TASER: Translation Assessment via Systematic Evaluation and Reasoning

Monishwaran Maheswaran[†‡*] **Marco Carini**[‡] **Christian Federmann**[‡] **Tony Diaz**[‡]
[†]University of California, Berkeley [‡]Apple
{monishwaran}@berkeley.edu, {m_carini, chrf, tonydiaz}@apple.com

Abstract

We introduce TASER (Translation Assessment via Systematic Evaluation and Reasoning), a metric that uses Large Reasoning Models (LRMs) for automated translation quality assessment. TASER harnesses the explicit reasoning capabilities of LRMs to conduct systematic, step-by-step evaluation of translation quality. We evaluate TASER on the WMT24 Metrics Shared Task across both reference-based and reference-free scenarios, demonstrating state-of-the-art performance. In system-level evaluation, TASER achieves the highest soft pairwise accuracy in both reference-based and reference-free settings, outperforming all existing metrics. At the segment level, TASER maintains competitive performance with our reference-free variant ranking as the top-performing metric among all reference-free approaches. Our experiments reveal that structured prompting templates yield superior results with LRMs compared to the open-ended approaches that proved optimal for traditional LLMs. We evaluate o3, a large reasoning model from OpenAI, with varying reasoning efforts, providing insights into the relationship between reasoning depth and evaluation quality. The explicit reasoning process in LRMs offers interpretability and visibility, addressing a key limitation of existing automated metrics. Our results demonstrate that Large Reasoning Models show a measurable advancement in translation quality assessment, combining improved accuracy with transparent evaluation across diverse language pairs.

1 Introduction

Large Language Models (LLMs) have been demonstrated in zero-shot and few-shot translation scenarios, achieving comparable results to dedicated machine translation systems (Jiao et al., 2023; Robinson et al., 2023). Previous work by (Kocmi and Federmann, 2023b) used Large Language Models (LLMs) through prompting to assess the quality of a machine translation. In their work, GEMBA-DA, they prompt a LLM such as GPT to assess the quality of the translation. Their investigation shows that with straightforward zero-shot prompting, LLMs show accuracy exceeding that of all other non-LLM metrics on the WMT22 (Kocmi et al., 2022) evaluation dataset. Their subsequent work, GEMBA-MQM, (Kocmi and Federmann, 2023a) expands on this investigation to detect granular translation quality errors. GEMBA-MQM uses a language agnostic prompting strategy with fixed three-shot prompting to query GPT-4 model to mark error quality spans. Their results indicate GEMBA-MQM achieves state-of-the-art accuracy for system ranking. In this paper, we introduce TASER. TASER builds on these recent findings by investigating Large Reasoning Models.

Large Reasoning Models (OpenAI et al., 2024; QwenTeam, 2024; DeepSeek-AI et al., 2025) use long chained reasoning to answer input queries. Reasoning models have shown abilities in problem-solving, coding, as well as scientific reasoning and multi-step logical inference (Zhou et al., 2025). Recent findings show that Large Reasoning Models can also be used in translation. (Liu et al., 2025) investigated LRMs at machine translation tasks. In their position paper, they identified three shifts brought about by LRMs: 1) contextual coherence, where LRMs resolve ambiguities and preserve discourse structure through explicit reasoning via context clues; 2) cultural intentionality, where models can adapt translations by inferring speaker intent, audience expectations, and socio-linguistic norms, and finally 3) self-reflection, where LRMs can iteratively refine translations during inference, correcting errors dynamically. These three shifts contribute to more nuanced translations.

In this paper, we present TASER. TASER uses LRMs with zero-shot prompting to arrive at a translation quality estimation. We define and investigate LRMs for the assessment of translation quality in

[*]Work done while interning at Apple Inc.

Proceedings of the Tenth Conference on Machine Translation, pages 1004–1010
November 8-9, 2025

both reference based and reference free scenarios. Starting with the evaluation of the prompts from earlier works that showed state-of-the-art result on non-reasoning LLMs, we iterated on the DA+SQM template used for the human assessment of the translation quality as implemented in the Appraise framework (Federmann, 2018) for WMT22 (Kocmi et al., 2022) and adapted it towards LRMs. We posit that the strengths of LRMs lead to translation quality estimation that is more aligned with human judgment, as measured in Tables 1 and 2 below.

