Creative and Context-Aware Translation of East Asian Idioms with GPT-4

Kenan Tang[1] *, **Peiyang Song**[2] *, **Yao Qin**[1], **Xifeng Yan**[1]
[1] UC Santa Barbara, [2] California Institute of Technology
kenantang@ucsb.edu, psong@caltech.edu, yaoqin@ucsb.edu, xyan@cs.ucsb.edu

Abstract

As a type of figurative language, an East Asian idiom condenses rich cultural background into only a few characters. Translating such idioms is challenging for human translators, who often resort to choosing a context-aware translation from an existing list of candidates. However, compiling a dictionary of candidate translations demands much time and creativity even for expert translators. To alleviate such burden, we evaluate if GPT-4 can help generate high-quality translations. Based on automatic evaluations of faithfulness and creativity, we first identify Pareto-optimal prompting strategies that can outperform translation engines from Google and DeepL. Then, at a low cost, our context-aware translations can achieve far more high-quality translations per idiom than the human baseline. We open-source all code and data to facilitate further research[1].

1 Introduction

Figurative language is a challenge for both linguistic analysis (Dancygier, 2014) and many natural language processing (NLP) tasks (Chakrabarty et al., 2022). One representative task is literary translation (Karpinska and Iyyer, 2023), where the translation of figurative language is one major difficulty. Among figurative language constructs, idioms are especially hard for a machine translation (MT) model due to their non-compositionality. For example, the meaning of the idiom "bite the bullet", deciding to do something difficult, is not simply composed of the meanings of "bite" and "bullet".

Of idioms in all languages, East Asian idioms constitute an interesting subset. Each of these idioms condenses its figurative meaning into a small number of characters, dominantly 4 characters (Chinese: *sizichengyu*, Japanese: *yojijukugo*, Korean: *sajaseong-eo*, all literally meaning "4-character idioms"). As the set of commonly accepted East Asian idioms and their meanings do *not* change largely over time, the challenge of translating such idioms can seemingly be tackled by using a fixed list of literal and figurative candidate translations. This strategy has been commonly adopted by human translators (Tang, 2022) and MT researchers (Li et al., 2024) alike.

However, this approach has a major limitation. The existing East Asian idiom translation datasets provide translations out of context, i.e., idioms were translated without being incorporated into a surrounding sentence or paragraph. Hence, the

1. Idiom
刮目相看
2. Sentences
小明的成绩提高得非常快，让老师和同学们都刮目相看。
3. Context-Aware Translations
Xiaoming's grades soared impressively, leaving both teachers and classmates in awe.
Extracted Spans (From Multiple Translations)
• leaving both teachers and classmates in awe • taken everyone by surprise • like a phoenix reborn from its ashes • a blazing comet • earned everyone's admiration • with newfound respect • . . .
Human Reference (Sentences Not Available)
• treat somebody with increased respect • look at somebody with new eyes • have a completely new appraisal of somebody • regard somebody with special esteem

Table 1: **With our methods, we generate far more context-aware translations than human reference.** The pipeline of our context-aware translation is shown in Steps 1-3. To show results more clearly, we automatically extract the span (continuous words) in the translation that corresponds to the original idiom (Section 3.3). More examples are available in Appendix D.

[1] https://github.com/kenantang/cjk-idioms-gpt
*Equal contributions.

Findings of the Association for Computational Linguistics: EACL 2024, pages 9285–9305
November 12-16, 2024

translations sometimes require significant rewording to be appropriate in a given context. An example of such limitation is shown in Table 1. Despite that all 4 human reference translations are correct, the first 2 are awkward, as they overexaggerate a teacher's attitude towards a student as "increased respect" or "special esteem", and the last 2 use the active voice that interrupts the flow of the sentence.

In this work, we alleviate this limitation by using a SoTA large language model (LLM), GPT-4, to generate a dataset of context-aware idiom translations. We prompt GPT-4 to use different strategies to translate each idiom within various contexts. Moreover, to avoid accumulating translations by a brute-force and costly repetition of prompting, we select a small subset of Pareto-optimal prompting strategies from a comprehensive set, including zero-shot instructions inspired by human expertise and few-shot prompts that reuse high-quality translations. Table 1 shows the steps that lead to a successful example, where our translations are superior in diversity and quality to the human reference. Our methods also beat commercial translation engines from Google and DeepL (Section 3.2).

2 Method

In this section, we elaborate on the methods and experiment details for each step shown in Table 1.

Step 1: Idioms We obtain idioms from a dictionary for Chinese (Tang, 2022) and online resources for Japanese[2] and Korean[3]. These sources cover commonly used idioms in the 3 East-Asian languages. To test if our method generalizes to uncommon or new words that have an idiom-like structure, we also curate a set of plausible Chinese idioms. Plausible idioms are GPT-4-generated words which are not real idioms, but can fool GPT-4 when we ask it if the word is an idiom (Appendix A.1). For convenience, we use "plausible Chinese" to refer to the language of these idioms for convenience. For the 4 source languages, we limit the target language to English. To our best knowledge, the only human baseline that provides multiple translations for each idiom is the Chinese-English dictionary we use.

Step 2: Generate Sentences To translate an idiom with context awareness, we need to translate a sentence that contains this idiom. Hence, we first generate multiple sentences containing a given idiom with GPT-4. For each of the 4 languages, we randomly sample 50 idioms and generate 10 sentences for each idiom, totalling 500 sentences.

Step 3: Context-Aware Translation Overall, we want multiple translations of each idiom within different contexts. This could be achieved if we only use a standard prompt (BASELINE) to generate one translation per sentence. However, the contextual information is provided not only by the sentence but also by the paragraph that surrounds it. For example, a sentence can be translated more vividly when it appears in an everyday conversation than in a history book, but the BASELINE translation is formal when no instructions are given (Table 2). To always have an option when a context is given, we generate multiple translations of each sentence by the following prompting strategies. Full prompts for each strategy can be found in Appendix B.

Sentence 他们通过威逼利诱，想要我放弃诉讼。
Baseline (History Book) They tried to get me to drop the lawsuit through threats and inducements.
Analogy Creative (Everyday Conversation) They tried to make me drop the lawsuit through a carrot and stick approach.

Table 2: **The same idiom in different contexts (paragraphs) requires different translation strategies, even when the sentence is the same.** The two English sentences are translations of the same Chinese sentence. In all three sentences, the parts corresponding to the idiom is highlighted. Our pipeline is able to offer abundant choices for different contexts (in parentheses) by utilizing a comprehensive set of strategies (in bold).

Creativity is the key to adaptation in different contexts. To invoke creativity naively, we use a two-turn prompt that asks GPT-4 for 5 translations of one sentence (DIVERSITY EXPLICIT) and then for another 5 translations (DIVERSITY DIALOG).

However, we should be able to get context-aware translations more efficiently. Instead of hoping for context-aware translations to be generated by sheer chance, we can directly ask for such translations. To do so, we explicitly ask GPT-4 to translate "creatively" (ZERO-SHOT CREATIVELY).

Furthermore, we can add detailed instructions based on human expertise, instead of letting GPT-4 implicitly choose its translation strategy. Inspired by common translation strategies (Molina and Hurtado Albir, 2002), we use the following prompts: assuming a sentence appears in a paragraph of

[2] https://dictionary.goo.ne.jp/idiom/
[3] https://github.com/LiF-Lee/idioms/

a certain genre (CONTEXT EXPLICIT), using an analogy that is common (ANALOGY NATURAL) or uncommon (ANALOGY CREATIVE), shuffling the order of clauses (SHUFFLE ORDER), rewriting the sentence without an idiom and then translating (TWO-STEP), avoiding using continuous spans (DISCONTINUOUS 1) or multi-word expressions (DISCONTINUOUS 2). For CONTEXT EXPLICIT, we use 4 genres: a news report, a romance novel, an everyday conversation, and a history book.

While we can generate an abundance of translations, many are expected to be mundane or repetitive. Thus, to select a small subset that quickly helps human translators, we need to evaluate all translations to identify the best ones. Due to the low reproducibility and high cost of human evaluation, we instead prompt GPT-3.5 to score each sentence translation on a 1-5 scale based on faithfulness or creativity (Appendix B.7). These two aspects are good proxies for the context-awareness of the idiom translation in the sentence. For each of the two aspects, the final score of a sentence translation is averaged from 5 runs, and the overall score of a prompting strategy is averaged from the scores of translations it produces. We use both aspects to select Pareto-optimal prompting strategies.

After the initial round of evaluation, we reuse high-quality zero-shot translations as examples for few-shot prompting, prioritizing creativity. We choose the most creative translations and randomly sample from them to construct 5-shot prompts. The most creative translations are ones that score a 5 on creativity in at least 1 out of 5 runs. We use a 5-shot prompt with the word "creatively" (FEW-SHOT CREATIVELY) or without (FEW-SHOT). The few-shot translations are evaluated using the same procedure for zero-shot translations.

Overall, we generate 27 translations per sentence with all prompting strategies, totalling 13,500 translations per language (Table 3). After identifying Pareto-optimal strategies, we further apply them to more idioms to expand our dataset (Section 3.3).

For generation, we use the GPT-4 API and GPT-3.5 API from OpenAI[4]. Google[5] and DeepL are used as commercial translation engine baselines.

In Appendix A, we discuss alternative choices of resources, including dictionaries, parallel corpora, evaluation metrics, and LLMs. While we only experiment on the set of resources we choose above, we already reach the goal of benefiting translators by generating sufficiently high-quality translations.

