Iterative Translation Refinement with Large Language Models

Anonymous ARR submission

Abstract

This paper argues that benefiting from vast pre-training data, large language models offer a means to improve translation fluency. We propose iterative refinement prompting, which is infeasible for conventional encoder-decoder models. In our experiments, multi-pass querying reduces string-based metric scores, but neural metrics suggest comparable or improved quality. Human evaluations indicate better fluency and naturalness compared to initial translations and even human references, all while maintaining quality. Ablation studies underscore the importance of anchoring the refinement to the source and a reasonable seed translation for quality considerations. We also discuss the challenges in evaluation and relation to human performance and translationese.

1 Introduction

Large language models (LLMs), e.g. generative pre-trained Transformers (GPT), have made notable advancements in natural language processing (Radford et al., 2019; Brown et al., 2020; Kaplan et al., 2020; Ouyang et al., 2022). In machine translation (MT), where the convention is to use an encoder-decoder architecture to deal with source and target sentences respectively (Bahdanau et al., 2015; Vaswani et al., 2017), recent papers have examined the feasibility of LLM prompting (Vilar et al., 2023; Zhang et al., 2023; Hendy et al., 2023).

With autoregressive decoding, MT yields output in a single attempt, and so does post-editing. Rather, a human translator can read and edit translations repeatedly. We explore such an iterative refinement process with LLMs, where the proposed method simply feeds a source-translation pair into an LLM for an improved translation in multiple rounds. It can be applied to an initial translation from any model. Our approach offers two insights from a fluency and naturalness perspective: 1) LLMs are pre-trained on natural texts that are orders of magnitude larger than traditional MT data, and 2) the method does not require complicated prompt engineering, yet allows for iterative and arbitrary rephrasing compared to automatic post-editing, which is limited to token-level error correction without style editing (Ive et al., 2020).

Empirical results show that the refinement process introduces significant textual changes reflected by the drop in BLEU and chrF++, but attains similar or higher COMET scores compared to initial translations. Native speakers prefer refined outputs in terms of fluency and naturalness when compared with GPT translations and even human references. Referenced-based human evaluation confirms that such gains are made without sacrificing general quality. As corroborated by recent works, automatic metrics like BLEU and COMET can move in opposite directions (Freitag et al., 2019, 2022). Our work makes an interesting contribution towards translation naturalness which can enhance utility as perceived by the target language users.

2 Methodology

Having an input source sentence x and an optimizable model θ_{mt}, the process to obtain a translation y can be modelled as $y = \mathrm{argmax}_y P(y|x, \theta_{mt})$. Next, an automatic post-editor θ_{ape} creates a refined translation y' through $y' = \mathrm{argmax}_{y'} P(y'|x, y, \theta_{ape})$. Conventional translation or automatic post-editing models are trained on (x, y) or (x, y, y') data pairs.

Extending prior work on LLM prompting, our study uses zero-shot prompting by affixing a task description to form a prompt p and querying an LLM θ_{LLM} to elicit a response (Brown et al., 2020). We introduce five prompts in our study:

1. *Translate*: it queries for a translation of a source input, extending the translation process with a prompt p: $y = \mathrm{argmax}_y P(y|p, x, \theta_{LLM})$. This is vanilla LLM prompting for MT.

Mode	Prompt
Translate	`Source: ${source}` `Please give me a translation in ${lang} without any explanation.`

Refine	`Source: ${source}`
	`Translation: ${prev_translation}`
	`Please give me a better ${lang} translation without any explanation.`

*Refine*Contrast	`Source: ${source}`
	`Bad` `translation: ${prev_translation}`
	`Please give me a better ${lang} translation without any explanation.`

*Refine*Random	`Source: ${source}`
	`Bad` `translation: ${random_target}` *if first-round, else* `${prev_translation}`
	`Please give me a better ${lang} translation without any explanation.`

Paraphrase	`Sentence: ${prev_translation}` `Please give me a paraphrase in ${lang} without any explanation.`

Table 1: Prompts used in our work, where `${variable}` is substituted with its corresponding content.

2. *Refine*: similar to post-editing, the LLM is given the source sentence and the previous translation to produce a better translation $y' = \text{argmax}_{y'} P(y'|p, x, y, \theta_{LLM})$.
3. *Refine*$_{\text{Contrast}}$: as a contrasting prompt to the above, we insert the word "bad" to hint that the previously translated text is unwanted, regardless of its actual quality.
4. *Refine*$_{\text{Random}}$: same prompt as *Refine*$_{\text{Contrast}}$, but in the first iteration, a random sentence is fed instead of a translation to imitate a genuinely "bad translation".
5. *Paraphrase*: to ablate the translation process, we prompt to rephrase a translation without feeding the source sentence x: $y'' = \text{argmax}_{y''} P(y''|p, y, \theta_{LLM})$.