The main contributions of this paper are as follows:

- We achieve state-of-the-art results using Large Reasoning Models for translation quality assessment on the latest WMT24 (Zerva et al., 2024) MQM metrics evaluation dataset.
- We evaluate a reasoning model from OpenAI: o3 (OpenAI, 2025) with different reasoning efforts: low and high. Reasoning efforts guide the model on how many reasoning tokens to generate before creating a response to the prompt. Our results show that for translation metric tasks, there isn't any advantage in using high reasoning effort as they both show comparable performance. Performance might however vary, if we had more fine-grained control over the reasoning effort budget.
- We conclude that TASER shows great promise and prompt further investigation into leveraging reasoning models for translation quality assessment.

2 TASER Metric

In this method, we prompt reasoning models from OpenAI with the following attributes: source language, target language, source text segment, translation segment, and optionally, the human reference segment, analogous to (Kocmi and Federmann, 2023b). After iterating and evaluating on different prompts, we observed that simple zero-shot open ended prompting does not result in the best overall assessment. The prompt that we settled on includes the attributes as listed above as well as includes more direction, particularly assessment instructions and details of what to look for during quality assessment. We leave evaluating other reasoning models and additional language pairs for future work.

3 Experimental Setup

Our experiments involve measuring the performance of TASER on the WMT24 Metrics shared task (Zerva et al., 2024), where automated metrics are evaluated against human gold labels. The goal is to predict a quality score for each segment in a given test set which can be a variant of Direct Assessment (DA) or Multidimensional Quality Metrics (MQM). We evaluate TASER across the evaluation set provided by WMT24. Similar to (Kocmi and Federmann, 2023a), we compare our method against the best-performing reference-based and reference free metrics of WMT24.

3.1 Evaluation Datasets

MQM datasets from the WMT24 (Zerva et al., 2024) are across three language pairs: English → German, English → Spanish, and Japanese → Chinese. The dataset contains the source sentences, output of machine translation systems, and reference translations. The quality of each source-translation pair is annotated by at least three independent expert annotators, using DA on a scale 0-100.

3.2 Evaluation Criteria

Our evaluation is the same process as the evaluation process followed in (Freitag et al., 2024).

At the system level, the evaluation is done with soft pairwise accuracy (SPA) (Thompson et al., 2024), which addresses some of the drawbacks of standard pairwise accuracy which does not account for the uncertainty of the system ranking. SPA addresses this problem by using p-values as a proxy for certainty, where p-values are calculated between two systems using both the metric and human scores, then taking 1.0 minus the absolute difference between the two p-values as the metric's score for that pair, resulting in the same statistical conclusion as the human scores. Moreover, SPA does not reward or penalize metrics with statistical ties rather the accuracy score is proportional to whether or not the metric and human have the same level of certainty in the ranking.

At the segment level, evaluation follows the same process as (Freitag et al., 2024, 2023) where pairwise accuracy is computed with tie calibration, that is, metrics are given credit for correctly predicting ties in human scores, while automatically calibrating for each metric's natural scale. The accuracy/correlation scores are then simply averaged

for the final score, placing the metric scores on an absolute scale and independent of the performance of other metrics.

4 Results

Metric	SPA
TASER-o3-low	0.872
TASER-o3-high	0.868
TASER-o3-high	0.867
TASER-o3-low	0.864
XCOMET	0.861
MetricX-24-Hybrid	0.856
MetaMetrics-MT	0.852
MetricX-24-Hybrid-QE	0.848
gemba-esa	0.846
XCOMET-QE	0.833
COMET-22	0.824
BLEURT-20	0.821
bright-qe	0.805
MetaMetrics-MT-QE	0.802
BLCOM-1	0.789
PrismRefMedium	0.766
PrismRefSmall	0.760
damonmonli	0.739
sentinel-cand-mqm	0.739
YiSi-1	0.735
CometKiwi	0.733
BERTScore	0.714
chrF	0.700
MEE4	0.696
chrfS	0.694
spBLEU	0.671
BLEU	0.663
sentinel-ref-mqm	0.570
sentinel-src-mqm	0.570
XLsimMqm	0.509

Table 1: System level average soft pairwise accuracy (SPA) for all metrics from the WMT24 across the main language pairs: English → German, English → Spanish, Japanese → Chinese. Metrics highlighted gray did not use a reference translation.