[4] `gpt-4-0125-preview` and `gpt-3.5-turbo`

[5] `translating-text-v3`

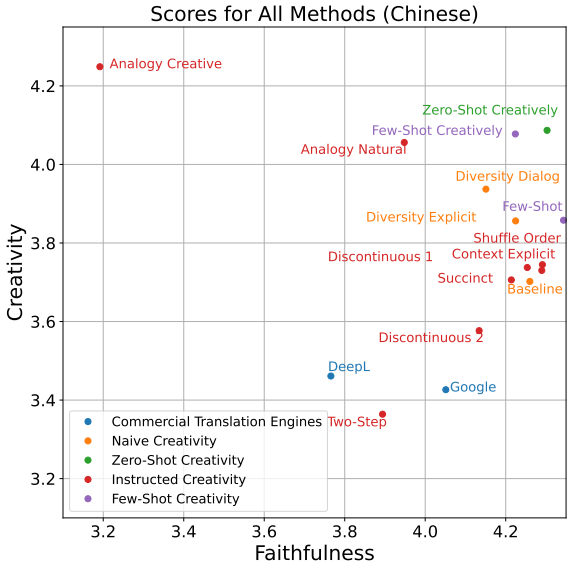

Figure 1: **Our strategies significantly differ in the mean faithfulness and creativity scores.** The strategies closer to the upper-right corner are better.

3 Results and Discussion

In this section, we first analyze the generated translations. Then, we examine how the optimal translation strategies work on a larger set of idioms. We also discuss how the dataset can be extended by an alternative pipeline using paragraph context.

3.1 Quantitative Analysis

To pick the Pareto-optimal translation strategies in each language, we aggregate the faithfulness and creativity scores of the translations from each strategy. The mean scores for Chinese idioms are visualized in Figure 1, while the visualizations for other languages and all numerical results are listed in Appendix E. Here, we summarize 5 trends that generally hold true for all 4 languages.

First, GPT-4 is better at idiom translation than commercial translation engines. Both faithfulness and creativity are higher for GPT-4 translations than for Google and DeepL translations. This ranking aligns with our qualitative observations and supports the validity of our evaluation method.

Secondly, by naively invoking creativity of GPT-4 (DIVERSITY EXPLICIT and DIVERSITY DIALOG), we are able to improve creativity over the BASELINE, at the cost of faithfulness. This result shows an inevitable trade-off between faithfulness and creativity without further instructions.

Thirdly, by simply adding the word "creatively" into the prompt (ZERO-SHOT CREATIVELY), we are able to improve over the naive strategy and overcome the trade-off. This result motivates the search for stronger and more cost-efficient prompts, instead of repeatedly using weak prompts.

Fourthly, few strategies based on human expertise result in a Pareto improvement. This suggests that human expertise does not necessarily transfer to strong prompts, at least in the form of short prompts we use to briefly describe human translators' strategy. While a longer prompt with detailed instructions may bring out the full potential of a certain strategy, we do not consider these forms of prompts due to their high cost.

Finally, few-shot prompting strategies (FEW-SHOT CREATIVELY and FEW-SHOT) are often Pareto-optimal. This result reveals the potential in using longer prompts and more sophisticated strategies to improve performance. However, from zero-shot to few-shot, the small improvement in scores costs many more tokens per idiom.

3.2 Qualitative Analysis

We show translation examples in Table 1 and Appendix D. For Chinese, we can compare our translations with the human reference. Thanks to the large number of different translations we get, most of them have not appeared in the dictionary. This shows that our method wins in diversity.

Regarding quality, GPT-4 almost always translate the idiom correctly[6], and the translation quality are comparable to that of the human baseline (Table 1). In contrast, despite producing fluent sentences, Google and DeepL noticably misinterpret some idioms. For example, DeepL mistranslates the Chinese idiom "威逼利诱" (literally "coercion and coaxing") as "bullying" in the sentence in Table 2.

While all GPT-4-based strategies are able to produce faithful and creative translations, the proportions of such translations apparently differ for each strategy. In the GPT-4 translations, we observe two major failure patterns that cause a quality drop. First, GPT-4 fails to follow instructions on the translation strategy, producing idiom translations that are the same as the one from the BASELINE prompt. While this behavior lowers the scores, these outputs are still valid translations, and outputs in undesired formats are rare (Appendix C). Secondly, in the cases where instructions are fully followed, GPT-4 sometimes uses the strategy to improve the translation of other parts of the sentence, but not necessarily of the idiom itself.

[6]We look at more than a total of 10,000 translations by GPT-4 for over 100 random idioms, and we do not see obvious misinterpretation.

Item	Count
Idioms	50
Sentences	$50 \times 10 = 500$
Translations	$50 \times 10 \times 27 = 13500$
Idioms	500
Sentences	$500 \times 10 = 5000$
Translations	$500 \times 10 \times 4 = 20000$

Table 3: **The total number of idiom translations we generated for Chinese.** We first use all 27 translation strategies on 50 idioms. To expand the dataset, we then use 4 Pareto-optimal strategies on another 500 idioms.

3.3 Extension to a Larger Set

To validate our methods on a larger set of idioms, we apply 4 Pareto-optimal strategies (ZERO-SHOT CREATIVELY, ANALOGY CREATIVE, FEW-SHOT, and FEW-SHOT CREATIVELY) on 500 top-frequency Chinese idioms (Appendix A.2). For the two few-shot methods, we reuse the most creative translations of the original 50 Chinese idioms. The numbers of Chinese idiom translations from all experiments are summarized in Table 3.

Table 4 shows that high scores are maintained for the new translations. We would like to further validate our evaluation strategy by comparing against the popular reference-free quality estimation (QE) metric COMETKIWI (Rei et al., 2023)[7]. Interestingly, the COMETKIWI ranking is only consistent with the one given by our faithfulness score, suggesting the limitation of traditional QE metrics when creativity is among the evaluation criteria.

Other than increasing the number of idioms, increasing the number of sentences containing each idiom is also a way to extend the dataset. Though we limit the number of sentences per idiom to 10, we observe that the number of different translations of each idiom keeps increasing with the number of sentences. To show this, we first extract the span in each translation that corresponds to the idiom (Appendix C.7). Then, we use the number of unique unigrams in the spans as a proxy for the number of different translations for each idiom. We count unique unigrams instead of unique spans in order to avoid counting trivially different translations that

[7]https://huggingface.co/Unbabel/wmt23-cometkiwi-da-xxl

Method	Faithfulness	Creativity	CometKiwi
ZSC	4.27 ± 0.62	4.08 ± 0.36	0.79 ± 0.09
AC	3.01 ± 1.02	$\mathbf{4.27 \pm 0.48}$	0.59 ± 0.14
FS	$\mathbf{4.31 \pm 0.63}$	3.80 ± 0.49	$\mathbf{0.83 \pm 0.08}$
FSC	4.17 ± 0.71	4.07 ± 0.40	0.77 ± 0.12

Table 4: **The scores (mean $\pm$ standard deviation) for the Pareto-optimal translation strategies (denoted by initials) are maintained on a larger set of idioms.** Faithfulness and creativity are in $[1, 5]$. CometKiwi is in $[0, 1]$. Highest (best) scores are boldfaced. The number of idiom translations from each method is 500 (sampled from 5,000). Faithfulness and CometKiwi give the same ranking. The strategy AC with the highest creativity is the lowest in faithfulness and CometKiwi.

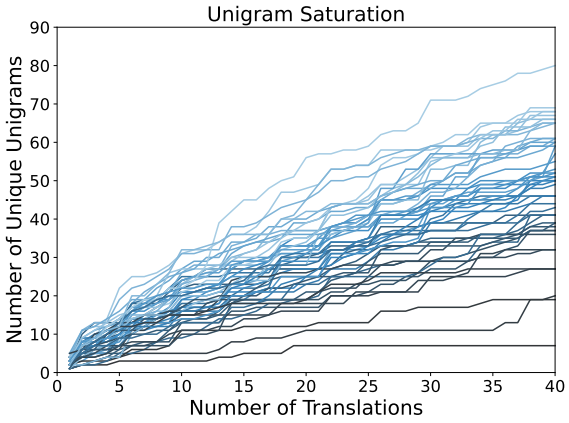

Figure 2: **The number of unique unigrams increases with the number of translations.** Increase rates and saturation points differ across 50 random idioms from 500 top-frequency Chinese idioms. Since there are 10 sentences for each idiom and 4 Pareto-optimal strategies, the total number of translations for each idiom is 40. For most idioms, the increase in the number of unique unigrams does not saturate at 40 translations.

only rearrange parts of other translations. For most idioms, the number of unique unigrams do not saturate as we increase the number of translations (Figure 2). This illustrates the potential of scaling up our methods to more translations (sentences).

3.4 Enhancing Context-Awareness

Contrary to our expectations, we notice that the Context Explicit prompts often do not change the translation (one failure pattern in Section 3.2). Hence, for this specific strategy, we try to enhance context-awareness with a stronger pipeline. First, we specify different genres and generate multiple paragraphs that contain the sentence. Then, we ask GPT-4 to translate the paragraph, using the following 4 types of prompts. The first type is a baseline translation prompt. The second type encourages GPT-4 to emphasize 3 evaluation aspects, namely faithfulness, creativity, and word choices that match the theme. The third type utilizes step-by-step instructions that are generated by Auto-CoT (Liu et al., 2023). The fourth type is a multi-turn dialog that asks GPT-4 to iteratively improve the translation based on each aspect. The full prompts can be found in Appendix B.5.

For each of the 4 languages, we choose 20 idioms and 1 sentence per idiom. We obtain a total of 800 translated paragraphs using 4 genres and 10 prompts. With the paragraph pipeline, we observe higher variation in translations from different genres than with our previous sentence pipeline. However, the different prompting strategies do not lead to meaningful variations for the same source paragraph (Appendix D.2). Due to the high cost of this pipeline, we leave the investigation of more idioms and prompting strategies as a future work.