We propose to iteratively call the refinement prompts, where the source stays the same but the previous translation is updated with the latest, to understand how quality changes. Through ablative prompts, we can analyse to what degree the source input and seed translations are helpful. The exact prompt texts are displayed in Table 1.

3 Experiments

3.1 Data and model details

We experiment with language pairs from the translation tasks hosted at WMT 2021 and 2022 (Farhad et al., 2021; Kocmi et al., 2022). In total, we tested seven translation directions: English↔{German, Chinese}, German→French, English→Japanese, and Ukrainian→Czech. We directly benchmark on the test sets, and in situations where multiple references are available, we use human reference "A" released by the WMT organizers.

We experiment with GPT-3.5, a powerful API from OpenAI that can be accessed by all users.[1] As the API is very slow to query, we randomly sample 200 instances from the official test set to form our own test. In the refinement and paraphrase experiments, we use the response from the *Translate* query as the base translation to be improved upon. In experiments later on, we also test with seed translations from encoder-decoder models. We do not keep the query history so as to prevent an LLM from seeing that the previous translation is produced by itself. Overall, translation refinement is iterated four times maximum.

3.2 Evaluation setup

We consider four automatic metrics: string-based BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017), as well as embedding-based COMET$_{\text{DA}}$ and COMET$_{\text{QE}}$ (Rei et al., 2020). The difference between DA and QE versions is that COMET$_{\text{DA}}$ requires a source, a translation, and a human reference, whereas COMET$_{\text{QE}}$ is reference-free.[2] These metrics are widely used to measure overall translation quality, yet Freitag et al. (2020) hint that too high a single-reference BLEU cannot imply high quality; we see it as an indicator of text variations from the reference. Further, we rely on COMET$_{\text{QE}}$ for two reasons: 1) we intend to compare with WMT references which could be sub-optimal; 2) it is reference-free, so it also serves as a "stopping criterion" amid iterations if performance does not improve. We report BLEU and COMET$_{\text{QE}}$ scores in the main content and attach chrF++ and COMET$_{\text{DA}}$ in Appendix A.

[1] We accessed a version of `gpt-3.5-turbo` with training data up to Sep 2021, so it should not have seen WMT 2021 or 2022 test references. Nevertheless, our findings are mostly drawn from reference-free metrics and human evaluation.

[2] BLEU and chrF++ from `sacrebleu` (Post, 2018). For COMET, we use `wmt-2022-da` and `wmt-2021-qe-da` respectively. We document details in Appendix E.

	WMT21 de→en		WMT21 en→de		WMT21 zh→en		WMT21 en→zh		WMT22 de→fr		WMT22 en→ja		WMT22 uk→cs	
	BLEU	COMET_{QE}	BLEU	COMET_{QE}	BLEU	COMET_{QE}	BLEU	COMET_{QE}	BLEU	COMET_{QE}	BLEU	COMET_{QE}	BLEU	COMET_{QE}
$\text{Reference}_{\text{A}}$	-	*.0919*	-	*.1127*	-	*.0708*	-	***.0956***	-	*.0772*	-	***.1345***	-	***.1273***
Translate	**30.90**	.1128	**25.39**	.1083	**25.64**	.0867	**29.28**	.0761	**36.25**	.0807	**23.00**	.1255	**29.91**	.1173
Refine	23.14	.1116	22.35	**.1153**	20.26	.0921	28.26	.0870	32.47	**.0851**	22.63	**.1305**	28.60	**.1183**
$\text{Refine}_{\text{Contrast}}$	22.88	**.1162**	22.54	.0929	24.81	**.1132**	**29.28**	**.0881**	33.12	.0805	22.82	.1282	28.90	.1151
$\text{Refine}_{\text{Random}}$	18.83	.0770	19.36	.0832	24.24	.1022	25.71	.0763	-	-	-	-	-	-
Paraphrase	11.01	.0919	13.60	.1006	12.76	.0885	21.95	.0716	16.06	.0682	17.69	.1086	13.59	.0969

Table 2: Automatic scores of different strategies on translation directions from WMT 2021 and 2022 news translation.

3.3 Refinement results

WMT21 We first experiment with en↔de and en↔zh from WMT21, and display results in Table 2. For iterative experiments, the best iteration is picked according to COMET_{QE}. We observe that the refined translations record a drastic drop in string-based metrics compared to initial translations, indicating lexical and structural variations. In terms of COMET_{QE}, refined outputs surpass all initial GPT translations, with substantial improvement for into-English directions. The ablative *Paraphrase* method sees a decline in all metrics, suggesting the importance of feeding the input as an anchor during iterations to prevent semantic drift.

