MAS-LitEval : Multi-Agent System for Literary Translation Quality Assessment

Anonymous EMNLP submission

Abstract

Literary translation requires preserving cultural nuances and stylistic elements, which traditional metrics like BLEU and METEOR fail to assess due to their focus on lexical overlap. This oversight neglects the narrative consistency and stylistic fidelity that are crucial for literary works. To address this, we propose **MAS-LitEval**, a multi-agent system using Large Language Models (LLMs) to evaluate translations based on terminology, narrative, and style. We tested **MAS-LitEval** on translations of *The Little Prince* and *A Connecticut Yankee in King Arthur's Court*, generated by various LLMs, and compared it to traditional metrics. **MAS-LitEval** outperformed these metrics, with top models scoring up to 0.890 in capturing literary nuances. This work introduces a scalable, nuanced framework for Translation Quality Assessment (TQA), offering a practical tool for translators and researchers.

1 Introduction

Literary translation is a complex task that goes beyond simple word-for-word conversion. It demands a deep understanding of cultural nuances and the preservation of the author's unique voice through creative adaptation for a new audience. Unlike technical translation, which prioritizes precision and clarity, literary translation requires fidelity to the stylistic essence, emotional resonance, and narrative depth of the source text. This complexity makes evaluation challenging, as the quality of a literary translation is subjective and varies depending on readers' preferences—some favor literal accuracy, while others prioritize capturing the original's spirit (Toral and Way, 2018; Thai et al., 2022).

Traditional evaluation metrics for machine translation, such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005), measure lexical overlap and syntactic similarity. While effective in technical contexts, these metrics struggle with literary texts, overlooking stylistic, discursive, and cultural factors critical to literature (Reiter, 2018). Neural-based metrics like BERTScore (Zhang et al., 2020) and COMET (Rei et al., 2020) enhance semantic analysis, yet they still fail to fully capture aesthetic and cultural nuances. This gap highlights the need for advanced methods tailored to the unique demands of literary translation (Yan et al., 2015; Freitag et al., 2021; Team et al., 2022).

Specialized metrics like Multidimensional Quality Metrics (MQM) (Lommel et al., 2014) and Scalar Quality Metric (SQM) (Blain et al., 2023) attempt to address these shortcomings by evaluating style and fluency alongside accuracy. However, MQM's reliance on human annotation limits its scalability, and SQM lacks the depth required for literary analysis. Large Language Models (LLMs) such as `gpt-4`, `claude`, and `gemini` show promise due to their advanced text generation and comprehension capabilities (Zhang et al., 2025). Nevertheless, no single LLM can comprehensively assess the multifaceted aspects of translation quality—accuracy, fluency, style, and cultural fidelity—necessitating a multi-agent system that leverages their combined strengths (Karpinska and Iyyer, 2023).

Our method introduces a multi-agent system where specialized agents evaluate distinct dimensions of literary translation quality. One agent ensures the consistency of terminology, such as character names; another verifies the alignment of narrative perspective; and a third assesses stylistic fidelity, including tone and rhythm. A coordinator integrates these evaluations into an Overall Translation Quality Score (OTQS), combining quantitative scores with qualitative insights. This approach capitalizes on the strengths of models like `claude` for style and `Llama` for customization, addressing the complex nature of literary TQA.

We evaluated this system on translations of *The Little Prince* and *A Connecticut Yankee in King Arthur's Court*, generated by LLMs including `gpt-4o` (OpenAI et al., 2024), `claude-3.7-sonnet`, `gemini-flash-1.5`, `solar-pro-preview` (Kim et al., 2024), `TowerBase-7B` (Alves et al., 2024), and `Llama-3.1-8B` (Grattafiori et al., 2024). The experimental setup compared our OTQS against traditional metrics (BLEU, METEOR, ROUGE-1, ROUGE-L, WMT-KIWI) using a diverse dataset and a rigorous process to ensure validity.