4 Related Work

Different from traditional MT models, LLMs can produce diverse and less literal translations (Raunak et al., 2023a). Hence, a series of work has targeted generating diverse translations with LLMs and selecting the best. On one hand, translations can be generated from scratch using various prompting strategies. Then, ensembling methods can be applied to select the best candidates (Farinhas et al., 2023). On the other hand, candidates can be further refined. Some examples include refining candidates generated by other machine translation systems (Raunak et al., 2023b), iterative editing (Chen et al., 2024; Briakou et al., 2024), self-correction (Feng et al., 2024a), and multi-agent debate (Liang et al., 2023). While existing work focused mostly on general translation, our work contributes in the more challenging task of generating diverse, high-quality translations for idioms.

5 Conclusion

In this work, we thoroughly test prompting strategies that generate different translations for an East Asian idiom. We identify the strategies that generate most creative and faithful translations. To our surprise, the prompts derived from human experience do not consistently generate quantitatively better translations. Finally, we use the Pareto optimal strategies to construct a dataset of high quality translations, which can help human translators.

Limitations

Limitations of our work include:

- **No external databases.** We discover that few-shot prompts could produce competitive translations, but are not able to use external translation as examples due to the lack of a high-quality and well-aligned parallel corpora.
- **No multi-agent or role-playing.** These orthogonal prompt-based directions provide effective methods that could possibly be combined with ours.
- **No language-specific strategies.** Our set of translation strategies is language-agnostic and thus not exhaustive.
- **No idiom categorization.** We do not consider the diverse linguistically motivated categorization of idioms in our general pipeline.
- **No expert evaluation.** We are not able to obtain the evaluation of translation quality from a large crowd of full-time professional translators due to cost and resource limits. Opinions from a smaller set of evaluators may carry personal biases, especially for the creative translation task we are investigating. Thus, we do not include human study in this work.

We would like to address these limitations in future work.

References

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+Dublin. Association for Computational Linguistics.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. *arXiv preprint arXiv:2409.06790*.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Pinzhen Chen, Zhicheng Guo, Barry Haddow, and Kenneth Heafield. 2024. Iterative translation refinement with large language models. *Preprint*, arXiv:2306.03856.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Barbara Dancygier. 2014. *Figurative language*. Cambridge University Press.

António Farinhas, José de Souza, and Andre Martins. 2023. An empirical study of translation hypothesis ensembling with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore. Association for Computational Linguistics.

Yunlong Feng, Yang Xu, Libo Qin, Yasheng Wang, and Wanxiang Che. 2024a. Improving language model reasoning with self-motivated learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8840–8852, Torino, Italia. ELRA and ICCL.

Zhaopeng Feng, Yan Zhang, Hao Li, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. 2024b. Improving llm-based machine translation with systematic self-correction. *Preprint*, arXiv:2402.16379.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, 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.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023. Exploring human-like translation strategy with large language models. *Preprint*, arXiv:2305.04118.

Yuchen Jiang, Tianyu Liu, Shuming Ma, Dongdong Zhang, Jian Yang, Haoyang Huang, Rico Sennrich, Ryan Cotterell, Mrinmaya Sachan, and Ming Zhou. 2022. BlonDe: An automatic evaluation metric for document-level machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1550–1565, Seattle, United States. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

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.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, Maja Popović, and Mariya Shmatova. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. Translate meanings, not just words: Idiomkb‘Ă‘Źs role in optimizing idiomatic translation with language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *Preprint*, arXiv:2305.19118.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Chenyang Lyu, Zefeng Du, Jitao Xu, Yitao Duan, Minghao Wu, Teresa Lynn, Alham Fikri Aji, Derek F. Wong, and Longyue Wang. 2024. A paradigm shift: The future of machine translation lies with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1339–1352, Torino, Italia. ELRA and ICCL.

Lucía Molina and Amparo Hurtado Albir. 2002. Translation techniques revisited: A dynamic and functionalist approach. *Meta*, 47(4):498–512.

Yongyu Mu, Abudurexiti Reheman, Zhiquan Cao, Yuchun Fan, Bei Li, Yinqiao Li, Tong Xiao, Chunliang Zhang, and Jingbo Zhu. 2023. Augmenting large language model translators via translation memories. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10287–10299, Toronto, Canada. Association for Computational Linguistics.

Hongbin Na, Zimu Wang, Mieradilijiang Maimaiti, Tong Chen, Wei Wang, Tao Shen, and Ling Chen. 2024. Rethinking human-like translation strategy: Integrating drift-diffusion model with large language models for machine translation. *Preprint*, arXiv:2402.10699.

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.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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 (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Awadalla, and Arul Menezes. 2023b. Leveraging GPT-4 for automatic translation post-editing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the

quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.

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

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. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sai Cheong Siu. 2023. Chatgpt and gpt-4 for professional translators: Exploring the potential of large language models in translation. *Available at SSRN 4448091*.

Kenan Tang. 2022. Petci: A parallel english translation dataset of chinese idioms. *arXiv preprint arXiv:2202.09509*.

Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2024. Cat-llm: Prompting large language models with text style definition for chinese article-style transfer. *Preprint*, arXiv:2401.05707.

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.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth'l'e Lacroix, Baptiste Rozi'Ire, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Danqing Wang and Lei Li. 2023. Learning from mistakes via cooperative study assistant for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10667–10685, Singapore. Association for Computational Linguistics.

Longyue Wang, Zhaopeng Tu, Yan Gu, Siyou Liu, Dian Yu, Qingsong Ma, Chenyang Lyu, Liting Zhou, Chao-Hong Liu, Yufeng Ma, Weiyu Chen, Yvette Graham, Bonnie Webber, Philipp Koehn, Andy Way, Yulin Yuan, and Shuming Shi. 2023. Findings of the WMT 2023 shared task on discourse-level literary translation: A fresh orb in the cosmos of LLMs. In *Proceedings of the Eighth Conference on Machine Translation*, pages 55–67, Singapore. Association for Computational Linguistics.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Empowering llm-based machine translation with cultural awareness. *Preprint*, arXiv:2305.14328.

A Resources

In this section, we compare alternative resources with the ones we chose.

A.1 Dictionaries

The dictionaries we used contain 4,310 idioms for Chinese, 2440 for Japanese, and 2,316 for Korean. These dictionaries are all available for public use. While other larger idiom dictionaries are available[8], our primary focus in this work was not a comprehensive coverage of all idioms in a language. The dictionaries we used should have included the idioms that are most frequently used.

It is worth noting that East Asian idioms include more than just the 4-character category. One example is the *xiehouyu* in Chinese that has a different format. While we did not consider these other categories for experiments, our methods are applicable.

An interesting question is whether we can retrieve the idioms from GPT-4 just like we can retrieve the translations. To mine idioms, we asked GPT-4 to list a given number of idioms with a given initial. The initials are all valid *pinyin* syllables in Chinese, and they can be written as one or more roman letters (e.g. "a", "an", and "ang"). We obtained a list of 416 *pinyin* syllables from Wikipedia[9]. One syllable might be a prefix of another, so different idiom lists returned by GPT-4 might partially overlap with each other. Based on various idiom dictionaries used in previous work in the NLP community (Li et al., 2024), we estimated that the total number of frequently used Chinese idioms does not exceed 10K. Hence, for each initial, we asked GPT-4 to list 200 idioms. We expected the sufficiently large number of queries would give

[8] https://github.com/pwxcoo/chinese-xinhua/blob/master/data/idiom.json

[9] https://en.wikipedia.org/wiki/Comparison_of_Standard_Chinese_transcription_systems

us a comprehensive list after deduplication. Furthermore, for each initial, we made 5 queries with different random seeds to improve stability, as we observed that GPT-4 produced results in some runs that were much worse than in other runs.

We found that when listing idioms, GPT-4 tended to provide explanation or pronunciation of the idioms. Since these information are irrelevant for the listing task and significantly increase the output length, we asked GPT-4 to only list idioms without explaining them. We also found that when we did not explicitly require GPT-4 to list different idioms, GPT-4 tended to repeat a small set of idioms during listing. Hence, we explicitly asked GPT-4 to list different idioms.

Still, with the constraints in the prompt, GPT-4 occasionally produced undesirable content. We summarize the failure patterns below.

First, not every query returned with 200 results. This can be expected, as there do not exist 200 idioms for some initials. However, we saw a large number of queries stopping at exactly 100 results. This indicated that GPT-4 was not interpreting the number in the instruction with full precision.

Secondly, when given a certain initial, GPT-4 returned idioms with this initial in the beginning of the list, but idioms with different initials appeared later in the list. This was another example showing that the instruction was not precisely understood, even for the very simple task of listing.

Thirdly, a majority of the returned expressions were not Chinese idioms. These fake idioms could be divided into two categories. For a idiom in the first category, when we asked GPT-4 whether this is a Chinese idiom, GPT-4 successfully identified the idiom to be fake. Similar to the previous failure pattern, this indicates that the classification ability of GPT-4 is weaker during listing than when classifying a single example. Some examples of the first category included general multi-word expressions (e.g. "半导体照明") and real idioms with a single character replaced (e.g. "按甲不动" from "按兵不动"). The second category was more intriguing, as GPT-4 identified the fake idioms to be real (e.g. "落翅螳螂"). In this category, the seemingly plausible idioms are constructed in a very similar way as real idioms. Hence, we manually selected 50 such idioms when we tested the translation strategies.

A.2 Parallel Corpora

We have observed a low occurrence rate of idioms in large scale parallel corpora, for example the training set of WMT'23 (Kocmi et al., 2023). Idioms appear more often in literary text. However, due to copyright restrictions, the available literary parallel corpus is usually very small (Thai et al., 2022). The largest literary parallel corpus we have found is the BWB corpus (Jiang et al., 2022), which is a publicly available dataset of the English translation of Chinese web novels. Furthermore, due to sentence rearranging in literary translation, the smallest unit of a source-target pair in BWB is paragraphs. This makes it difficult to use BWB as a baseline to compare with the translations we get. Hence, we only used BWB to estimate the frequency ranking of idioms, where the frequency of an idiom is defined as the number of sentence pairs that contain this idiom. For the Chinese-English translation direction, GuoFeng (Wang et al., 2023) is another large-scale literary parallel corpus. Since the dataset is also derived from web novels, we do not assume a large difference when the dataset is used for frequency estimation instead of BWB.

