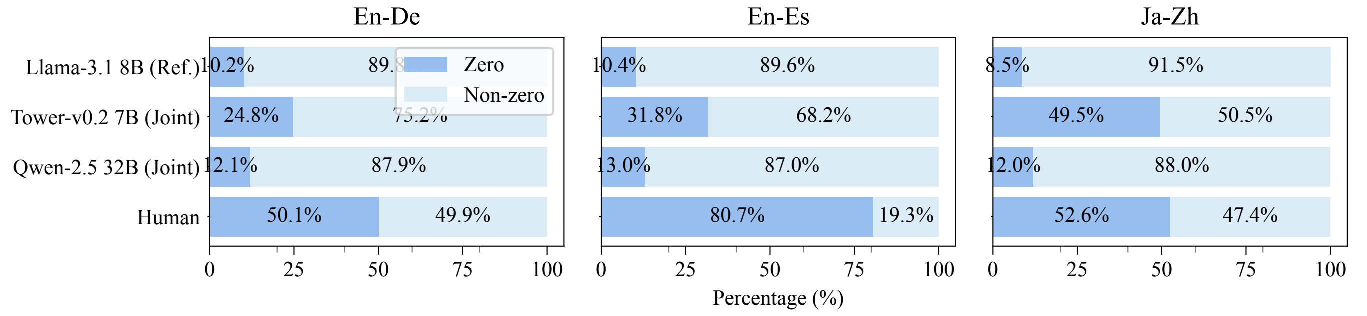
Motivation

The black-box nature of LLM-based MT metrics remains a question.

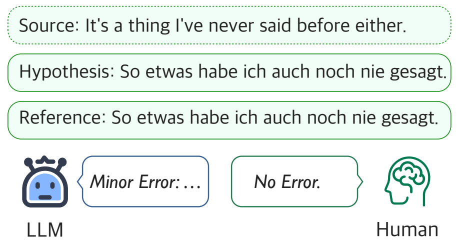
- ➢ Do the behaviors of LLM-based MT metrics align with human evaluation?
- ➢ How can insights from interpretability analyses enhance the MT metrics?

Goal:

Investigate how evaluation materials are processed by LLM-based MT metrics.
Enhance alignment with human evaluation for better metric reliability.



✘ Identify Misalignment Problem


Source: It's a thing I've never said before either.

Hypothesis: So etwas habe ich auch noch nie gesagt.

Reference: So etwas habe ich auch noch nie gesagt.

Key Misalignment:

Overestimation: LLM-based metrics detect more errors (Non-zero) than there actually are.

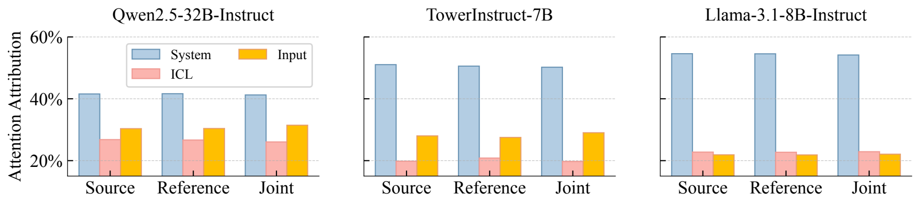
Tracing Misalignment: Unpacking the Information Flow

How is Input Information Processed?

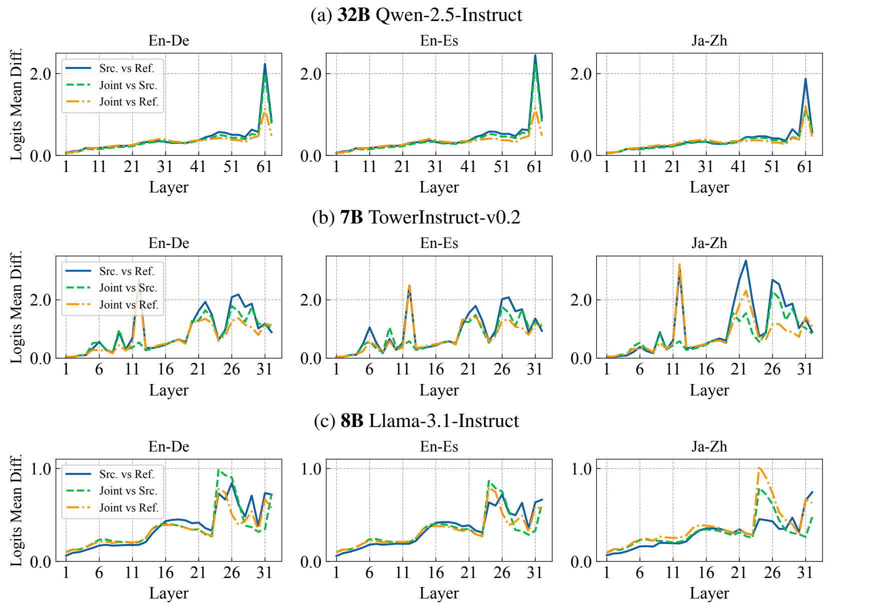
- Gap in both entropy and output probability between Correct and Overestimated predictions.