To investigate the behaviour of different refinement strategies, we plot BLEU, COMET_{DA}, and COMET_{QE} at different iterations in Appendix C Figure 2. We see that *Refine* and $\textit{Refine}_{\text{Contrast}}$ usually attain their best after the first iteration, but in almost all *Paraphrase* experiments, scores decrease monotonically, indicating that semantics drift away as paraphrasing iterates. Moreover, $\textit{Refine}_{\text{Random}}$ results start low, gradually catch up, but never reach as high as *Refine* or $\textit{Refine}_{\text{Contrast}}$. This means that iterative refinement is indeed useful in fixing translations, but starting with a reasonable translation is also crucial for obtaining a strong result.

WMT22 For non-English translation, we pick three directions from WMT22. Since $\textit{Refine}_{\text{Random}}$ results are not desirable for WMT21, we omit experiments with this. We find that *Refine* works best, obtaining higher COMET_{QE} than vanilla translations and $\textit{Refine}_{\text{Contrast}}$. Also, the reduction in string-based scores becomes less obvious, which might be attributed to seed GPT translations in lesser-resourced languages being lower in quality.

Online, encoder-decoder systems, and human translations In addition to translation refinement from GPT-3.5 itself, we also apply our refinement calls to outputs from conventional MT systems and human translators. These translations can represent genuine errors, if any, introduced during the translation process. We experiment with seven different submissions in the WMT 2021 German-to-English news translation track as a starting point. Due to the space constraint, we introduce the systems and report automatic metric scores in Appendix B.

A pattern similar to previous GPT refinement is noticed: for five out of seven WMT entries, the refinement strategy reaches a higher COMET_{QE} score, surprisingly, with up to one-third drop in BLEU. $\textit{Refine}_{\text{Contrast}}$ in all but one system surpass *Refine*, and without the initial translation, *Paraphrase* iterations record the lowest scores compared to the original submissions and refinements.

4 Human Evaluation

String-based and neural scores are observed to vary in opposite directions, which may suggest volatile changes in texts. We set up human evaluations to measure two characteristics in the refined translations: text naturalness and overall quality.

4.1 Fluency and naturalness

We mimic the human evaluation of fluency in (Lembersky et al., 2012, p. 819). Native speakers of the target language are with two translations but without the source sentence; then we ask "`Please choose the translation that is more fluent, natural, and reflecting better use of ${language}`". The evaluators can select one of the two translations, or a "tie" if they consider both equally (un)natural. We conduct such pairwise evaluation to compare the first-round output from $\textit{Refine}_{\text{Contrast}}$ against human references, as well as against *Translate* separately.

We evaluate 50 samples from en↔de and en↔zh experiments in Section 3.3, and report in Figure 1 (left). Native speakers prefer $\textit{Refine}_{\text{Contrast}}$ to vanilla *Translate* in all four directions, and even favour $\textit{Refine}_{\text{Contrast}}$ over human references when translating into English. It demonstrates that our simple strategy enhances the naturalness of GPT

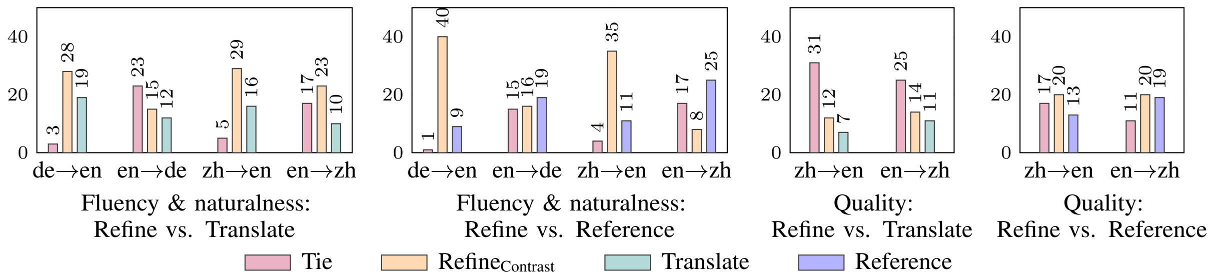

Figure 1: Human preferences on fluency and naturalness (source-free, left) and overall quality (source-based, right).

outputs and that WMT human references could be less favourable than GPT outputs in some cases.

4.2 Overall quality

We also evaluate for general quality as a safeguard. In this setup, a source sentence and two translations are given to an evaluator who is fluent in both languages. They are asked to pick the translation with better quality or indicate a tie. We only evaluated two translation directions, English to and from Chinese, due to the limited availability of bilingual speakers. Similar to the previous evaluation, we compare *Refine*$_{\text{Contrast}}$ against human references, as well as *Refine*$_{\text{Contrast}}$ against *Translate* separately.