A.3 Evaluation Metrics

Translation quality estimation has long been studied in the NLP community. Currently, the popular metrics are COMET (Rei et al., 2020), COMETKIWI (Rei et al., 2022), BLEURT (Sellam et al., 2020), BLEU (Papineni et al., 2002), and chrF++ (Popović, 2017)[10]. Among these metrics, the reference-free COMETKIWI is the most suitable for a creative generation task. We used the `wmt-23-cometkiwi-da-xxl` version of COMETKIWI (Rei et al., 2023).

LLM-based metrics have been shown to perform well on text summarization and dialog generation (Liu et al., 2023). For translation, LLMs were also applied to generate scores and textual evaluations (Kocmi and Federmann, 2023; Fernandes et al., 2023) based on MQM (Freitag et al., 2021). These work validated our choice of GPT-4 as an automatic evaluator of translation quality. We also explored the novel setting of evaluating creativity of translation, an aspect not covered by MQM.

Imperfect as they are, LLM-based automatic metrics are suitable for the assumed purpose of our dataset. We would like to provide a set of relatively good translations for a human translator to choose from. The automatic metrics reduce the cost of retrieving translations from LLMs and the time of human translators reading through the list.

[10]Both BLEU and chrF++ are implemented in SacreBLEU (Post, 2018).

Automatic metrics may produce false negatives (high-quality translations that are excluded due to low scores) and false positives (low-quality translations that are chosen due to high scores). However, on the one hand, false negatives are not concerning due to the sheer number of translations we are able to retrieve. On the other hand, false positives can be easily identified by human translators. Hence, we chose automatic metrics to help data collection.

A.4 Large Language Models

LLMs other than GPT-4 have been widely used on translation-related tasks. Some examples are Claude-2[11], Gemini-Pro[12], Flan-T5 (Chung et al., 2024), GPT-NeoX (Black et al., 2022), and LLaMA (Touvron et al., 2023). Since the prompting strategies we used in this work is model agnostic, it would be beneficial to use them on any model from the rapidly evolving set of LLMs. There is not a wide agreement on which model to choose. A consistent gap in translation quality between models was reported under some prompting strategies (Wang and Li, 2023) and datasets but not others (Feng et al., 2024b).

B Prompts

In this section, we list all prompts we used. While a variety of prompts have been used for translation (Table 5), we used the prompts that more directly describes both our idiom translation task and relevant translation instructions. For all the prompts, we use a temperature of 1.0. In our pilot study, we observe that changing the temperature in the API call does not produce meaningful variations in the generated translations. The total cost of all GPT-related experiments in this paper, including pilot studies, was $480.55.

B.1 Idioms

For Chinese idiom mining, we used the prompt:

- Give 200 Chinese idioms that begin with `<PINYIN>`. Only list idioms. Do not explain them. No duplicates.

Here, `<PINYIN>` is chosen from a list of 416 *pinyin* syllables obtained from Wikipedia[13].

For checking if a result is a true Chinese idiom, we used the prompt:

- Is `<IDIOM>` a Chinese idiom? Output yes or no.

For explanation generation, we used the prompt:

- Is `<PLAUSIBLE IDIOM>` a Chinese idiom? Please explain.

B.2 Sentences

For sentence generation, we used the prompt:

- Can you make 10 `<LANGUAGE>` sentences with the `<LANGUAGE>` idiom `<IDIOM>`? Only list sentences. Do not explain.

For plausible Chinese, we use "Chinese" as the `<LANGUAGE>`.

B.3 Zero-Shot Translations

For zero-shot translation, we used the prompts in Table 6.

B.4 Few-Shot Translations

For few-shot translation, we used the prompts in Table 7. The example sentence pairs are randomly chosen from a set of sentence pairs with highly creative translations.

B.5 Paragraph Translations

For instruction generation, we used the following 3 prompts:

- If you are asked to translate a paragraph that contains a `<LANGUAGE>` idiom, what would you do to ensure that the translation of the idiom is faithful?

- If you are asked to translate a paragraph that contains a `<LANGUAGE>` idiom, what would you do to ensure that the translation of the idiom is creative?

- If you are asked to translate a paragraph that contains a `<LANGUAGE>` idiom, what would you do to ensure that the translation of the idiom matches the theme of its context?

For paragraph translation, we used the prompts in Table 8.

[11] https://www.anthropic.com/index/claude-2

[12] https://cloud.google.com/vertex-ai/docs/generativeai/learn/models

[13] https://en.wikipedia.org/wiki/Comparison_of_Standard_Chinese_transcription_systems

Reference	Strategies
Jiao et al. (2023)	ChatGPT generated templates
Siu (2023)	Instructions in multi-turn dialogs
Lyu et al. (2024)	Specifying a poetic style
He et al. (2023)	Providing keywords, topics, or demonstrations mined by ChatGPT
Na et al. (2024)	Using Skopos, functional equivalence, or text typology theory
Tao et al. (2024)	Specifying style by several word-level and sentence-level statistics
Mu et al. (2023)	Extracting most similar sentences from a translation database
Yao et al. (2023)	Providing the whole sentence and a literal translation of a cultural-specific entity

Table 5: A summary of prompting strategies for translation.

Name	Prompt
BASELINE	Please translate the following sentence from <LANGUAGE> to English: <SENTENCE>
DIVERSITY EXPLICIT	Please generate 5 different translations of the following sentence from <LANGUAGE> to English: <SENTENCE>
DIVERSITY DIALOG	Please generate another 5 different translations.
ZERO-SHOT CREATIVELY	Please creatively translate the following sentence from <LANGUAGE> to English: <SENTENCE>
CONTEXT EXPLICIT	The sentence below comes from <GENRE>. Please translate it from <LANGUAGE> to English: <SENTENCE>
ANALOGY NATURAL	Please translate the following sentence from <LANGUAGE> to English: <SENTENCE> In the translation, please use an analogy commonly used in English.
ANALOGY CREATIVE	Please translate the following sentence from <LANGUAGE> to English: <SENTENCE> In the translation, please create a new analogy that has not been commonly used in English.
SHUFFLE ORDER	Please translate the following sentence from <LANGUAGE> to English: <SENTENCE> Please try to change the order of clauses to make the translation more natural.
SUCCINCT	Please translate the following sentence from <LANGUAGE> to English: <SENTENCE> Please translate the <LANGUAGE> idiom appeared in the sentence as succinctly as possible.
TWO-STEP	Please rewrite the following sentence in <LANGUAGE> without using a <LANGUAGE> idiom: <SENTENCE> Please translate the rewritten sentence to English.
DISCONTINUOUS 1	Please translate the following sentence from <LANGUAGE> to English: <SENTENCE> Please do not use a continuous span to translate the <LANGUAGE> idiom appeared in the sentence.
DISCONTINUOUS 2	Please translate the following sentence from <LANGUAGE> to English: <SENTENCE> Please do not use a multi-word expression to translate the <LANGUAGE> idiom appeared in the sentence.

Table 6: The prompts we used for zero-shot translation.

Name	Prompt
FEW-SHOT	Please translate the following sentences from <LANGUAGE> to English: <LANGUAGE>: <SOURCE 1> English: <TARGET 1> <LANGUAGE>: <SOURCE 2> English: <TARGET 2> <LANGUAGE>: <SOURCE 3> English: <TARGET 3> <LANGUAGE>: <SOURCE 4> English: <TARGET 4> <LANGUAGE>: <SOURCE 5> English: <TARGET 5> <LANGUAGE>: <SENTENCE> English:
FEW-SHOT CREATIVELY	Please creatively translate the following sentences from <LANGUAGE> to English: <LANGUAGE>: <SOURCE 1> English: <TARGET 1> <LANGUAGE>: <SOURCE 2> English: <TARGET 2> <LANGUAGE>: <SOURCE 3> English: <TARGET 3> <LANGUAGE>: <SOURCE 4> English: <TARGET 4> <LANGUAGE>: <SOURCE 5> English: <TARGET 5> <LANGUAGE>: <SENTENCE> English:

Table 7: The prompts we used for few-shot translation.

B.6 Span Extraction

For span extraction, we used the following prompt:

- Given the English translation of the <LANGUAGE> sentence, please only output the span that corresponds to the <LANGUAGE> idiom.
  <LANGUAGE> sentence: <SOURCE>
  English translation: <TARGET>
  <LANGUAGE> idiom: <IDIOM>
  Span:

We only tested the prompt on Chinese sentences.

B.7 Automatic Evaluation

For automatic evaluation of faithfulness, we used the following prompt:

- Please rate the faithfulness of the following idiom translation within a sentence.
  Idiom to be translated: <IDIOM>
  Original sentence containing this idiom: <SOURCE>
  Translation: <TARGET>
  Your faithfulness rating should be a score from 1 to 5, where 1 is not faithful at all and 5 is perfectly faithful. Return a single number as your rating.

For automatic evaluation of creativity, we used the following prompt:

- Please rate the creativity of the following idiom translation within a sentence.
  Idiom to be translated: <IDIOM>
  Original sentence containing this idiom: <SOURCE>
  Translation: <TARGET>
  Your creativity rating should be a score from 1 to 5, where 1 is not creative at all (just plain language) and 5 is perfectly creative. Return a single number as your rating.

C Cleaning and Parsing

In this section, we list the details for cleaning and parsing the model output.

