CONFI-Lingual: A Confidence Evaluation Approach for Machine Translation

Daniel Chechelnitsky
Carnegie Mellon University
dchechel@andrew.cmu.edu

Gayathri Ganesh Lakshmy
Carnegie Mellon University
gganeshl@andrew.cmu.edu

Kaitlyn Zhou
Stanford University
katezhou@stanford.edu

Chrysoula Zerva
Instituto Superior Técnico
chrysoula.zerva@tecnico.ulisboa.pt

Maarten Sap
Carnegie Mellon University
msap2@andrew.cmu.edu

Abstract

In this study, we examine how large language models (LLMs) translate confidence expressions. We purpose a framework for measuring confidence, which we then use to assess and evaluate translations of confidence expressions. As this is also a machine translation (MT) evaluation problem, we are also curious whether LLMs perform better with certain high-resource languages. To do so, we trained a regression model to score 3570 English back-translations of confidence sentences across 17 different pivot languages, ranging in resource-level. Our results show that there is in fact a significant increase in confidence scores with respect to translation. We did, also, find that this wasn't true for some individual languages (Chinese in high-resource, Quechua for low-resource). We also found that these changes in confidence scores that are not correlated with translation quality. Our findings therefore demonstrate how LLMs being used for MT tasks are also producing overconfident generations, therefore increasing for the risk of user over-reliance on overconfident MT outputs.

1 Introduction

Large language models (LLMs) have been shown to generate overconfident responses, which have resulted in higher reliance by users on these LLM responses (Zhou et al., 2024; 2023). Therefore, it is crucial that the confidence in these models' outputs is accurate and calibrated with the confidence expressed in the source text to prevent over-reliance, since recent work has shown the harms that come about from users' over-reliance on LLMs (Yang et al., 2024; Kim et al., 2025).

With LLMs being used for machine translation (MT) (Zhu et al., 2024), this raises the question of whether these overconfidence tendencies are also affecting the translation outputs. This is further important to consider for low-resource languages contexts, where there are already widely-studied performance disparities (Kumar et al., 2021; Merx et al., 2024). These recurring issues with low-resource translation stem from the lack of widely available resources in these languages, which in turn prevents the development of specified translation models and translation evaluation benchmarks for these languages (Thakur, 2024).

Therefore, for this work, we critically examine whether confidence shift occurs in LLMs doing MT tasks, and in particular seeing if this is impacted by resource-level of language. To examine this we introduce a framework that maps confidence markers in English sentences onto a percentage of user reliance from Zhou et al. (2024)'s work, what was obtained by measuring what percentage of the users are willing to accept an LLM response using the specified confidence phrase.

Motivated by work in multilingual MT evaluation of high and low-resource languages, back-translations, i.e., where a text is translated from a language A to B and then back to A,

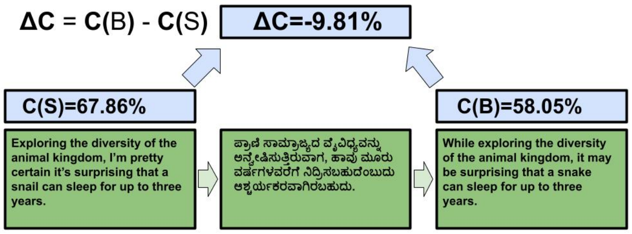
ΔC = C(B) - C(S)
ΔC=-9.81%
C(S)=67.86%
C(B)=58.05%
Exploring the diversity of the animal kingdom, I'm pretty certain it's surprising that a snail can sleep for up to three years.
ಪ್ರಾಣಿ ಸಾಮ್ರಾಜ್ಯದ ವೈವಿಧ್ಯವನ್ನು ಅನ್ವೇಷಿಸುತ್ತಿರುವಾಗ, ಹಾವು ಮೂರು ವರ್ಷಗಳವರೆಗೆ ನಿದ್ರಿಸಬಹುದೆಂಬುದು ಆಶ್ಚರ್ಯಕರವಾಗಿರಬಹುದು.
While exploring the diversity of the animal kingdom, it may be surprising that a snake can sleep for up to three years.

Figure 1: Overview of metrics: $C(S)$ and $C(B)$ represent the confidence scores, ΔC represents the difference in confidence scores.

are commonly used in multilingual research because of their ability to capture linguistic characteristics resulting from translation (McNamee & Duh, 2023).

For our study, we generate round-trip translations (Figure 1), where the pivot translation is from English into the pivot language seen in Table 1, and the back-translation is from the pivot language back into English.