Results demonstrate that our system outperforms traditional metrics, with top models achieving OTQS scores up to 0.890, capturing nuances like stylistic consistency that BLEU (0.28) misses. Open-source models lagged behind, revealing gaps in their training. These findings confirm our approach's effectiveness in tackling the complexities of literary TQA.

The significance of this work lies in its contributions: (1) a scalable multi-agent TQA framework that enhances literary evaluation, (2) a comparative analysis of LLM performance in translation, and (3) a practical system adaptable for human-in-the-loop refinement. This advances TQA beyond conventional methods, providing a valuable tool for translators and researchers to improve literary translation quality.

2 Method : MAS-LitEval

MAS-LitEval employs specialized LLMs to assess literary translations, with agents focusing on terminology consistency, narrative perspective, and stylistic fidelity.

Overall Architecture. Three agents process the source and translated texts in parallel, with the texts segmented into 4096-token chunks. A coordinator combines their scores and feedback into an Overall Translation Quality Score(OTQS) and a detailed report, ensuring consistency across the entire text.

Roles of Each Agent. The roles of the agents are as follows:

- **Terminology Consistency Agent**: This agent ensures that key terms, such as character names or recurring motifs, remain consistent throughout the translation. Using named entity recognition (NER), it identifies these terms and assigns a score (ranging from 0 to 1) based on their uniformity across the text.

- **Narrative Perspective Consistency Agent**: This agent confirms that the narrative voice (e.g., first-person or omniscient) aligns with the source text across all chunks. An LLM analyzes the segments, assigns a score (ranging from 0 to 1), and flags deviations, such as perspective shifts, to preserve narrative integrity.

- **Stylistic Consistency Agent**: This agent evaluates tone, rhythm, and aesthetic fidelity by comparing stylistic traits between the source and target texts, assigning a fidelity score (ranging from 0 to 1).

Collaboration Mechanism. The coordinator computes the OTQS using a weighted average:

$$\text{OTQS} = w_T \cdot S_T + w_N \cdot S_N + w_S \cdot S_S$$

where S_T, S_N, and S_S represent the scores from the terminology, narrative, and stylistic agents, respectively, and w_T, w_N, and w_S are their corresponding weights. Given the emphasis on preserving the artistic essence of literary works, the weight for stylistic consistency ($w_S = 0.4$) is higher than those for terminology consistency ($w_T = 0.3$) and narrative consistency ($w_N = 0.3$), reflecting its pivotal role in literary translation quality (Yan et al., 2015; Freitag et al., 2021).

Rationale for Multi-Agent Approach. Literary translation quality encompasses multiple dimensions—terminology, narrative, and style—that a single LLM cannot fully evaluate. By employing specialized agents, MAS-LitEval harnesses diverse LLM capabilities, enhancing accuracy and efficiency compared to traditional metrics (Wu et al., 2024). This method ensures consistency is assessed across the entire text, overcoming the limitations of chunk-based evaluations where local consistency might obscure global discrepancies.

Implementation Details. MAS-LitEval is implemented in Python, integrating spaCy for preprocessing and LLMs via APIs. Although texts are segmented into 4096-token chunks for processing, the agents maintain a global context: the Terminology Consistency Agent tracks terms across all chunks, the Narrative Perspective Consistency Agent ensures voice continuity, and the Stylistic Consistency Agent evaluates tone and rhythm holistically.

3 Experiment

We tested MAS-LitEval on translations of excerpts from *The Little Prince* and *A Connecticut Yankee in King Arthur's Court*, generated by a mix of closed-source and open-source LLMs.

Dataset. We selected two works for evaluation: a 5,000-word excerpt from the Korean translation of *The Little Prince* (originally in French) and a 4,000-word excerpt from the Korean translation of *A Connecticut Yankee in King Arthur's Court*

Work	#paras	#sent pairs	Avg. sent/para (src)	Avg. sent/para (tgt)
The Little Prince (Kr-En)	274	1812	6.6	7.0
A Connecticut Yankee in King Arthur's Court (Kr-En)	205	2545	12.2	12.8

Table 1: Dataset Statistics for Specific Works in Korean to English Translation.