	En-De				En-Es				Ja-Zh			
	$\mathrm{Ent}_{\mathrm{OV.}}$	$\mathrm{Ent}_{\mathrm{Cor.}}$	$\mathrm{Prob}_{\mathrm{OV.}}$	$\mathrm{Prob}_{\mathrm{Cor.}}$	$\mathrm{Ent}_{\mathrm{OV.}}$	$\mathrm{Ent}_{\mathrm{Cor.}}$	$\mathrm{Prob}_{\mathrm{OV.}}$	$\mathrm{Prob}_{\mathrm{Cor.}}$	$\mathrm{Ent}_{\mathrm{OV.}}$	$\mathrm{Ent}_{\mathrm{Cor.}}$	$\mathrm{Prob}_{\mathrm{OV.}}$	$\mathrm{Prob}_{\mathrm{Cor.}}$
Qwen2.5-32B-Instruct												
Src	0.66	1.73	**0.85**	0.53	0.57	1.74	**0.88**	0.53	0.54	1.55	**0.88**	0.60
Ref	1.03	2.03	**0.68**	0.52	0.98	2.09	**0.72**	0.50	0.88	1.93	**0.72**	0.55
Joint	0.94	1.81	**0.73**	0.49	0.86	1.68	**0.77**	0.52	0.74	1.71	**0.80**	0.54
Avg.	0.88	1.86	**0.75**	0.51	0.80	1.84	**0.79**	0.52	0.72	1.73	**0.80**	0.56
TowerInstruct-7B-v0.2												
Src	0.88	1.11	**0.66**	0.55	0.93	1.05	**0.61**	0.57	0.90	0.92	0.58	**0.64**
Ref	1.15	1.45	**0.57**	0.48	1.15	1.37	**0.56**	0.51	1.06	1.29	**0.53**	0.51
Joint	0.94	1.20	**0.62**	0.56	0.84	1.10	**0.69**	0.63	0.53	0.63	**0.86**	0.83
Avg.	0.99	1.25	**0.62**	0.53	0.97	1.18	**0.62**	0.57	0.83	0.95	**0.66**	**0.66**
Llama-3.1-8B-Instruct												
Src	0.52	1.25	**0.87**	0.52	0.39	1.18	**0.91**	0.57	0.20	1.44	**0.96**	0.54
Ref	0.60	1.17	**0.81**	0.61	0.57	1.19	**0.83**	0.58	0.41	1.36	**0.89**	0.54
Joint	0.36	1.03	**0.92**	0.64	0.36	1.09	**0.92**	0.58	0.13	1.10	**0.98**	0.66
Avg.	0.49	1.15	**0.87**	0.59	0.44	1.15	**0.88**	0.58	0.25	1.30	**0.94**	0.58

- High-performing evaluators tend to weigh input materials more than ICL.



Where is Input Information Processed?


(a) 32B Qwen2.5-Instruct

(b) 7B TowerInstruct-v0.2

(c) 8B Llama3.1-Instruct

Logit Lens

$$z^{(l)} = \mathrm{Norm}(\underbrace{\mathbf{h}^{(l)}}_{\text{Hidden State}}) \cdot \underbrace{\mathbf{E}_U^T}_{\text{Unembed.}}$$

Mean Absolute Difference

$$\mathrm{MAD}(z_{\texttt{Src.}}, z_{\texttt{Ref.}}) = \frac{1}{\underbrace{V}_{\text{Vocabulary Length}}} \sum_{i=1}^{V} |z^{l,i}_{\texttt{Src.}} - z^{l,i}_{\texttt{Ref.}}|$$

- Evaluation task is processed by mid-to-high layers
  ➠ can be used in sparse fine-tuning.

Alleviating Misalignment: Insights into Improvements

Improving Alignment Effectiveness

- **Error-Free ICL**: Replace one demonstration with an error-free one.
- **MQM SFT**: Supervised fine-tuning with human MQM annotations.

	En-De		En-Cs		Ja-Zh		Avg.		
Models	SPA (%)	Acc^*_{eq}	SPA (%)	Acc^*_{eq}	SPA (%)	Acc^*_{eq}	SPA (%)	Acc^*_{eq}	All
Llama-3.1-8B (Src.)	70.5	44.5	71.8	68.0	77.7	43.5	73.3	52.0	62.7
+ *EF ICL*	73.0	45.4	73.7	68.0	77.8	43.5	74.8	52.3	63.6
+ *MQM SFT*	**78.9**	**48.3**	**81.7**	68.0	**80.5**	**43.5**	**80.4**	**53.3**	**66.8**
Llama-3.1-8B (Ref.)	85.4	45.5	72.6	68.0	89.2	43.5	82.4	52.3	67.4
+ *EF ICL*	**85.9**	45.1	72.4	68.0	89.7	43.5	82.7	52.2	67.4
+ *MQM SFT*	84.7	**47.7**	**79.3**	68.0	**91.8**	**48.7**	**85.3**	**54.8**	**70.0**
Llama-3.1-8B (Joint.)	74.4	45.4	71.2	68.0	84.4	48.0	76.7	53.8	65.2
+ *EF ICL*	74.1	46.0	78.4	68.0	86.3	**45.9**	79.6	53.3	66.5
+ *MQM SFT*	**78.8**	**49.0**	**80.6**	68.0	**90.5**	44.8	**83.3**	**53.9**	**68.6**

"SPA" and "Acc" are correlation metrics (w.r.t. human evaluation).