We report evaluator preferences in Figure 1 (right). It shows that GPT *Refine* attains slightly better performance in zh→en and similar performance in en→zh when compared with human references. On the other hand, it is more favourable than GPT *Translate* in terms of human judgements. Combining evaluation outcomes, we conclude that the refinement strategy could improve the target-side naturalness without undermining general quality.

5 Discussions

5.1 Automatic evauation

In Appendix D Table 5 we show outputs from different strategies for a single source input, where a native speaker marked preference for *Refine*$_{\text{Contrast}}$. It illustrates that the word choice is diverse for both directions and specifically for Chinese→English, there are substantial structural changes. The huge variety in expressions across translations can result in low BLEU with respect to human references, but without much change in meaning, for instance, as in Table 2 where BLEU can decline up to one-third, but neural metric scores change little. In the field of MT, a leap in BLEU is usually associated with performance improvement; however, in our case, a drop cannot be simply interpreted as performance degradation. This can be attributed to the lexical and structural diversity in the refined translations.

5.2 Human performance

A human translator is deemed to be fluent in their native language, which intuitively is difficult for a model to compete with. We offer two explanations. First, the WMT references might have been created by translators with varying expertise, which may not represent upper-bound human performance, especially when compared with advanced LLMs. More importantly, translations can exhibit awkwardness in word and syntax choices, potentially due to source language interference or "shining through" (Gellerstam, 1986; Teich, 2003).

5.3 Relation to translationese

On the target end, translations might be more explicit, language-normalized, and simpler (Baker, 1996; Koppel and Ordan, 2011). On a broad scope, translationese is regarded as the distinct features in translations to include both the source and target influences. Although MT normally learns from human translation data, researchers found that human and machine translation patterns do not fully overlap (Bizzoni et al., 2020). From a narrow aspect, our method relates to machine translationese mitigation in terms of reducing unnaturalness and literalness, instead of focusing on state-of-the-art metric scores. It may be viable to create diverse translations as shown in huge BLEU changes. Measuring these using automatic metrics at the moment is challenging. Finally, the concept of iterative refinement or post-editing is not new. In addition to the key related work already introduced, we detail other works in Appendix F.

6 Conclusion and Future Work

We presented a simple way to leverage an LLM for translation refinement, which greatly helps fluency and naturalness. It is shown that our method maintains translation quality and introduces lexical and structural changes, especially for high-resource into-English translation. Future work can explore sentence-level refinement decisions to reduce cost.

7 Limitations

We only experimented with GPT-3.5 without replicating with open-source LLMs. However, we argue that our intention is not to achieve state-of-the-art translation results, but to pose a new perspective that a simple iterative strategy can help translation naturalness. Therefore, using a powerful LLM is necessary, and open-sourced models might not be as effective. Finally, involving GPT in an iterated process is costly. We think that GPT is useful in showcasing our proposed approach, but smarter refinement strategies need to be investigated for practical use cases.

8 Ethical Statement

The texts we analyse are machine-generated. We are not able to manually examine all model outputs, but we are fairly confident that the generated texts do not include harmful or inappropriate elements that will make readers uncomfortable. Our human evaluators are university students recruited by the authors. They are paid an hourly rate higher than their local legal minimum wage.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations*.

Mona Baker. 1996. *Corpus-based Translation Studies: The Challenges that Lie Ahead*, Benjamins Translation Library. John Benjamins Publishing Company.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*.

Rajen Chatterjee, Matteo Negri, Raphael Rubino, and Marco Turchi. 2018. Findings of the WMT 2018 shared task on automatic post-editing. In *Proceedings of the Third Conference on Machine Translation*.

Kehai Chen, Masao Utiyama, Eiichiro Sumita, Rui Wang, and Min Zhang. 2022. Synchronous refinement for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Pinzhen Chen, Jindřich Helcl, Ulrich Germann, Laurie Burchell, Nikolay Bogoychev, Antonio Valerio Miceli Barone, Jonas Waldendorf, Alexandra Birch, and Kenneth Heafield. 2021. The University of Edinburgh's English-German and English-Hausa submissions to the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*.

Shamil Chollampatt, Raymond Hendy Susanto, Liling Tan, and Ewa Szymanska. 2020. Can automatic post-editing improve NMT? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. PaLM: Scaling language modeling with pathways. *arXiv preprint*.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation*.

Markus Freitag, George Foster, David Grangier, and Colin Cherry. 2020. Human-paraphrased references improve neural machine translation. In *Proceedings of the Fifth Conference on Machine Translation*.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, Eleftherios Avramidis, Tom Kocmi, George Foster, Alon Lavie, and André F. T. Martins. 2022. Results of WMT22 metrics shared task: Stop using BLEU – neural metrics are better and more robust. In *Proceedings of the Seventh Conference on Machine Translation*.