(originally in English). These texts were chosen for their stylistic richness and narrative complexity, making them ideal for assessing literary translation nuances. The LLMs generated translations from Korean to English. We also extracted Korean-English parallel data from additional literary works on Project Gutenberg Korea (`http://projectgutenberg.kr/`) and Project Gutenberg (`https://www.gutenberg.org/`), enriching the dataset. Table 1 provides statistics for the specific works used.

Models. Six LLMs were tested: closed-source models (`gpt-4o`, `claude-3.7-sonnet`, `gemini-flash-1.5`, `solar-pro-preview`) and open-source models (`TowerBase-7B`, `Llama-3.1-8B`). These models were chosen for their diverse strengths in language generation and comprehension, enabling a robust performance comparison.

Baselines. MAS-LitEval was compared against BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-1, ROUGE-L (Lin, 2004), and WMT-KIWI (Rei et al., 2023). Human reference translations, sourced from professional translations of the selected works, were used for baseline metrics to ensure a fair comparison.

Evaluation Process. Translations generated by the LLMs were assessed using MAS-LitEval. Texts were segmented into 4096-token chunks, but agents evaluated consistency across all chunks to capture global quality. For instance, the Terminology Consistency Agent assessed term uniformity across the entire text, addressing limitations of chunk-based evaluations where intra-chunk consistency might mask cross-chunk discrepancies. Baseline metrics were calculated against human references, while MAS-LitEval operated reference-free, using only the source and machine-generated translations.

Technical Setup. Experiments were conducted on an NVIDIA A100 GPU. Closed-source models were accessed via APIs, while open-source models were hosted locally with 4-bit quantization to optimize memory usage. The temperature was set to 0.1 to ensure deterministic outputs, guaranteeing reproducibility across runs.

4 Findings

MAS-LitEval evaluated translations of *The Little Prince* and *A Connecticut Yankee in King Arthur's Court*, generated by four closed-source and two open-source models. The results, presented in Table 2, highlight performance differences and our system's ability to detect nuances overlooked by traditional metrics.

Performance of Top Models. `claude-3.7` and `gpt-4o` achieved the highest OTQS scores: 0.890 and 0.875 for *The Little Prince*, and 0.880 and 0.860 for *A Connecticut Yankee in King Arthur's Court*. `claude-3.7-sonnet` excelled in stylistic fidelity (0.93) and narrative consistency (0.91), key aspects of literary quality. For the phrase "On ne voit bien qu'avec le cœur," it translated it as "It is only with the heart that one can see rightly" (stylistic score: 0.92), preserving poetic nuance, while `gpt-4o`'s "One sees clearly only with the heart" (0.87) was less evocative according to agent feedback. In *A Connecticut Yankee in King Arthur's Court*, `claude-3.7-sonnet` maintained the medieval tone across chunks (narrative consistency: 0.90), whereas `gpt-4o` occasionally introduced modern phrasing (0.85).

Comparison of Open-Source and Closed-Source Models. Closed-source models outperformed their open-source counterparts. For *The Little Prince*, `claude-3.7-sonnet` (0.890) and `gpt-4o` (0.875) surpassed `TowerBase-7B` (0.745) and `Llama-3.1-8B` (0.710). Stylistic scores for `TowerBase-7B` (0.70) indicated flatter translations compared to `claude-3.7-sonnet`'s nuanced output (0.92), suggesting limitations in open-source model resources.