C.1 Sentences

GPT-4 failed to generate sentences for very few idioms (4 out of all idioms). In these cases, GPT-4 was unable to identify the given idiom as a real word. Interestingly, this happened for real idioms, but not plausible idioms. For convenience of implementation, we save 10 empty sentences to the file when sentences are not generated.

Another failure pattern was that GPT-4 failed to include the idiom in the sentence. In these cases,

Name	Prompt
BASELINE	Please translate the following paragraph from <LANGUAGE> to English. <PARAGRAPH>
FAITHFUL SIMPLE	Please translate the following paragraph from <LANGUAGE> to English. Please translate the idiom <IDIOM> faithfully. Do not explain. <PARAGRAPH>
CREATIVE SIMPLE	Please translate the following paragraph from <LANGUAGE> to English. Please translate the idiom <IDIOM> creatively. Do not explain. <PARAGRAPH>
THEME SIMPLE	Please translate the following paragraph from <LANGUAGE> to English. Please translate the idiom <IDIOM> in a way that matches the theme. Do not explain. <PARAGRAPH>
FAITHFUL COT	Please translate the following paragraph from <LANGUAGE> to English. Please translate the idiom <IDIOM> faithfully. Do not explain. <PARAGRAPH> Please follow the instructions below: <FAITHFUL INSTRUCTIONS>
CREATIVE COT	Please translate the following paragraph from <LANGUAGE> to English. Please translate the idiom <IDIOM> creatively. Do not explain. <PARAGRAPH> Please follow the instructions below: <CREATIVE INSTRUCTIONS>
THEME COT	Please translate the following paragraph from <LANGUAGE> to English. Please translate the idiom <IDIOM> in a way that matches the theme. Do not explain. <PARAGRAPH> Please follow the instructions below: <THEME INSTRUCTIONS>
FAITHFUL MULTI-TURN	Please translate the following paragraph from <LANGUAGE> to English. Please translate the idiom <IDIOM> faithfully. Do not explain. <PARAGRAPH>
CREATIVE MULTI-TURN	Could you provide an alternative translation of the paragraph, where the idiom is translated more creatively? The translation you provided has been widely used elsewhere.
THEME MULTI-TURN	Could you provide an alternative translation of the paragraph, where the idiom is translated with words that better match the context? The translation you provided can be used verbatim in a different context.

Table 8: The prompts we used for paragraph translation.

GPT-4 split the idiom into two parts, or used synonyms to represent the meaning of the idiom. The statistics for all languages are listed in Table 9.

C.2 Translations (Sentences)

In general, the output translations are clean. In the cases where GPT-4 was prompted to generate multiple translations in a single response, we parsed the response to get the list of translations. For the succinct prompt, GPT-4 tended to provide an explanation, which we removed from the response to get the clean translation.

C.3 Scores

GPT-3.5 generated scores in different formats, including different prefixes and suffixes. We observed that all the irrelevant output can be cleaned by taking the first digit appearing in the response string as the score.

C.4 Auto-CoT Instructions

We repeatedly asked GPT-4 for translation instructions. Given the same prompt, GPT-4 returned different steps. The number of steps ranged from 6 to 8. The name and the description of each step also differed. However, each different set of steps from a single response was reasonable. Hence, for each combination of aspect and language, we only used one response as the Auto-CoT instruction.

C.5 Paragraphs

GPT-4 sometimes does not include the sentence verbatim in the paragraph. This is due to punctuation and language-specific phenomena, such as conjugation in Japanese and Korean. However, in most cases, the paragraph contains the idiom. The statistics for all languages are listed in Table 10.

C.6 Translations (Paragraphs)

No noise was observed.

C.7 Spans

GPT-4 was able to locate precisely the span in the translated sentence that corresponds to the idiom in the original sentence. The identified span is a substring of the translated sentence for 1994 out of 2000 Chinese-English sentence pairs, translated using the optimal strategies. The few failures were due to the change in capitalization and punctuation.

D More Results

In this section, we show more examples from our results. No spans are highlighted for these examples, as we did not perform manual labeling and did not run the span queries (Appendix B.6) for these translations. The total number of sentence translations is 13,500 for each language. For Chinese, we further generate 20,000 translations with Pareto-optimal strategies. For each translation, we generate 5 faithfulness and 5 creativity scores.

D.1 Sentence Translations

We show more translations for all languages we used in Tables 11 (Chinese), 12 (Japanese), 13 (Korean), and 14 (plausible Chinese). All other translations can be found in the published data. The total number of paragraph translations is 800 for each language.

D.2 Paragraph Translations

For the paragraph pipeline, we show more examples in Tables 15 (Chinese), 16 (Japanese), 17, and 18 (plausible Chinese). The paragraphs were generated with the same sentences in Appendix D.1. All shown examples are in the genre “a news report" (Section 2), while examples in the other genre can be found in the published data.

E Scores

We summarize the faithfulness and creativity scores for all languages and all strategies in Table 19. We also visualize the scores using a heat map in Figure 3 and scatter plots in Figure 4.

Language	# Idioms	# Sentences	# w/ Idiom	% w/ Idiom
Chinese	4310	43100	42873	99.71
Japanese	2440	24400	24247	99.37
Korean	2316	23160	22943	99.06
Plausible Chinese	50	500	495	99.00

Table 9: The statistics for noise in sentence generation.

Language	# Paragraphs	# w/ Idiom	% w/ Idiom	# w/ Sentence	% w/ Sentence
Chinese	80	80	100.00	51	63.75
Japanese	80	79	98.75	37	46.25
Korean	80	78	97.50	36	45.00
Plausible Chinese	80	73	91.25	56	70.00

Table 10: The statistics for noise in paragraph generation.

Method	Sentence	Faithfulness	Creativity
ORIGINAL	他们通过威逼利诱，想要我放弃诉讼。	N/A	N/A
GOOGLE	They used threats and inducements to try to get me to drop the lawsuit.	4.8	3.8
DEEPL	They wanted me to drop the lawsuit through bullying.	2.2	2.6
BASELINE	They tried to get me to drop the lawsuit through threats and inducements.	4.2	3.8
DIVERSITY EXPLICIT	Through intimidation and temptation, they want me to give up the lawsuit.	5.0	3.6
DIVERSITY EXPLICIT	They employed threats and bribes, trying to get me to abandon the lawsuit.	4.4	4.0
DIVERSITY EXPLICIT	By wielding both threats and incentives, they aim to make me drop the lawsuit.	4.4	4.0
DIVERSITY EXPLICIT	They attempted to make me forsake the lawsuit by means of coercion and enticement.	4.4	3.2
DIVERSITY EXPLICIT	They're using a mix of scare tactics and promises in hopes I'll relinquish the lawsuit.	4.0	4.2
DIVERSITY DIALOG	They are using a combination of threats and rewards to persuade me to withdraw from the lawsuit.	4.6	4.0
DIVERSITY DIALOG	With a strategy of intimidation and allure, they seek to have me forego the lawsuit.	4.0	4.0
DIVERSITY DIALOG	They're applying pressure and offering inducements, hoping I will abandon my legal action.	4.8	4.0
DIVERSITY DIALOG	By employing coercion and temptation, their goal is for me to give up my lawsuit.	4.2	3.0
DIVERSITY DIALOG	Through both menacing and luring tactics, they wish for me to relinquish the lawsuit.	4.4	4.0
ZERO-SHOT CREATIVELY	Through a mix of threats and temptations, they sought to make me abandon my lawsuit.	4.6	4.0
CONTEXT EXPLICIT	They tried to make me give up the lawsuit through threats and bribes.	4.4	2.8
CONTEXT EXPLICIT	They, through intimidation and inducement, want me to give up the lawsuit.	4.6	3.0
CONTEXT EXPLICIT	They tried to make me give up the lawsuit through threats and bribes.	4.6	3.0
CONTEXT EXPLICIT	They tried to make me give up the lawsuit through intimidation and bribery.	4.6	3.2
ANALOGY NATURAL	They tried to make me drop the lawsuit through a carrot and stick approach.	4.0	4.0
ANALOGY CREATIVE	They tried to make me drop the lawsuit by dangling carrots laced with threats, like offering honey wrapped in thorns.	3.6	4.2
SHUFFLE ORDER	They tried to get me to drop the lawsuit by using threats and bribes.	4.0	3.8
SUCCINCT	They, through threats and bribes, want me to give up the lawsuit.	4.2	3.0
TWO-STEP	They used threats and temptations, trying to make me give up the lawsuit.	4.0	3.6
DISCONTINUOUS 1	They, through threats and promises, want me to give up the lawsuit.	4.8	3.0
DISCONTINUOUS 2	They want me to give up the lawsuit by coercing and bribing.	4.0	2.8
FEW-SHOT	They tried to make me give up the lawsuit through threats and bribes.	4.6	3.2
FEW-SHOT CREATIVELY	They attempted to sway my resolve through a mix of threats and temptations, hoping I would abandon the lawsuit.	5.0	4.0

Table 11: The example translations for Chinese.