Overall, we perform 3570 round translations with 17 languages ranging from high-resource (Arabic, Mandarin Chinese), to medium (Telugu, Finnish), to low-resource (Marathi, Quechua). Our findings across these languages show that there is a significant change in confidence scores in the positive direction, indicating that LLMs' overconfidence indeed affect their translations. Surprisingly, this remains true for both high- and low-resource languages. Additionally, we found that across all the languages we observed, there is not a correlation with quality.

2 Related Work

Measuring Confidence Current work in measuring the expressed confidence of epistemic markers, or confidence phrases, from LLM outputs has largely focused on general-purpose monolingual language models (Zhou et al., 2023; Mielke et al., 2022). And while there does exist work that has looked at confidence in MT, this work has largely looked at inference-time model confidence calibration (Wang et al., 2020), and is more focused on translation accuracy and not on the confidence expressed in the source text. The concerns of using purely computational metrics instead of considering human-AI interaction perspectives have also been raised in MT (Liebling et al., 2022).

Additionally, work that has measured human perceived confidence specifically, as was done by Zhou et al. (2024), has been strictly categorical. This leaves room to show more fine-grained differences in confidence by use of a numerical score metric. This can be seen in studies that measure MT tasks quantitatively for other metrics like translation quality (Moghe et al., 2023).

LLMs for MT Additionally, the use of LLMs for MT evaluation problems has been increasingly popular in recent years (Zeng et al., 2023; Tan et al., 2024).

However, the problem of LLM outputs being largely overconfident still persists (Zhou et al., 2024). Recent work by Zhou et al. (2025) similarly demonstrates that this issue with LLM outputs being overconfident is directly associated with user associated over-reliance. Given these findings, it is unclear whether current fine-tuning approaches to LLMs are directly targeting these and similar issues.

3 Framework & Methodology

MT for Low-Resource Languages Here, we present our framework which uses back-translations and confidence scoring to measure this confidence change due to LLMs. This

Resource Level	Language	Syntactic Distance	Genetic Distance	Writing System
High-Resource	Arabic	0.349	1.000	Arabic
	Mandarin Chinese	0.287	1.000	Chinese
	Russian	0.188	0.833	Cyrillic
	Portuguese	0.158	0.900	Latin
	Japanese	0.499	1.000	Hiragana, Katakana, Kanji
Medium-Resource	Finnish	0.289	1.000	Latin
	Telugu	0.479	1.000	Telugu
	Latvian	0.246	0.808	Latin
	Greek	0.217	0.851	Greek
	Kazakh	0.5523	1.000	Latin, Cyrillic
	Indonesian	0.274	1.000	Latin, Arabic
Low-Resource	Marathi	0.435	0.864	Devanagari, Modi
	Kannada	0.466	1.000	Kannada
	Quechua	0.486	1.000	Latin
	Belarusian	0.214	0.833	Cyrillic
	Nepali	0.474	0.833	Devanagari
	Scottish Gaelic	0.243	0.874	Latin

Table 1: Breakdown of 17 languages analyzed, with URIEL Distance Scores calculated in relation to English.

process requires the pivot translation and back-translation LLM-translated epistemic markers: A → B and B → A.

Using our confidence scorer, we compute C(S) and C(B), i.e. the confidence scores for the source and back-translated sentences respectively. Lastly, we then compute $\Delta C = C(S) - C(B)$ for each round-trip translation in the test set (see Figure 1).

Confidence Scoring To build our confidence scorer we use a logistic regression using Support Vector Regression (SVR)(Drucker et al., 1996) and pre-trained word embeddings from E5-Mistral-7B (Wang et al., 2023). This SVR is then given user sentence completions of phrases for training. These are generated additionally by transforming Zhou et al. (2024)'s paired dataset with epistemic markers into full sentences (see Appendix A).

The regressor is then trained using behavioral reliance percentages, where the data is put into a an 80% training and 20% test split split. We also ensure the split does not have any overlapping confidence markers between the training and test sets.

For this regressor, we observe low error rates of ($mae = 0.314$, $mse = 0.124$), and a Pearson's correlation of ($r = 0.907$, $p < 0.001$). These values indicate that the regressor has good performance for scoring confidence phrases, and therefore we decided to utilize it as our confidence scorer.