Comparison with Baseline Metrics. OTQS showed a strong correlation with WMT-KIWI (0.93) but weaker correlations with BLEU (0.62), METEOR (0.70), ROUGE-1 (0.68), and ROUGE-L (0.65), indicating it captures distinct quality aspects. For *The Little Prince*, `gpt-4o` outperformed

Model	Type	Work	BLEU	METEOR	ROUGE-1	ROUGE-L	WMT-KIWI	OTQS
claude-3.7-sonnet	Closed	LP	0.28	**0.65**	0.55	0.45	**0.87**	**0.890**
		KA	0.27	**0.64**	0.54	0.44	**0.86**	**0.880**
gpt-4o	Closed	LP	**0.30**	0.67	**0.57**	**0.47**	0.85	0.875
		KA	**0.29**	0.66	**0.56**	**0.46**	0.84	0.860
gemini-flash-1.5	Closed	LP	0.25	0.60	0.50	0.40	0.83	0.820
		KA	0.24	0.59	0.49	0.39	0.82	0.810
solar-pro-preview	Closed	LP	0.23	0.58	0.48	0.38	0.81	0.790
		KA	0.22	0.57	0.47	0.37	0.80	0.775
TowerBase-7B	Open	LP	0.20	0.55	0.45	0.35	0.78	0.745
		KA	0.19	0.54	0.44	0.34	0.77	0.730
Llama-3.1-8B	Open	LP	0.18	0.53	0.43	0.33	0.76	0.710
		KA	0.17	0.52	0.42	0.32	0.75	0.695

Table 2: Evaluation Results for the two literary works: LP (*The Little Prince*) and KA (*A Connecticut Yankee in King Arthur's Court*). The highest scores for each metric and work are bolded.

`claude-3.7-sonnet` in BLEU (0.30 vs. 0.28), but OTQS favored the latter (0.890 vs. 0.875) for its stylistic depth. ROUGE-1 and ROUGE-L exhibited similar patterns, missing narrative inconsistencies in models like `TowerBase-7B` (OTQS: 0.745). MAS-LitEval's cross-chunk evaluation identified issues like tone shifts that baselines overlooked, underscoring its advantage in literary quality assessment.

5 Discussion

MAS-LitEval provides a sophisticated framework for literary Translation Quality Assessment (TQA). Below, we explore its strengths, limitations, and implications.

Advantages of the Multi-Agent Approach. MAS-LitEval's multi-dimensional evaluation—covering terminology, narrative, and style—surpasses single-metric methods. For *The Little Prince*, BLEU favored `gpt-4o` (0.30) over `claude-3.7-sonnet` (0.28), but OTQS prioritized `claude-3.7-sonnet` (0.890 vs. 0.875) for its lyrical fidelity. This mirrors human-like judgment, valuing literary essence over lexical overlap. By evaluating consistency across chunks, it detects global issues, such as narrative drift, that chunk-based approaches miss, offering a comprehensive assessment.

Challenges and Refinement Opportunities. Subjectivity in stylistic scoring poses a challenge. The difference between `claude-3.7-sonnet`'s 0.93 and `gpt-4o`'s 0.87 reflects potential LLM biases, which could lead to inconsistency. Averaging scores from multiple LLMs or calibrating with human annotations could improve reliability. Additionally, incorporating domain-specific training or a cultural fidelity agent could address cultural nuances.

Implications for Literary Translation. MAS-LitEval's scalability offers practical benefits. Publishers can use it to pre-screen translations, while educators can leverage its feedback to train translators. Its reference-free design suits literary contexts with multiple valid translations, unlike BLEU or ROUGE, which depend on fixed references. Future enhancements, such as human-in-the-loop integration, could further refine its accuracy, establishing it as a key tool for AI-supported literary TQA.

6 Limitations and Future Works

MAS-LitEval's dataset, restricted to two works, limits its generalizability; expanding to include genres like poetry, drama, and non-fiction is necessary. Stylistic scoring remains subjective and may reflect LLM training biases; averaging scores from multiple LLMs or using standardized rubrics could improve consistency. The absence of human evaluation leaves its alignment with expert judgment unconfirmed; integrating feedback from professional translators or scholars and correlating OTQS with human ratings would validate its reliability. Human input could also refine agent prompts and OTQS weightings. Future efforts should focus on expanding the dataset, incorporating human evaluation, refining stylistic scoring, and addressing cultural concerns to improve MAS-LitEval's reliability and versatility in literary translation quality assessment.