Martin Gellerstam. 1986. Translationese in Swedish novels translated from English. In L. Wollin and H. Lindquist, editors, *Translation studies in Scandinavia: Proceedings from the Scandinavian Symposium on Translation Theory II*. CWK Gleerup.

Jiatao Gu, Changhan Wang, and Junbo Zhao. 2019. Levenshtein transformer. In *Advances in Neural Information Processing Systems*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? a comprehensive evaluation. *arXiv preprint*.

Julia Ive, Lucia Specia, Sara Szoc, Tom Vanallemeersch, Joachim Van den Bogaert, Eduardo Farah, Christine Maroti, Artur Ventura, and Maxim Khalilov. 2020. A post-editing dataset in the legal domain: Do we underestimate neural machine translation quality? In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint*.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. 2018. MS-UEdin submission to the WMT2018 APE shared task: Dual-source transformer for automatic post-editing. In *Proceedings of the Third Conference on Machine Translation*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint*.

Kevin Knight and Ishwar Chander. 1994. Automated postediting of documents. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation*.

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

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho. 2018. Deterministic non-autoregressive neural sequence modeling by iterative refinement. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Gennadi Lembersky, Noam Ordan, and Shuly Wintner. 2012. Language Models for Machine Translation: Original vs. Translated Texts . *Computational Linguistics*.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on ChatGPT. *arXiv preprint*.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of the 26th International Conference on Computational Linguistics*.

Roman Novak, Michael Auli, and David Grangier. 2016. Iterative refinement for machine translation. *arXiv preprint*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Santanu Pal, Hongfei Xu, Nico Herbig, Sudip Kumar Naskar, Antonio Krüger, and Josef van Genabith. 2020. The transference architecture for automatic post-editing. In *Proceedings of the 28th International Conference on Computational Linguistics*.

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*.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Openai.com.

Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023a. Do GPTs produce less literal translations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Vikas Raunak, Amr Sharaf, Hany Hassan Awadallah, and Arul Menezes. 2023b. Leveraging GPT-4 for automatic translation post-editing. *arXiv preprint*.

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 Processing*.

Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*.

Elke Teich. 2003. *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. De Gruyter Mouton.

Antonio Toral. 2019. Post-editese: An exacerbated translationese. In *Proceedings of Machine Translation Summit XVII*.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. 2021. Facebook AI's WMT21 news translation task submission. In *Proceedings of the Sixth Conference on Machine Translation*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Longyue Wang, Mu Li, Fangxu Liu, Shuming Shi, Zhaopeng Tu, Xing Wang, Shuangzhi Wu, Jiali Zeng, and Wen Zhang. 2021. Tencent translation system for the WMT21 news translation task. In *Proceedings of the Sixth Conference on Machine Translation*.

Daimeng Wei, Zongyao Li, Zhanglin Wu, Zhengzhe Yu, Xiaoyu Chen, Hengchao Shang, Jiaxin Guo, Minghan Wang, Lizhi Lei, Min Zhang, et al. 2021. HW-TSC's participation in the WMT 2021 news translation shared task. In *Proceedings of the Sixth Conference on Machine Translation*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Weijia Xu and Marine Carpuat. 2021. EDITOR: An edit-based transformer with repositioning for neural machine translation with soft lexical constraints. *Transactions of the Association for Computational Linguistics*.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. 2023. INSTRUCTSCORE: Towards explainable text generation evaluation with automatic feedback. *arXiv preprint*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *Proceedings of the 40th International Conference on Machine Learning*.

A Additional scores for GPT refinement

Due to the space constraint, we are not able to display all metric scores in the main content, so we attach chrF++ and COMET_{DA} scores here for reference. We observe the same patterns in BLEU and chrF++ across all language pairs. Regarding COMET_{DA}, it is conditioned on the human reference, which 1) can be imperfect itself, and 2) is a subject in our comparison. Hence it might be not indicative. The Additional scores for GPT refinement experiments are listed in Table 3.

B WMT system refinement

Out of the seven WMT21 submissions, we select outputs from four models built by research labs that, based on human evaluation, have been ranked at significantly different positions on the German-to-English leaderboard: Tencent (Wang et al., 2021), Facebook AI (Tran et al., 2021), Edinburgh (Chen et al., 2021), and Huawei TSC (Wei et al., 2021). These are competitive systems built with data augmentation, multilingualism, ensembling, re-ranking, etc. We then include two online commercial systems tested in WMT 2021: Online-A and Online-Y.[3] Finally, human reference "B" is added so that we can experiment with our refinement strategy with human translations.[4] References "A" and "B" are sourced from different translation agencies (Farhad et al., 2021).