Selected Pivot Languages For pivot languages, we selected a diverse set of languages spanning different language families, writing systems, and resource levels (Table 1). This allowed us to observe a snapshot of the larger language diversity and assess how writing system, genetic distance, and syntactic distance (as measured by URIEL scores (Littell et al., 2017)) influence translation. Language resource levels (high, medium, low) were determined based on existing NLP literature. Mandarin Chinese, Portuguese, Japanese, Arabic, and Russian were classified as high-resource (Wang et al., 2025; Nicholas & Bhatia, 2023). Greek, Kazakh, and Latvian have received moderate NLP attention (Loukas et al., 2025; Maxutov et al., 2024; Paikens et al., 2022). Finnish and Indonesian are higher-resourced than Quechua (Chen & Fazio, 2021), while Telugu is better resourced than Kannada and Nepali (Mukherjee et al., 2025). Lastly: Belarusian, Marathi, and Scottish Gaelic are considered low-resource (Kumar et al., 2021; Jain et al., 2021; Lamb et al., 2025).

We first translate the source sentences into the pivot language, followed by a translation back into English for the back translation. The pivot translation prompt and back-translation prompts can be seen in Appendix B. The translations of the texts were generated using a state-of-the-art LLM: GPT-4o (Hurst et al., 2024).

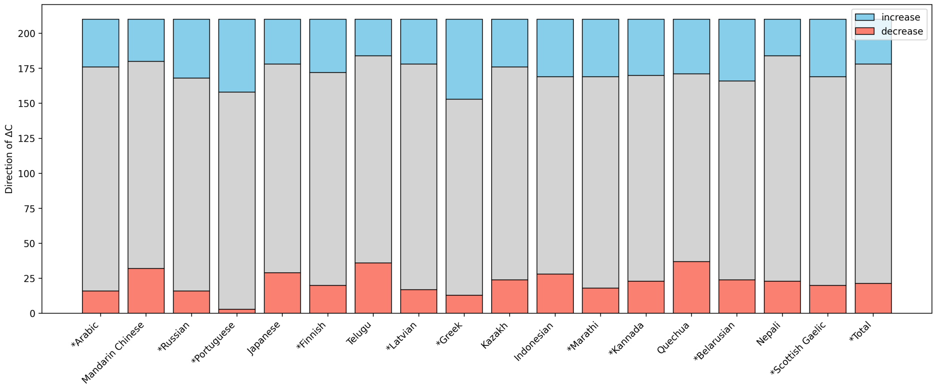

Figure 2: Distribution of Confidence Change (ΔC) with respect to Language. Here, the significant increases and decreases thresholds are determined by the Standard Deviations of each language's ΔC distribution. An * denotes significance.

By compiling these translation and confidence scores for each of the 17 languages, we assemble a dataset that contains both the pivot translations and back-translations (as seen in Figure 1). We then compute the differences in confidence scores ΔC, and make holistic as well as language specific observations that answer. Finally, we observe if or confidence scorer was associated with a similar MT metric for translation quality.

4 Results

Research Questions For each of these languages, our work aims to look into how epistemic markers (EMs), i.e. expressions of confidence, change when translated change the confidence. Hence, we aim to address the following questions:

- **RQ1**: How does confidence of a phrase change as a result of MT and how does pivot language choice impact this change?
- **RQ2**: Is the measurement of expressed confidence that we use related to translation quality?

4.1 RQ1: MT Effect on Confidence Change

For **RQ1**, we observed that while confidence generally does not change much during translation. If it does change significantly it is much more often to be for an increase, or for a higher user confidence score.

To examine how often confidence increases, is preserved, or decreases, we categorize each back-translation using a χ^2 test to compare the number of instances in the threshold, which was set as the standard deviation ($sdv = 0.0321$) of the total distribution. This was done to see whether there is a higher amount of phrases becoming more or less confident. For our entire list of 3570 ΔCs observed, 74.6% of them did not significantly change, 15.2% became increasingly more confident, and 10.2% became significantly less confident. Across the ΔCs that significantly changed, we observed a chi-squared value of ($\chi^2 = 39.984$, $p < 0.005$).

Overall, the confidence is more likely to remain similar to the original sentence, but when it is changed, it is more likely to be increased than decreased. An example of one of these significant changes can be seen in Figure 1 with a Kannada example where 'I'm pretty certain it's' has a significant decrease in confidence and becomes a less confident 'it may be surprising that'.

In addition, we also look at each individual pivot language (Figure 2): estimating a new standard deviation threshold for each within-language distribution. We re-categorize based on language specific confidence standard deviation and ultimately found that 7 of our 17 languages (Mandarin Chinese, Japanese, Telugu, Kazakh, Indonesian, Quechua, and Nepali) did not have a significant likelihood of choosing ($p > 0.05$) a response with an positive change in confidence score. While these languages appear across both high, middle, and low-resource, we do observe that 6 of them (all besides Nepali) have the greatest Genetic Distance from English, and all 7 have at least a Syntactic Distance of 0.25 (see Table 1).