The results for TASER on the WMT24 test dataset is reported under both reference based and reference free scenarios. The results are compared against the **MQM** gold labels. TASER is evaluated under two configurations: TASER-o3-low (low reasoning effort setting) and TASER-o3-high (high reasoning effort setting). The low-effort variant corresponds to settings where there are possibly fewer inference steps or less inference time compute as defined by (OpenAI, 2025), while the high-effort variant leverages more inference time compute. Table 1 reports soft pairwise accuracy (SPA) on the system level scenario averaged across the main language pairs: English → German, English → Spanish, Japanese → Chinese. The results in Table 1 show that TASER achieves the best performance under both reference free and reference based scenarios. The reference-free **TASER-o3-low** attains state-of-the-art results. The reference based **TASER-o3-high** outperforms all other metrics including other reference based metrics, only behind reference free **TASER-o3-low**. Table 2 reports the segment level accuracy with tie calibration. TASER achieves competitive performance overall with **TASER-o3-low**, which did not use a reference translation, achieving best overall accuracy among all non reference based metrics.

5 Conclusion

In this paper, we introduced TASER: Translation Assessment via Systematic Evaluation and Reasoning, a novel approach that uses Large Reasoning Models (LRMs) for automated translation quality assessment. Our work demonstrates that LRMs can measurably outperform traditional Large Language Models (LLMs) and existing automated metrics in evaluating translation quality. TASER achieves state-of-the-art performance on the WMT24 Metrics Shared Task when evaluated against the MQM24 dataset. TASER's performance demonstrates that the explicit reasoning capabilities of LRMs provide tangible benefits for translation assessment tasks.

In the near future, we plan to focus on exploring the interpretability advantages offered by the TASER reasoning process and how they might address the limitations of existing automated metrics. In addition, we plan to investigate TASER under open-source reasoning models.

In conclusion, our results suggest that the integration of explicit reasoning processes into evaluation metrics will play a crucial role in advancing the field of machine translation evaluation, ultimately contributing to more reliable and trustworthy automated translation systems across diverse languages and applications.

Limitations

TASER uses off the shelf Large Reasoning Models from OpenAI through prompting. The closed

Metric	Accuracy
MetaMetrics-MT	0.596
MetricX-24-Hybrid	0.586
TASER-o3-low	0.584
TASER-o3-high	0.584
TASER-o3-high	0.582
TASER-o3-low	0.581
MetricX-24-Hybrid-QE	0.580
gemba-esa	0.576
XCOMET	0.576
MetaMetrics-MT-QE	0.566
sentinel-cand-mqm	0.560
bright-qe	0.557
XCOMET-QE	0.557
COMET-22	0.554
BLEURT-20	0.550
CometKiwi	0.547
BLCOM-1	0.541
damonmonli	0.532
PrismRefMedium	0.526
YiSi-1	0.525
PrismRefSmall	0.524
XLsimMqm	0.523
BERTScore	0.522
MEE4	0.522
chrfS	0.520
chrF	0.516
spBLEU	0.516
BLEU	0.515
sentinel-ref-mqm	0.515
sentinel-src-mqm	0.515

Table 2: Segment level average accuracy with tie calibration for all metrics from the WMT24 across the main language pairs: English → German, English → Spanish, Japanese → Chinese. Metrics highlighted gray did not use a reference translation.

source nature of these models prevent fine-grained control over the reasoning chain and restrict the user from accessing the intermediate reasoning steps, which can limit the interpretability of the model's decision for the quality estimate. Moreover, with off the shelf, closed source models, there is uncertainty on whether models from OpenAI are trained on standard evaluation datasets such as those from WMT24. Therefore, we caution the reader to be mindful of potential data contamination when interpreting the provided results. WMT24 contains a limited set of language pairs which our testing is limited to and results in other language pairs could differ. TASER specific prompts were only used in TASER's evaluation, and were not used in the other LLM-based metrics we compared in Table 1 and 2. Some of the performance we saw could be attributed to the prompt alone. Finally, while LRMs can offer tangible benefits in a variety of tasks, including translation, it does come with increased inference cost when compared to LLMs.