We report automatic scores from the refinement process in Table 4. We explain the results in the main content Section 3.3. Overall, we observe patterns similar to refining GPT translations. The string-based metrics see significant drops, but COMET_{QE} improves for five out of seven original entries.

C Score changes through iterations

We plot the changes in BLEU, COMET_{DA}, and COMET_{QE} in Figure 2. Apart from scores from our translate and refinement queries, we also include the human reference performance in the COMET_{QE} plot.

[3]The online systems were anonymized by WMT21 organizers, so we do not have knowledge about them. The time of access is believed to be in 2021.

[4]The overview paper of WMT 2021 states that "for German↔English, the 'B' reference was found to be a post-edited version of one of the participating online systems". We discover that it refers to English→German only, and German→English is not affected.

	WMT21 de→en		WMT21 en→de		WMT21 zh→en		WMT21 en→zh		WMT22 de→fr		WMT22 en→ja		WMT22 uk→cs	
	chrF++	COMET$_{\text{DA}}$	chrF++	COMET$_{\text{DA}}$	chrF++	COMET$_{\text{DA}}$	chrF++	COMET$_{\text{DA}}$	chrF++	COMET$_{\text{DA}}$	chrF++	COMET$_{\text{DA}}$	chrF++	COMET$_{\text{DA}}$
Reference$_{\text{A}}$	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Translate	**57.55**	**.8606**	**53.54**	.8427	**53.74**	.8199	**20.61**	.8300	**59.50**	**.8395**	25.89	.8863	**54.64**	**.9074**
Refine	51.91	.8525	50.57	**.8478**	49.06	.8156	19.28	**.8417**	55.83	.8353	**27.30**	**.8941**	53.06	.9040
Refine$_{\text{Contrast}}$	52.47	.8452	51.21	.8211	51.77	**.8538**	19.69	.8395	56.37	.8308	26.71	.8928	54.29	.9036
Refine$_{\text{Random}}$	51.79	.7777	46.56	.7906	47.11	.8323	17.49	.8126	-	-	-	-	-	-
Paraphrase	40.05	.8044	43.54	.8197	40.92	.7931	17.14	.8144	44.28	.7937	23.18	.8592	40.04	.8625

Table 3: Additional automatic scores of different strategies on translation directions from WMT 2021 and 2022 news translation.

		BLEU	chrF++	COMET$_{\text{DA}}$	COMET$_{\text{QE}}$
	Reference$_A$	-	-	-	*.0919*
Reference$_{\text{B}}$	Submission	**30.05**	**56.00**	.8497	.1050
	Refine	23.39	51.80	.8527	**.1123**
	Refine$_{\text{Contrast}}$	25.10	53.82	**.8566**	.1116
	Paraphrase	12.52	41.03	.8031	.0894
Online$_{\text{A}}$	Submission	**34.45**	**60.78**	**.8582**	.1061
	Refine	23.37	51.67	.8494	.1098
	Refine$_{\text{Contrast}}$	25.14	52.84	.8534	**.1137**
	Paraphrase	12.22	41.34	.8097	.0942
Online$_{\text{Y}}$	Submission	**32.70**	**59.32**	.8500	.0981
	Refine	22.92	50.85	**.8522**	.1080
	Refine$_{\text{Contrast}}$	24.40	53.32	.8517	**.1134**
	Paraphrase	11.97	40.29	.8054	.0892
Tencent	Submission	**35.35**	**61.28**	**.8584**	.1055
	Refine	23.75	52.16	.8488	.1095
	Refine$_{\text{Contrast}}$	26.89	54.75	.8553	**.1116**
	Paraphrase	12.43	41.35	.8116	.0947
Facebook	Submission	**34.67**	**60.78**	**.8677**	**.1146**
	Refine	22.97	51.05	.8505	.1113
	Refine$_{\text{Contrast}}$	25.74	53.88	.8548	.1130
	Paraphrase	11.80	40.99	.8099	.0922
Edinburgh	Submission	**34.20**	**60.03**	**.8588**	.1087
	Refine	22.04	50.29	.8496	.1097
	Refine$_{\text{Contrast}}$	25.24	52.87	.8546	**.1147**
	Paraphrase	12.79	40.18	.8067	.0921
Huawei	Submission	**35.13**	**61.17**	**.8643**	**.1126**
	Refine	22.24	50.82	.8519	.1097
	Refine$_{\text{Contrast}}$	24.95	52.47	.8560	**.1124**
	Paraphrase	12.20	40.74	.8078	.0909

Table 4: Automatic scores of refining WMT 2021 news shared task German-to-English submissions.

D Example outputs

We place two examples in Table 5 as a case study. The cases illustrate significant string changes, but the meaning of sentences does not vary too much. This signifies the inability to use automatic string-based metrics in distinguishing translation quality or the degree of naturalness when the outputs are relatively high-quality.