Method	Sentence	Faithfulness	Creativity
ORIGINAL	私たちの小さなプロジェクトが、まさに燎原之火となり、急速に広がりました	N/A	N/A
GOOGLE	Our small project became a wildfire and spread rapidly.	4.2	4.0
DEEPL	Our little project became a veritable bonfire and spread rapidly.	2.8	4.0
BASELINE	Our small project, just like a spark in a field, rapidly spread far and wide.	3.4	4.2
DIVERSITY EXPLICIT	Our small project has indeed become a spark that quickly spread like wildfire.	4.4	4.2
DIVERSITY EXPLICIT	Our modest project has truly become a blaze that rapidly expanded.	3.6	3.8
DIVERSITY EXPLICIT	The little project of ours has indeed turned into a raging fire, spreading swiftly.	4.0	4.0
DIVERSITY EXPLICIT	Our small project has indeed become a prairie fire, spreading rapidly.	4.8	4.2
DIVERSITY EXPLICIT	Our minor project quickly became a spark that turned into a widespread inferno.	2.8	4.0
DIVERSITY DIALOG	Our little project has truly blossomed into a widespread blaze, spreading quickly.	2.8	3.8
DIVERSITY DIALOG	Our small initiative indeed became a conflagration that rapidly expanded.	4.4	4.0
DIVERSITY DIALOG	The small project we started has really caught on like wildfire, spreading rapidly.	4.4	4.0
DIVERSITY DIALOG	Our humble project turned into a blazing fire, quickly covering ground.	4.4	4.0
DIVERSITY DIALOG	Our tiny project has, indeed, turned into a spark that has rapidly spread far and wide.	3.6	4.0
ZERO-SHOT CREATIVELY	Our humble project, akin to a spark in a dry field, swiftly ignited, spreading far and wide with rapid fervor.	4.4	4.2
CONTEXT EXPLICIT	Our small project has indeed become a spark that quickly spread like wildfire.	4.2	4.2
CONTEXT EXPLICIT	Our small project became like a spark in a prairie, spreading rapidly.	3.2	4.0
CONTEXT EXPLICIT	Our small project has indeed become a spark that ignites the plains, spreading rapidly.	3.2	4.2
CONTEXT EXPLICIT	Our small project became like a spark in a prairie, spreading rapidly.	3.4	4.0
ANALOGY NATURAL	Our little project caught fire like wildfire and spread rapidly.	3.6	4.2
ANALOGY CREATIVE	Our modest project, just like a spark in a sea of whispers, quickly spread far and wide.	1.6	4.2
SHUFFLE ORDER	Our small project has indeed become a blaze that quickly spread far and wide.	4.0	3.6
SUCCINCT	Our small project has indeed become a spark that started a prairie fire, spreading rapidly.	5.0	4.4
TWO-STEP	Our small project really spread rapidly.	1.8	2.0
DISCONTINUOUS 1	Our small project, just like a spark in a field, rapidly spread far and wide.	3.4	4.2
DISCONTINUOUS 2	Our small project has indeed become a wildfire and spread rapidly.	4.2	3.8
FEW-SHOT	Our small project became like a wildfire, spreading rapidly.	4.6	4.0
FEW-SHOT CREATIVELY	Our modest project became a wildfire, spreading rapidly far and wide.	4.0	4.0

Table 12: The example translations for Japanese.

Method	Sentence	Faithfulness	Creativity
ORIGINAL	진짜 그 사람이 싫어서 안 만나는 건지, 아니면 담인인지 모르겠어.	N/A	N/A
GOOGLE	I don't know if I'm not dating that person because I really hate him, or if he's just a temperamental person.	2.2	3.6
DEEPL	I don't know if I'm not seeing him because I really don't like him, or if he's a dummy.	2.0	3.8
BASELINE	I'm not sure if it's really because I dislike that person that I don't meet them, or if it's because I'm just busy.	3.2	3.0
DIVERSITY EXPLICIT	I'm not sure if it's because I really dislike that person or if I'm just being indifferent.	3.6	3.8
DIVERSITY EXPLICIT	I can't tell if I'm not meeting them because I really hate them, or if I'm just unbothered.	3.6	4.0
DIVERSITY EXPLICIT	It's unclear whether I'm avoiding them because I genuinely don't like them, or if I'm just being nonchalant.	3.0	4.0
DIVERSITY EXPLICIT	I'm uncertain if my not seeing them is due to real distaste or simply indifference.	4.0	4.0
DIVERSITY EXPLICIT	It's hard to say whether my avoidance is because of genuine dislike or if I am just indifferent.	4.0	4.0
DIVERSITY DIALOG	I can't figure out if my refusal to meet is because I truly detest that person, or if it's just apathy.	4.2	3.6
DIVERSITY DIALOG	I'm unsure whether my avoidance is rooted in real animosity towards that person or mere indifference.	4.4	4.0
DIVERSITY DIALOG	It's difficult to determine if I'm not seeing them because I genuinely loathe them, or if I'm just detached.	3.6	4.0
DIVERSITY DIALOG	I'm puzzled whether my lack of desire to meet stems from actual aversion or simply from being disinterested.	3.8	4.2
DIVERSITY DIALOG	I find it challenging to discern if my avoidance is due to a genuine dislike or if it's just a matter of being indifferent.	4.2	4.0
ZERO-SHOT CREATIVELY	I can't tell if they're genuinely avoiding that person because they dislike them, or if it's because they're shy.	3.8	4.0
CONTEXT EXPLICIT	I'm not sure if it's because they really dislike that person that they're not meeting, or if it's because they're just playing hard to get.	4.0	4.2
CONTEXT EXPLICIT	I can't tell if it's because they really dislike that person that they're not meeting, or if it's because they're playing hard to get.	3.8	3.8
CONTEXT EXPLICIT	I can't tell if they're really avoiding meeting because they dislike that person, or if they're just busy.	3.6	3.4
CONTEXT EXPLICIT	I'm not sure if it's because they really dislike that person or if it's because they are just busy.	3.6	3.6
ANALOGY NATURAL	"I can't tell if they're really avoiding someone because they dislike them, or if they're just playing hard to get."	3.4	4.0
ANALOGY CREATIVE	"I can't tell if they're really avoiding meeting because they dislike that person, or if it's like trying to decipher the flavor of a mystery soda."	2.2	4.4
SHUFFLE ORDER	I can't tell if they're really avoiding meeting because they dislike that person, or if they're just busy.	3.8	3.4
SUCCINCT	I can't tell if it's really because they dislike that person that they're not meeting up, or if they're just playing hard to get.	4.0	4.0
TWO-STEP	I don't know if it's really because I dislike that person that I'm not meeting them, or if there is another reason.	4.2	3.2
DISCONTINUOUS 1	I can't tell if they're really not meeting because they dislike that person, or if it's because they're playing hard to get.	3.8	4.0
DISCONTINUOUS 2	I'm not sure if it's really because I don't like that person that I'm not meeting them, or if it's something else.	4.2	3.6
FEW-SHOT	I can't tell if they're not meeting because they genuinely dislike the person or if it's just an excuse.	4.0	4.0
FEW-SHOT CREATIVELY	I can't tell if it's genuinely because they dislike that person that they don't meet up, or if it's just an excuse.	4.0	3.8

Table 13: The example translations for Korean.

Method	Sentence	Faithfulness	Creativity
ORIGINAL	每年的春天，园子里春荷盛开，秋天则是秋菊竞艳。	N/A	N/A
GOOGLE	Every spring, lotus flowers bloom in the garden, and in autumn, chrysanthemums bloom.	3.2	2.2
DEEPL	Every spring, the garden is in full bloom with spring lotus, and in the fall, fall chrysanthemums compete.	4.2	4.0
BASELINE	Every spring, the garden is filled with blooming spring lotuses, while in autumn, it's the chrysanthemums that vie for beauty.	4.0	4.0
DIVERSITY EXPLICIT	Every spring, the garden blooms with spring lotuses, while in autumn, the chrysanthemums vie for beauty.	4.2	3.6
DIVERSITY EXPLICIT	In the garden, spring sees the lush blossoming of lotuses, and autumn showcases the competitive bloom of chrysanthemums.	4.0	4.0
DIVERSITY EXPLICIT	Yearly, spring brings the luxuriant opening of lotuses in the garden, and autumn displays the dazzling chrysanthemums.	3.2	3.8
DIVERSITY EXPLICIT	Every year, the springtime garden is filled with blooming lotuses, and autumn with the splendid display of chrysanthemums.	3.6	3.8
DIVERSITY EXPLICIT	The garden annually celebrates spring with the flourishing of lotuses, and autumn with the striking rivalry of chrysanthemums.	3.8	4.0
DIVERSITY DIALOG	Each spring, the garden is adorned with burgeoning lotuses, while autumn brings forth the competing beauty of chrysanthemums.	4.4	4.0
DIVERSITY DIALOG	Annually, the spring season witnesses the garden's lotuses in full bloom, followed by the chrysanthemums' vibrant display in autumn.	4.0	4.2
DIVERSITY DIALOG	The garden experiences a bounty of spring lotuses every year, and in autumn, the chrysanthemums burst into competitive radiance.	3.6	3.8
DIVERSITY DIALOG	Every spring, lotuses thrive in the garden, with autumn presenting a spectacle of chrysanthemums competing in splendor.	3.8	4.0
DIVERSITY DIALOG	In the garden, each year, spring is marked by the blossoming of lotuses and autumn by the riotous beauty of chrysanthemums in competition.	4.4	4.0
ZERO-SHOT CREATIVELY	Every spring, the garden blooms with vibrant spring lotus, while in autumn, the chrysanthemums vie in beauty.	3.8	4.0
CONTEXT EXPLICIT	Every spring, the lotus blooms abundantly in the garden, while in autumn, the chrysanthemums vie in beauty.	4.0	4.0
CONTEXT EXPLICIT	Every spring, the lotus blooms abundantly in the garden, while in autumn, the chrysanthemums compete in beauty.	4.0	4.0
CONTEXT EXPLICIT	Every spring, the lotus blooms abundantly in the garden, while in autumn, the chrysanthemums compete in beauty.	4.0	4.0
CONTEXT EXPLICIT	Every spring, the lotus blooms profusely in the garden, while in autumn, the chrysanthemums vie in beauty.	3.6	3.8
ANALOGY NATURAL	Every spring, the garden blooms with spring lotuses, much like a sea of blossoms, and in autumn, autumn chrysanthemums vie in beauty, akin to a painter's vibrant palette.	3.8	4.2
ANALOGY CREATIVE	In the springtime each year, the garden blossoms into a sea of spring lotuses, while in autumn, it transforms into a stage where autumn chrysanthemums vie for the spotlight.	3.6	4.2
SHUFFLE ORDER	Every spring, the lotus blossoms flourish in the garden, while in autumn, the chrysanthemums vie in beauty.	4.0	4.0
SUCCINCT	Every spring, the lotus flowers bloom in abundance in the garden, and in autumn, the chrysanthemums vie in beauty.	3.6	3.8
TWO-STEP	Every spring, the lotuses in the garden bloom in abundance; come autumn, it is the chrysanthemums' turn to bloom competitively.	3.6	4.0
DISCONTINUOUS 1	Every spring, the garden is in full bloom with spring lotuses, and in autumn, autumn chrysanthemums vie in beauty.	3.6	3.8
DISCONTINUOUS 2	Every spring, the garden is full of blooming lotuses, and in autumn, autumn chrysanthemums vie in beauty.	3.6	4.0
FEW-SHOT	Every spring, the garden blossoms with spring lotuses, while in autumn, the autumn chrysanthemums vie in beauty.	4.0	4.0
FEW-SHOT CREATIVELY	Every spring, the garden is a spectacle of blooming spring lotuses, while autumn boasts a competitive display of vibrant chrysanthemums.	4.0	4.2

Table 14: The example translations for plausible Chinese.