Acknowledgements

We acknowledge gracious support from Apple without which this project would not have been completed. The authors are grateful for their peers for their feedback throughout the life cycle of this project. The authors also acknowledge their team's leadership, particularly Adam Archer and Tim Shaw for their invaluable guidance.

References

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang,

Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Christian Federmann. 2018. Appraise evaluation framework for machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023a. Gemba-mqm: Detecting translation quality error spans with gpt-4.

Tom Kocmi and Christian Federmann. 2023b. Large language models are state-of-the-art evaluators of translation quality.

Sinuo Liu, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New trends for modern machine translation with large reasoning models.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil

Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. Openai o1 system card.

OpenAI. 2025. Openai o3 and o4-mini system card.

QwenTeam. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages.

Brian Thompson, Nitika Mathur, Daniel Deutsch, and Huda Khayrallah. 2024. Improving statistical significance in human evaluation of automatic metrics via soft pairwise accuracy. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1222–1234, Miami, Florida, USA. Association for Computational Linguistics.

Chrysoula Zerva, Frederic Blain, José G. C. De Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Orasan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André Martins. 2024. Findings of the quality estimation shared task at WMT 2024: Are LLMs closing the gap in QE? In *Proceedings of the Ninth Conference on Machine Translation*, pages 82–109, Miami, Florida, USA. Association for Computational Linguistics.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Shreedhar Jangam, Jayanth Srinivasa, Gaowen Liu, Dawn Song, and Xin Eric Wang. 2025. The hidden risks of large reasoning models: A safety assessment of r1.

A TASER Prompts

Below we provide the prompt template used for the experiments described in this paper. There are two prompt templates with minimal variations to account for reference free and reference based scenarios.

A.1 Reference Free Prompt Template

```
{source_lang} Source: ```{source_seg}```
{target_lang} Machine Translation: ```{target_seg}```
Evaluate the quality of a machine translation for a given segment, using the provided source text,
machine-translated text, source language, and target language.

You must analyze the translation without access to any human reference, considering the following:
- Fluency of the translation in the target language.
- Accuracy and completeness of using the information in the source segment.
- Appropriateness of terminology and style for the target language.
- Possible mistranslations, omissions, or additions.

Think step by step:
1. First, compare the source and translation for meaning preservation, fidelity, and missing/additional
content.
2. Then, analyze fluency, grammar, and naturalness in the target language.
3. Finally, synthesize your findings into a final judgment of quality, including a justification.

Continue evaluating as above until all elements have been considered before presenting your final
output.
The output should follow this structure:"Score: <your numerical score>"

Important:
- Only use the source and MT segment for evaluation (no references).
- Always provide your reasoning before the final rating and justification.
- Output MUST be valid and must follow the structure.
- Use a continuous scale from 1 (worst) to 100 (best)
```

A.2 Reference Based Prompt Template

```
{source_lang} Source: ```{source_seg}```
{target_lang} Human Reference Translation: ```{reference_seg}```
{target_lang} Machine Translation: ```{target_seg}```
Evaluate the quality of a machine translation for a given segment, using the provided source text,
human reference translation, machine-translated text, source language, and target language.

You must analyze the machine translation in comparison to the human reference, considering the following:
- Fluency of the translation in the target language.
- Accuracy and completeness of using the information in the source segment and the human reference.
- Appropriateness of terminology and style for the target language.
- Possible mistranslations, omissions, or additions.

Think step by step:
1. First, compare the source and machine translation for meaning preservation, fidelity, and
missing/additional content.
2. Then, compare the machine translation with the human reference to analyze fluency, grammar, and
naturalness in the target language.
3. Finally, synthesize your findings into a final judgment of quality, including a justification.

Continue evaluating as above until all elements have been considered before presenting your final
output.
The output should follow this structure:"Score: <your numerical score>"

Important:
- Use the source and MT segment with respect to the human reference for evaluation.
- Always provide your reasoning before the final rating and justification.
- Output MUST be valid and must follow the structure.
- Use a continuous scale from 1 (worst) to 100 (best)
```