E Evaluation metric details

BLEU and chrF++ are as implemented in the `sacrebleu` toolkit.[5] We also use this toolkit to obtain test sets with references as well as past WMT systems' outputs. Specifically for tokenization in BLEU calculation, we use "zh" for Chinese, "ja-mecab" for Japanese, and "13a" for the rest. The BLEU signature is `nrefs:1 | case:mixed | eff:no | smooth:exp | version:2.3.1`, and the chrF++ signature is `nrefs:1 | case:mixed | eff:yes | nc:6 | nw:2 | space:no | version:2.3.1`. For COMET metrics, we used the official implementation released by the authors.[6]

[5] https://github.com/mjpost/sacrebleu
[6] https://github.com/Unbabel/COMET

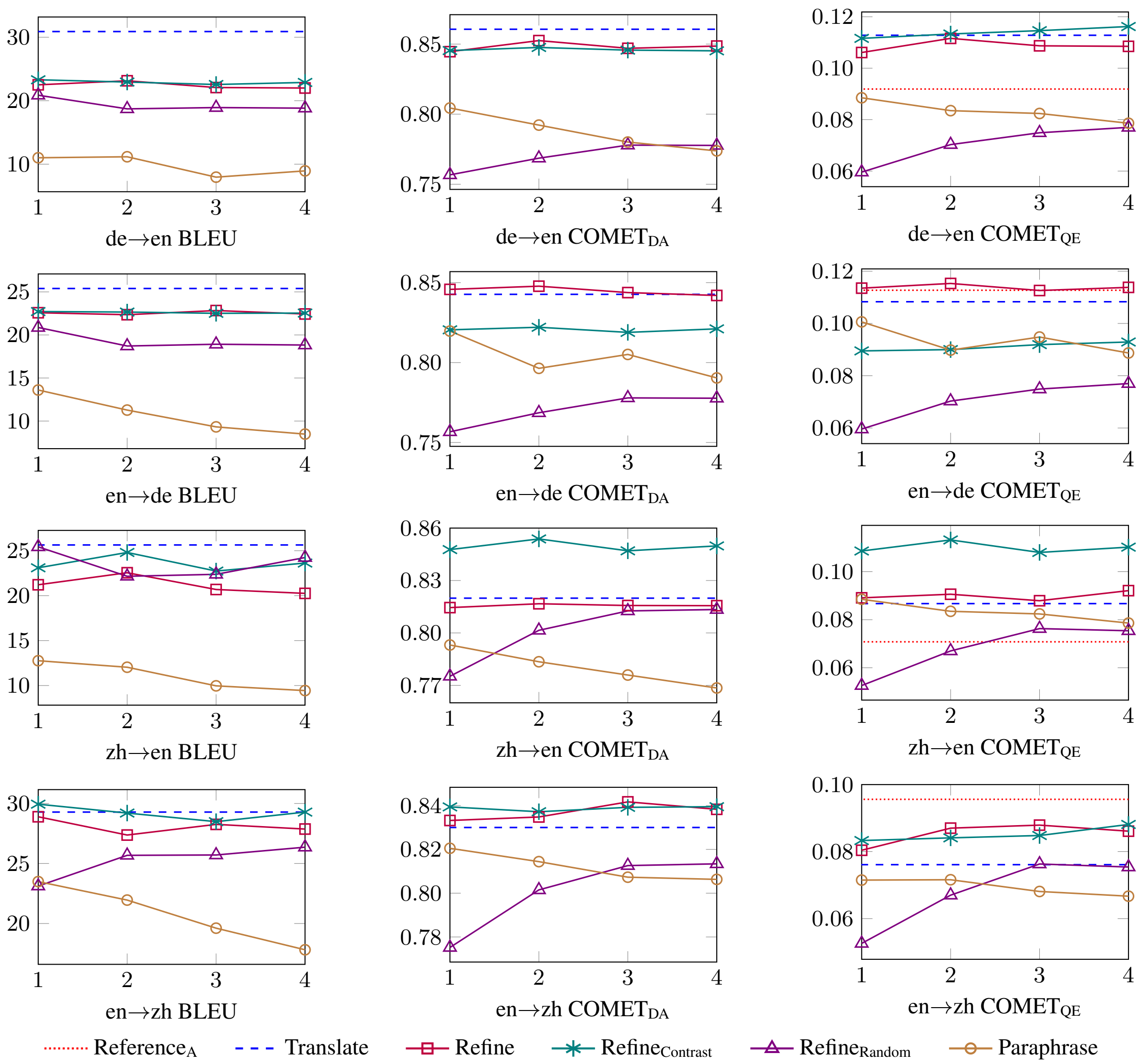

Figure 2: BLEU, COMET$_{\text{DA}}$, and COMET$_{\text{QE}}$ at different refinement and paraphrase iterations for high-resource translation.