Method	Paragraph
ORIGINAL	在今日下午的新闻发布会上，被告李先生向媒体披露了一起备受关注的诉讼案件的最新进展。据李先生称，“他们通过威逼利诱，想要我放弃诉讼。”他详细说明了对方如何施加压力，试图迫使其撤销对公司不正当行为的指控。这起事件引起了公众的广泛关注，许多人表示，像李先生这样勇敢站出来揭露真相的行为，值得社会给予更多的支持和保护。
BASELINE	At this afternoon's press conference, Mr. Li, the defendant, disclosed the latest developments in a highly publicized lawsuit case to the media. According to Mr. Li, "They tried to make me give up the lawsuit through threats and bribes." He detailed how the other party applied pressure to force him to withdraw his accusations of the company's misconduct. This incident has garnered widespread public attention, with many people expressing that acts of bravery like Mr. Li's in exposing the truth deserve more societal support and protection.
FAITHFUL SIMPLE	At the press conference this afternoon, Mr. Li, the defendant, disclosed to the media the latest developments in a highly publicized lawsuit. According to Mr. Li, "They tried to make me give up the lawsuit through coercion and bribery." He detailed how the other party applied pressure, attempting to force him to withdraw his allegations of impropriety against the company. This incident has attracted widespread public attention, with many people stating that brave actions like Mr. Li's, in stepping forward to reveal the truth, deserve more societal support and protection.
CREATIVE SIMPLE	At this afternoon's press conference, Mr. Li, the defendant, disclosed to the media the latest developments in a highly publicized lawsuit. According to Mr. Li, "They tried to sway me with a mix of threats and temptations, wanting me to drop the lawsuit." He elaborated on how the other party applied pressure, attempting to force him to withdraw his accusations of improper conduct by the company. This incident has sparked widespread public interest, with many expressing that courageous acts like Mr. Li's in revealing the truth deserve more societal support and protection.
THEME SIMPLE	At this afternoon's press conference, Mr. Li, the defendant, revealed to the media the latest developments in a lawsuit that has garnered significant attention. According to Mr. Li, "They tried to strong-arm and entice me into dropping the lawsuit." He detailed how the other party applied pressure in an attempt to force him to withdraw his accusations of improper conduct by the company. This incident has sparked widespread public interest, with many expressing that courageous acts like Mr. Li's in exposing the truth deserve more societal support and protection.
FAITHFUL COT	At this afternoon's press conference, Mr. Li, the defendant, disclosed to the media the latest developments in a highly watched lawsuit. According to Mr. Li, "They tried to make me give up the lawsuit through threats and bribes." He detailed how the other party attempted to pressure him into withdrawing his allegations of the company's improper conduct. This incident has garnered widespread public attention, and many people believe that acts of bravery like Mr. Li's, in coming forward to reveal the truth, deserve more societal support and protection.
CREATIVE COT	At this afternoon's press conference, Mr. Li, the defendant, unveiled the latest developments in a highly anticipated lawsuit to the media. Mr. Li claimed, "They resorted to a mix of intimidation and enticement, aiming for me to drop the lawsuit." He elaborated on how the other party applied pressure, attempting to coerce him into retracting his allegations of misconduct against the company. This incident has drawn widespread public attention, with many voicing that courageous actions like Mr. Li's, stepping forward to expose the truth, deserve greater societal support and protection.
THEME COT	At the press conference this afternoon, Mr. Li, the defendant, disclosed to the media the latest developments in a highly publicized lawsuit. According to Mr. Li, ‘Ä‘ÍJThey tried to make me drop the lawsuit through a combination of threats and inducements.‘Ä‘ÍHe detailed how the other party applied pressure, attempting to force him to withdraw his allegations of misconduct against the company. This incident has attracted widespread public attention, with many people expressing that courageous actions like Mr. Li's, in revealing the truth, deserve more support and protection from society.
FAITHFUL MULTI-TURN	At this afternoon's press conference, Mr. Li, the defendant, disclosed to the media the latest developments in a highly publicized lawsuit. According to Mr. Li, "They tried to make me give up the lawsuit through coercion and bribery." He detailed how the other party applied pressure, attempting to force him to withdraw his allegations of impropriety against the company. This incident has attracted widespread public attention, with many people stating that brave actions like Mr. Li's, in stepping forward to reveal the truth, deserve more support and protection from society.
CREATIVE MULTI-TURN	At the news conference this afternoon, Mr. Li, the accused, unveiled to the press the latest developments in a closely watched lawsuit. Mr. Li claimed, "They attempted to sway me into dropping the lawsuit by a combination of threats and temptations." He elaborated on how the opposition exerted pressure, trying to compel him to rescind his accusations of misconduct against the company. The saga has captivated public interest, with many voicing that courageous disclosures like Mr. Li's merit greater societal endorsement and safeguarding.
THEME MULTI-TURN	In this afternoon's press conference, Mr. Li, the defendant, shared with the media the recent developments of a lawsuit that has captured widespread attention. Mr. Li stated, "They resorted to threats and allurements in an effort to make me abandon the lawsuit." He went into detail about how the other side applied pressure, aiming to make him withdraw his claims against the company's wrongful actions. The incident has sparked considerable public interest, with many expressing that the courageous act of coming forward like Mr. Li has done should be met with greater societal support and protection.

Table 15: The example paragraph translations for Chinese.

Method	Paragraph
ORIGINAL	日本、東京 - 地元の起業家たちが始めた画期的な取り組みが話題になっています。彼らの開発した環境に優しい再生可能エネルギー製品が予想外の成功を収め、大きな注目を集めています。「私たちの小さなプロジェクトが、まさに燎原之火となり、急速に広がりました。」と、プロジェクトリーダーは興奮を隠せません。この成功は、持続可能な開発と環境保護に関心が集まっている今、他の起業家や企業にも大きな刺激を与えています。
BASELINE	Japan, Tokyo - A groundbreaking initiative started by local entrepreneurs is gaining attention. Their development of environmentally friendly renewable energy products has achieved unexpected success and is drawing significant attention. “Our small project has truly become a blazing wildfire, spreading rapidly," says the project leader, unable to hide their excitement. This success is providing a major stimulus to other entrepreneurs and companies at a time when there is a growing focus on sustainable development and environmental protection.
FAITHFUL SIMPLE	Japan, Tokyo - An innovative initiative started by local entrepreneurs is making waves. Their development of eco-friendly renewable energy products has achieved unexpected success and garnered significant attention. “Our small project has indeed become a wildfire, spreading rapidly," the project leader could hardly contain his excitement. This success is providing a major stimulus to other entrepreneurs and companies at a time when interest in sustainable development and environmental protection is growing.
CREATIVE SIMPLE	Tokyo, Japan - A groundbreaking initiative started by local entrepreneurs is becoming a hot topic. Their development of environmentally friendly renewable energy products has achieved unexpected success and has drawn significant attention. “Our small project has, quite literally, become a wildfire, spreading rapidly," says the project leader, unable to hide his excitement. This success is providing great motivation to other entrepreneurs and corporations at a time when interest in sustainable development and environmental protection is growing.
THEME SIMPLE	Japan, Tokyo - An innovative initiative started by local entrepreneurs is becoming a hot topic. Their development of eco-friendly renewable energy products has achieved unexpected success and attracted significant attention. “Our small project has, indeed, spread like wildfire, rapidly expanding," the project leader said, unable to hide his excitement. This success is providing a substantial stimulus to other entrepreneurs and companies at a time when there is a growing focus on sustainable development and environmental protection.
FAITHFUL COT	Japan, Tokyo - A groundbreaking initiative started by local entrepreneurs is becoming a topic of conversation. Their developed environmentally friendly renewable energy products have achieved unexpected success and are attracting significant attention. “Our small project has indeed become a raging wildfire, spreading rapidly," says the project leader, unable to hide his excitement. This success is providing a substantial stimulus to other entrepreneurs and companies at a time when there is a strong focus on sustainable development and environmental protection.
CREATIVE COT	Tokyo, Japan - A groundbreaking initiative started by local entrepreneurs is making waves. Their development of environmentally friendly renewable energy products has achieved unexpected success and garnered significant attention. “Our small project has truly become a wildfire, spreading rapidly," says the project leader, unable to hide their excitement. This success has given a substantial boost to other entrepreneurs and companies at a time when there's a growing focus on sustainable development and environmental protection.
THEME COT	Japan, Tokyo - A groundbreaking initiative started by local entrepreneurs is making headlines. Their development of environmentally friendly renewable energy products has achieved unexpected success and garnered significant attention. “Our small project has indeed become a wildfire, spreading rapidly," the project leader could not hide their excitement. This success provides a major stimulus to other entrepreneurs and companies at a time when there is growing interest in sustainable development and environmental protection.
FAITHFUL MULTI-TURN	Japan, Tokyo - An innovative initiative started by local entrepreneurs is gaining attention. Their development of eco-friendly renewable energy products has achieved unexpected success and garnered significant interest. “Our small project has indeed become a wildfire, spreading rapidly," the project leader could hardly contain his excitement. This success is providing a major stimulus to other entrepreneurs and companies at a time when interest in sustainable development and environmental protection is growing.
CREATIVE MULTI-TURN	Japan, Tokyo - An innovative initiative started by local entrepreneurs is making waves. Their development of eco-friendly renewable energy products has unexpectedly soared to success, capturing widespread attention. “Our modest project has, quite literally, turned into a blaze across the prairie, expanding swiftly," the project leader is elated. This triumph is giving a significant boost to other entrepreneurs and corporations at a moment when the spotlight is on sustainable development and conservation of the environment.
THEME MULTI-TURN	Japan, Tokyo - A groundbreaking initiative launched by local entrepreneurs is creating a buzz. Their invention of eco-friendly renewable energy products has manifested into an unexpected triumph, attracting major interest. “Our humble project has indeed sparked a conflagration, spreading at an incredible pace," the project leader is visibly thrilled. This success is fueling a significant impetus among other entrepreneurs and businesses at a time when the focus on sustainable development and environmental preservation is intensifying.