Improving Alignment Efficiency

- Only update the parameters of mid-to-high layers when SFT.

Model	Source	Reference	Joint
Llama-3.1-8B	62.7	67.4	65.2
+ Full	66.8	70.0	68.6
+ Sparse	66.8	70.0	67.8
Recovery	100.0%	100.0%	98.8%
+ LoRA Full	66.8	68.6	68.7
+ LoRA Sparse	66.8	67.8	68.6
Recovery	100.0%	98.8%	99.9%

Correlation Performance Recovery%

Is Overestimation Misalignment Addressed by Common Practices?

- **False Negative (FN)**: The model incorrectly predicts a non-zero score (an error) for a translation that was actually error-free.
- **FNR (False Negative Rate)**:

1- Recall, measure of overestimation.

	Avg.	En-De	En-Es	Ja-Zh
Llama-3.1-8B (Src.)	91.2	86.0	91.2	96.5
+ *EF ICL*	86.4	80.0	84.2	95.0
+ *MQM SFT*	**57.7**	**46.0**	**44.0**	**83.2**
Llama-3.1-8B (Ref.)	86.1	83.2	88.2	86.7
+ *EF ICL*	84.3	81.0	86.8	85.2
+ *MQM SFT*	**67.5**	**62.0**	**64.6**	**75.8**
Llama-3.1-8B (Joint.)	90.1	86.1	91.6	92.6
+ *EF ICL*	86.7	83.0	88.2	89.0
+ *MQM SFT*	**51.7**	**40.0**	**47.5**	**67.7**

- From the perspective of correlation metrics, Yes. But...
- **Entropy & Output Probability**: the Gap between Correct and Overestimated predictions still exists.

	En-De				En-Es				Ja-Zh			
	$\mathrm{Ent}_{\mathrm{OV.}}$	$\mathrm{Ent}_{\mathrm{Cor.}}$	$\mathrm{Prob}_{\mathrm{OV.}}$	$\mathrm{Prob}_{\mathrm{Cor.}}$	$\mathrm{Ent}_{\mathrm{OV.}}$	$\mathrm{Ent}_{\mathrm{Cor.}}$	$\mathrm{Prob}_{\mathrm{OV.}}$	$\mathrm{Prob}_{\mathrm{Cor.}}$	$\mathrm{Ent}_{\mathrm{OV.}}$	$\mathrm{Ent}_{\mathrm{Cor.}}$	$\mathrm{Prob}_{\mathrm{OV.}}$	$\mathrm{Prob}_{\mathrm{Cor.}}$
Llama-3.1 8B												
Src.	0.52	1.25	**0.87**	0.52	0.39	1.18	**0.91**	0.57	0.20	1.43	**0.96**	0.54
Ref.	0.60	1.17	**0.81**	0.60	0.57	1.19	**0.83**	0.58	0.40	1.36	**0.89**	0.54
Joint.	0.36	1.03	**0.92**	0.64	0.36	1.09	**0.92**	0.58	0.13	1.10	**0.98**	0.66
Avg.	0.49	1.15	**0.87**	0.59	0.44	1.15	**0.88**	0.58	0.24	1.30	**0.94**	0.58
+ *EF ICL*												
Src	0.31	1.1	**0.94**	0.69	0.22	0.92	**0.96**	0.75	0.11	1.39	**0.98**	0.63
Ref	0.29	1.38	**0.93**	0.56	0.26	1.30	**0.94**	0.59	0.19	1.58	**0.96**	0.50
Joint	0.14	0.63	**0.98**	0.85	0.13	0.58	**0.98**	0.87	0.05	0.63	**0.99**	0.88
Avg.	0.24	1.04	**0.95**	0.70	0.20	0.94	**0.96**	0.73	0.11	1.20	**0.98**	0.67
+ *MQM SFT*												
Src	0.23	0.68	**0.95**	0.78	0.15	0.52	**0.97**	0.83	0.04	0.35	**0.99**	0.91
Ref	0.24	0.84	**0.94**	0.72	0.27	0.78	**0.93**	0.75	0.13	0.54	**0.97**	0.86
Joint	0.14	0.29	**0.97**	0.93	0.13	0.27	**0.98**	0.94	0.05	0.12	**0.99**	0.98
Avg.	0.20	0.60	**0.95**	0.81	0.18	0.52	**0.96**	0.84	0.07	0.34	**0.99**	0.91

Conclusions

We have Identified key internal factors crucial to evaluation decisions.
However, addressing the overestimation issue remains challenging through common practices.

Contact

rzzhan_ovo

nlp2ct.runzhe@gmail