Acknowledgements

References

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. 2023. Findings of the WMT 2023 shared task on quality estimation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore. Association for Computational Linguistics.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-

Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. In *Proceedings of the Eighth Conference on Machine Translation*, pages 419–451, Singapore. Association for Computational Linguistics.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024. Solar 10.7b: Scaling large language models with simple yet effective depth upscaling.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Mqm: Un marc per declarar i descriure mètriques de qualitat de la traducció. *Tradumàtica: traducció i tecnologies de la informació i la comunicació*, (12):455–463.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin

Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up cometkiwi: Unbabel-ist 2023 submission for the quality estimation shared task.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Process-*

ing (EMNLP), pages 2685–2702, Online. Association for Computational Linguistics.

Ehud Reiter. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Antonio Toral and Andy Way. 2018. What level of quality can neural machine translation attain on literary text?

Minghao Wu, Jiahao Xu, and Longyue Wang. 2024. TransAgents: Build your translation company with language agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 131–141, Miami, Florida, USA. Association for Computational Linguistics.

Rongjie Yan, Chih-Hong Cheng, and Yesheng Chai. 2015. Formal consistency checking over specifications in natural languages. In *2015 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pages 1677–1682. IEEE.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025. How good are llms for literary translation, really? literary translation evaluation with humans and llms.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Prompts Used in MAS-LitEval

A.1 Translation Prompt

Translate the following literary text from [source language] to [target language]. Ensure that the translation preserves the original's style, tone, and cultural nuances. Pay special attention to maintaining the narrative voice and literary devices used in the source text.

A.2 Terminology Consistency Agent Prompt

You are an expert in literary translation evaluation. Given a source text in [source language] and its translation in [target language], your task is to ensure that key terms, such as character names, place names, and recurring motifs, are translated consistently throughout the text. Follow these steps:

1. Identify key terms in the source text that appear multiple times.

2. For each key term, check how it is translated in the target text across all occurrences.

3. Calculate a consistency score (0 to 1), where 1 indicates that all occurrences of a term are translated identically, and 0 indicates no consistency.

4. Provide feedback highlighting any inconsistencies, specifying the terms and their varying translations.

Your output should include the consistency score and the detailed feedback.

A.3 Narrative Perspective Consistency Agent Prompt

You are an expert in literary analysis. Given a source text in [source language] and its translation in [target language], your task is to verify that the narrative perspective (e.g., first-person, third-person limited, omniscient) is consistently maintained in the translation. Follow these steps:

1. Determine the narrative perspective of the source text.

2. Analyze the translation to identify its narrative perspective.

3. Compare the two and assess whether the translation accurately reflects the source's perspective.

4. Assign a score (0 to 1) indicating the degree of consistency, where 1 means

perfect alignment, and 0 means complete mismatch.

5. Provide feedback on any deviations, citing specific examples from the text.

Your output should include the consistency score and the detailed feedback.

A.4 Stylistic Consistency Agent Prompt

You are an expert in literary style and translation. Given a source text in [source language] and its translation in [target language], your task is to evaluate how well the translation preserves the stylistic elements of the original, such as tone, rhythm, imagery, and literary devices. Follow these steps:

1. Identify the key stylistic features of the source text.

2. Analyze the translation to see if these features are adequately captured.

3. Assign a score (0 to 1) indicating the level of stylistic fidelity, where 1 means the translation perfectly preserves the style, and 0 means it completely fails to do so.

4. Provide feedback with specific examples where the translation succeeds or falls short in maintaining the style.

Your output should include the fidelity score and the detailed feedback.