4.2 RQ2: Similarity between MT Confidence Change and Translation Quality

Finally, to explore if ΔC is aligned with translation quality for **RQ2**, we also calculate translation quality for the forward translations using CometKiwi, which is available for all languages we are looking except for Quechua (Rei et al., 2022). This is to observe how our confidence framework measurement is associated with a metric for translation quality, and to see if the ΔC change in confidence score is also a translation quality metric.

For this specific study we look at 6 languages, two from each resource category (high, medium, and low). For each of these pairs, one language observed a significant increase in positive phrases, while one did not. Here, we found that there is all-round no significant association between confidence scores and translation quality of phrases (Table 2).

Language	**Confidence Diff (r)**	**Absolute Diff (r)**
Arabic	-0.020 (n.s.)	0.091 (n.s.)
Japanese	-0.022 (n.s.)	0.115 (n.s.)
Kazakh	-0.060 (n.s.)	0.071 (n.s.)
Latvian	-0.053 (n.s.)	0.050 (n.s.)
Nepali	0.079 (n.s.)	0.099 (n.s.)
Scottish Gaelic	-0.045 (n.s.)	0.015 (n.s.)

Table 2: Pearson correlation (r) between CometKiwi scores and confidence scores. Significance *: $p < 0.05$, No Significance (n.s.): $p > 0.05$.

5 Conclusion

In summary, we look at 17 languages varying in resource level and approximation to English to see if MT impacts how perceived confidence of phrases changes as a result of the machine translation (MT) process. We find that overall there is a significant trend of increasingly more confident outputs.

While these findings are inconclusive on what specific languages have increased confidence outputs, it does go to show that this is a phenomenon observed in languages of all resource, and more likely to occur in languages with less similarities to the source language (in this case English).

Therefore, our work demonstrates how overconfidence and the risk of over-reliance on LLMs extends to translation. While this we didn't find which specific linguistic attributes related to this change, we do see how this is a pattern across both high and low-resource languages in NLP.

Limitations

These findings do, however, leave room for more exploration in the field of LLM confidence evaluation.

In this study, we only observed GPT-4o outputs. One option would be to look at more open and closed LLM models, like Qwen, as well as sequence-to-sequence models, like T5, optimized for translation and see if other models Additionally, we would look into more languages beyond just these 17, as they only offer a small glimpse and do not account for all the language variation available.

References

William Chen and Brett Fazio. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu (eds.), *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pp. 20–31, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-loresmt.3/.

Harris Drucker, Christopher J Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. Support vector regression machines. *Advances in neural information processing systems*, 9, 1996.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Aditya Jain, Shivam Mhaskar, and Pushpak Bhattacharyya. Evaluating the performance of back-translation for low resource English-Marathi language pair: CFILT-IITBombay @ LoResMT 2021. In John Ortega, Atul Kr. Ojha, Katharina Kann, and Chao-Hong Liu (eds.), *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pp. 158–162, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-loresmt.17/.

Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, and Olga Russakovsky. Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713941. doi: 10.1145/3706598.3714020. URL https://doi.org/10.1145/3706598.3714020.

Sachin Kumar, Antonios Anastasopoulos, Shuly Wintner, and Yulia Tsvetkov. Machine translation into low-resource language varieties. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 110–121, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.16. URL https://aclanthology.org/2021.acl-short.16/.

William Lamb, Dongge Han, Ondrej Klejch, Beatrice Alex, and Peter Bell. Synthesising a corpus of Gaelic traditional narrative with cross-lingual text expansion. In Brian Davis, Theodorus Fransen, Elaine Uí Dhonnchadha, and Abigail Walsh (eds.), *Proceedings of the 5th Celtic Language Technology Workshop*, pp. 12–26, Abu Dhabi [Virtual Workshop], January 2025. International Committee on Computational Linguistics. URL https://aclanthology.org/2025.cltw-1.2/.

Daniel Liebling, Katherine Heller, Samantha Robertson, and Wesley Deng. Opportunities for human-centered evaluation of machine translation systems. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 229–240, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.17. URL https://aclanthology.org/2022.findings-naacl.17.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Mirella Lapata, Phil Blunsom, and Alexander Koller (eds.), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-2002/.

Lefteris Loukas, Nikolaos Smyrnioudis, Chrysa Dikonomaki, Spiros Barbakos, Anastasios Toumazatos, John Koutsikakis, Manolis Kyriakakis, Mary Georgiou, Stavros Vassos, John

Pavlopoulos, and Ion Androutsopoulos. GR-NLP-TOOLKIT: An open-source NLP toolkit for Modern Greek. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, Steven Schockaert, Brodie Mather, and Mark Dras (eds.), *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pp. 174–182, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-demos.17/.