Source	Der 17-Jährige floh zunächst vom Tatort, seine Personalien konnten aber im Nachhinein ermittelt werden.
Reference	The 17 year-old proceeded to flee the crime scene, however, his personal details could be retrieved later.
Translate	The 17-year-old initially fled from the crime scene, but his personal information was later determined.
Refine$_{\text{Contrast}}$	The 17-year-old initially fled from the scene of the crime, but his personal details could later be identified.
Paraphrase	At first, the 17-year-old ran away from where the crime occurred, but eventually, the authorities were able to identify him by his personal details.
Source	新法令规定，坎帕尼亚大区自即日起室内公共场所必须戴口罩，违者最高可处以1000欧元罚金。
Reference	According to a new decree, people must wear masks in indoor public places in Campania from now on, and offenders can be fined up to 1,000 euros.
Translate	A new regulation stipulates that in Campania, indoor public places must wear masks. Violators can be fined up to 1000 euros.
Refine$_{\text{Contrast}}$	A new regulation states that in the Campania region, masks must be worn in indoor public places, with a maximum fine of 1000 euros for those who violate the rule.
Paraphrase	A new rule in Campania requires people to wear masks in indoor public places, and those who don't follow this rule may be charged up to 1000 euros.

Table 5: German→English and Chinese→English examples showing rich lexical variations across translation strategies.

F Other related works

F.1 Translation post-editing

Closely related to translation refinement is automatic post-editing (APE), which trains a neural network to fix translation errors by learning from human correction data (Knight and Chander, 1994). While it has shown notable developments in statistical machine translation, it could become less effective in the deep learning era due to original translations being high-quality and lack of post-editing data (Junczys-Dowmunt and Grundkiewicz, 2018; Chatterjee et al., 2018). Whilst one way to facilitate this is more data provision (Chollampatt et al., 2020; Ive et al., 2020), our workaround utilizes a large language model, which possesses the post-editing capability without being specifically tuned. Furthermore, post-editing models have limited power to alleviate awkwardness, because human editing data is collected from annotators who are usually instructed to not make style improvements (Ive et al., 2020). Compared to APE, our method allows LLMs to re-generate an entirely different translation, which could escape the "post-editese" phenomenon, where Toral (2019) demonstrated that human-edited machine translations still exhibit translationese features.

Some post-editing works do not rely on the source translation or human editing data (Simard et al., 2007). For instance, Freitag et al. (2019) trained a post-editor solely on monolingual data by reconstructing the original text given its round-trip translation. In our work, we incorporate stronger natural language modelling into post-editing by employing LLMs. Other translation refinement research includes combining statistical and neural systems (Novak et al., 2016; Niehues et al., 2016), merging APE into the NMT framework (Pal et al., 2020; Chen et al., 2022), and debiasing translationese in the latent embedding space (Dutta Chowdhury et al., 2022). The iterative editing mechanism is not commonly employed in autoregressive translation or translation editing. Its use cases mostly lie in non-autoregressive translation, where each output token is independent of other target positions and iterative decoding enhances output quality (Lee et al., 2018; Gu et al., 2019; Xu and Carpuat, 2021).

F.2 Large language models

Large language models have recently become highly effective tools for various NLP tasks (Radford et al., 2019; Brown et al., 2020; Chowdhery et al., 2022; Ouyang et al., 2022). Nowadays, optimising LLMs directly for specific tasks becomes infeasible yet unnecessary since they generalize to downstream tasks without explicit supervision. With more parameters and training data, LLMs may offer stronger performance than dedicated translation or post-editing models. The method we use to elicit a response from GPT is zero-shot hard prompting (Brown et al., 2020), which means affixing a description to the original task input to form a query to the model. Researchers have benchmarked LLMs' capability to translate (Vilar et al., 2023; Zhang et al., 2023; Jiao et al., 2023; Hendy et al., 2023), and to evaluate translations (Kocmi and Federmann, 2023; Lu et al., 2023; Xu et al., 2023).

Recent findings show that GPT produces less literal translations, especially for out-of-English translations (Raunak et al., 2023a), which to some extent stands in contrast with our evaluation outcome. Concurrent with our study, Raunak et al. (2023b) formalized post-editing as a chain-of-thought process (Wei et al., 2022) with GPT-4 and showed promising results. Different from their focus, our work features the iterative refinement process as a means to enhance naturalness and fluency. We have shown that iterated refinement is better than one-off editing. Our improvement, especially for into-English, may be attributed to the abundant English pre-training data available for LLMs. To the best of our knowledge, although the concept of iterative refinement is not new, ours is the pioneering paper in applying such strategies to LLMs for translation.