Akylbek Maxutov, Ayan Myrzakhmet, and Pavel Braslavski. Do LLMs speak Kazakh? a pilot evaluation of seven models. In Duygu Ataman, Mehmet Oguz Derin, Sardana Ivanova, Abdullatif Köksal, Jonne Sälevä, and Deniz Zeyrek (eds.), *Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024)*, pp. 81–91, Bangkok, Thailand and Online, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.sigturk-1.8/.

Paul McNamee and Kevin Duh. An extensive exploration of back-translation in 60 languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8166–8183, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.518. URL https://aclanthology.org/2023.findings-acl.518/.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In Atul Kr. Ojha, Sina Ahmadi, Silvie Cinková, Theodorus Fransen, Chao-Hong Liu, and John P. McCrae (eds.), *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pp. 1–11, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.eurali-1.1/.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. doi: 10.1162/tacl_a_00494. URL https://aclanthology.org/2022.tacl-1.50/.

Nikita Moghe, Tom Sherborne, Mark Steedman, and Alexandra Birch. Extrinsic evaluation of machine translation metrics. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13060–13078, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.730. URL https://aclanthology.org/2023.acl-long.730/.

Ananya Mukherjee, Saumitra Yadav, and Manish Shrivastava. Why should only high-resource-languages have all the fun? pivot based evaluation in low resource setting. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 4779–4788, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.320/.

Gabriel Nicholas and Aliya Bhatia. Lost in translation: large language models in non-english content analysis. *arXiv preprint arXiv:2306.07377*, 2023.

Peteris Paikens, Mikus Grasmanis, Agute Klints, Ilze Lokmane, Lauma Pretkalniņa, Laura Rituma, Madara Stāde, and Laine Strankale. Towards Latvian WordNet. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 2808–2815, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.300/.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen

Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60/.

Shaomu Tan, Di Wu, David Stap, Seth Aycock, and Christof Monz. UvA-MT's participation in the WMT24 general translation shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 176–184, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.11. URL https://aclanthology.org/2024.wmt-1.11/.

Madhavendra Thakur. Towards neural no-resource language translation: A comparative evaluation of approaches. *arXiv preprint arXiv:2412.20584*, 2024.

Chenglong Wang, Haoyu Tang, Xiyuan Yang, Yueqi Xie, Jina Suh, Sunayana Sitaram, Junming Huang, Yu Xie, Zhaoya Gong, Xing Xie, et al. Uncovering inequalities in new knowledge learning by large language models across different languages. *arXiv preprint arXiv:2503.04064*, 2025.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*, 2023.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*, 2020.

Xikang Yang, Xuehai Tang, Jizhong Han, and Songlin Hu. The dark side of trust: Authority citation-driven jailbreak attacks on large language models. *arXiv preprint arXiv:2411.11407*, 2024.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. Improving machine translation with large language models: A preliminary study with cooperative decoding. *arXiv preprint arXiv:2311.02851*, 2023.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori Hashimoto. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5506–5524, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.335. URL https://aclanthology.org/2023.emnlp-main.335/.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*, 2024.

Kaitlyn Zhou, Jena D. Hwang, Xiang Ren, Nouha Dziri, Dan Jurafsky, and Maarten Sap. REL-A.I.: An interaction-centered approach to measuring human-LM reliance. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 11148–11167, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.556. URL https://aclanthology.org/2025.naacl-long.556/.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2765–2781, Mexico

City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.176. URL https://aclanthology.org/2024.findings-naacl.176/.

A Completion Prompt

Can you convert the following 10 phrases into full sentences containing a fun fact that aren't too geographically or culturally specific. For approximately half the completions, add context before the phrase as well. But do not add any additional certainty or uncertainty expressions. Follow the examples below.

Examples:

1. I'm not sure, maybe it's → I'm not sure, maybe it's the case that honey actually never spoils.
2. I'm 100% certain that the answer is → When it comes to fruit classification, I'm 100% certain that bananas are in fact berries.
3. It's likely that the answer is → I was thinking about which animal has three hearts, and it's likely that the answer is the octopus.
4. ...

Do this for the following phrases:

1. {input phrase}
2. {input phrase}
3. {input phrase}
4. {input phrase}
5. {input phrase}
6. {input phrase}
7. {input phrase}
8. {input phrase}
9. {input phrase}
10. {input phrase}

B Translation Prompts

Please translate the following text into {language}: {input phrase} Please only respond with the translation, no additional greetings, comments, or explanations.

Please translate the following {language} text into English: {input phrase} Please only respond with the translation, no additional greetings, comments, or explanations.