Table 16: The example paragraph translations for Japanese.

Method	Paragraph
ORIGINAL	서울, 대한민국 - 지역 사회에서는 최근 한 개인의 대인 관계 문제가 관심사로 떠올랐습니다. 주변 사람들은 “진짜 그 사람이 싫어서 안 만나는 건지, 아니면 다른 진짜 이유가 있는 것인지 모르겠어” 라고 말하며, 이 문제에 대한 의견을 나누고 있습니다. 이 문제는 한 사람의 갑작스런 사회적 활동 감소와 그 이유에 대한 추측으로 인해 생겨났습니다. 해당 개인이나 그의 친구들은 현재 이와 관련된 공식적인 입장을 밝히지 않고 있으며, 이로 인해 더욱 많은 추측이 이어지고 있습니다. 지역사회는 이 문제에 대해 더 많은 정보를 원하며, 이에 대한 진실을 알기 위해 관심을 기울이고 있습니다.
BASELINE	Seoul, South Korea - A personal relationship issue of an individual has recently become a matter of interest in the community. People around are sharing their opinions, saying, “I’m not sure if they’re really not meeting because they dislike that person, or if there’s another real reason." This issue arose from speculation about the sudden decrease in one person’s social activities and the reasons behind it. Neither the individual in question nor their friends have made an official statement regarding the matter, leading to even more speculation. The community is seeking more information on this issue and is keen to find out the truth.
FAITHFUL SIMPLE	Seoul, South Korea - In the local community, the personal relationship issue of an individual has recently become a matter of interest. People around are saying, “I don’t know if it’s really because they dislike that person and don’t want to meet, or if there’s another real reason," and are sharing their opinions on this issue. This issue arose due to a sudden decrease in one person’s social activities and speculation about the reason. The individual in question or their friends have not made any official statements regarding this matter, leading to further speculation. The community wants more information about this issue and is keen to find out the truth.
CREATIVE SIMPLE	Seoul, South Korea - In the local community, the personal relationships of an individual have recently become a matter of interest. People around are saying, “I wonder if it’s really because they dislike that person or if there’s another real reason," as they share their opinions on this matter. The issue emerged due to a sudden decrease in social activities by the individual and speculation about the reasons behind it. Neither the individual in question nor their friends have made any official statement regarding this, leading to further speculation. The community is eager for more information on the matter and is paying close attention in hopes of uncovering the truth.
THEME SIMPLE	Seoul, South Korea - A personal relationship issue of an individual has recently become a topic of interest in the community. People around are saying, “I can’t tell if they genuinely dislike the person and that’s why they’re not meeting, or if there’s another real reason," and are sharing their opinions on this matter. This issue arose from sudden decreases in one person’s social activities and speculation about the reasons behind it. Neither the individual in question nor their friends have made any official statement on the matter, leading to further speculation. The community wants more information on this issue and is keen to uncover the truth.
FAITHFUL COT	Seoul, South Korea - In the local community, the interpersonal relationship issue of an individual has recently arisen as a topic of interest. People around them are saying, ‘Ä‘IJI don’t know if they are not meeting them because they really dislike the person, or if there’s another real reason,‘Ä‘Ísharing their opinions on this matter. This issue sprung up due to the sudden decrease in social activities of one person and the speculation about the reasons behind it. The individual in question or their friends have not officially addressed the matter, leading to even more speculation. The community wants more information on this issue and is paying attention to uncover the truth.
CREATIVE COT	Seoul, South Korea - In the local community, the interpersonal issues of an individual have recently become a hot topic. People around are saying, “I can’t tell if they’re actually giving the cold shoulder, or if there’s some other real reason behind it," sharing their thoughts on the matter. This issue emerged due to a sudden decrease in social activities by an individual and the speculation around the reasons for such withdrawal. Neither the person in question nor their friends have made any official statements regarding the situation, leading to even more speculation. The community is eager for more information and is paying close attention in hopes of uncovering the truth.
THEME COT	Seoul, South Korea - A personal relationship issue of an individual has recently emerged as a concern within the community. People around are saying, “I don’t know whether they’re not meeting because they genuinely dislike each other, or if there’s some other real reason," sharing their opinions on the matter. This issue arose from speculation about the sudden decrease in social activities of a person and the reasons behind it. Neither the individual in question nor their friends have made any official statement regarding this, leading to even more speculation. The community seeks more information on this issue, paying close attention to uncover the truth.
FAITHFUL MULTI-TURN	Seoul, South Korea - In the local community, the personal relationship issue of an individual has recently become a matter of interest. People around are saying, “I don’t know if it’s really because they dislike the person and don’t want to meet, or if there’s another real reason," and are sharing their opinions on this issue. This issue arose due to a sudden decrease in one person’s social activities and speculation about the reason. The individual in question or their friends have not made any official statements regarding this matter, leading to further speculation. The community wants more information about this issue and is keen to find out the truth.
CREATIVE MULTI-TURN	Seoul, South Korea - A personal relationship dilemma involving an individual has recently captured the community’s attention. Neighbors are speculating, “Is it genuinely because they despise the person so much they avoid meeting, or is there another underlying truth?" This conversation is buzzing around as people exchange views on the matter. Sparked by one individual’s abrupt retreat from social engagements and the swirling guesses about its cause, this topic has become hot. Neither the person at the heart of the rumors nor their close associates have clarified their stance, fueling the rumor mill even more. The community is eager for more clarity on this issue, showing heightened interest to unearth the actual story.
THEME MULTI-TURN	Seoul, South Korea - In the local community, there’s been a rising curiosity about an individual’s inter-personal conflict. People around are left wondering, “Is it true aversion driving them apart or is there something deeper at play?" This speculation has grown from observing a notable drop in one’s social interactions, sparking conversations. Neither the individual at the center of these rumors nor their confidants have come forward with any explanations, leading to evolving conjecture. The community is showing a keen interest, hoping to peel back the layers of this mystery and uncover the truth.

Table 17: The example paragraph translations for Korean.

Method	Paragraph
ORIGINAL	【环球新闻网】随着季节的更迭，大自然总是有着它独特的方式来展现生命之美。在我国南方的一个古老园林中，这种自然的循环得到了完美的体现。每年的春天，园子里春荷盛开，犹如一片片翠绿的玉盘，静静浮在水面上，吸引了无数游人驻足欣赏。而到了秋天，则是秋菊竞艳，黄、紫、白交织的菊花在秋风中摇曳生姿，成为这古园一道别样的风景线。这一切似乎在提醒着人们，无论世事如何变迁，自然的美好总是值得我们去珍惜和维护。
BASELINE	[Global News Network] As seasons change, nature always has its unique way of displaying the beauty of life. In an ancient garden in the south of our country, this natural cycle is perfectly embodied. Every spring, the garden is full of blooming spring lotuses, resembling pieces of emerald discs quietly floating on the surface of the water, attracting countless visitors to stop and admire. By autumn, the chrysanthemums outshine each other in beauty; yellow, purple, and white chrysanthemums sway in the autumn breeze, becoming a distinct scenic line in this ancient garden. All this seems to remind us that, no matter how the world changes, the beauty of nature is always worth cherishing and protecting.
FAITHFUL SIMPLE	[Global News Network] With the changing of seasons, nature always has its unique way to display the beauty of life. In an ancient garden in the south of our country, this natural cycle is perfectly embodied. Every spring, the garden is full of spring lotus, like pieces of green jade discs, quietly floating on the water surface, attracting countless visitors to stop and admire. And in autumn, it's the time for the autumn chrysanthemums to compete in beauty, with yellow, purple, and white chrysanthemums swaying in the autumn breeze, becoming a different landscape in this ancient garden. All of this seems to remind people that no matter how the world changes, the beauty of nature is always worth our appreciation and preservation.
CREATIVE SIMPLE	[Global News Network] As seasons transition, nature always has its unique way of showcasing the beauty of life. In an ancient garden in the southern part of our country, this natural cycle is perfectly embodied. Each spring, the garden blooms with lotus, like emerald discs quietly floating on the water, attracting countless visitors to stop and admire. Come autumn, it's a competition of chrysanthemums' beauty, with yellow, purple, and white chrysanthemums swaying in the autumn breeze, forming a distinctive landscape in this ancient garden. All these seem to remind us that, no matter how the world changes, the beauty of nature is always worth cherishing and protecting.
THEME SIMPLE	[Global News Network] With the change of seasons, Nature always has its unique way of showcasing the beauty of life. In an ancient garden in the southern part of our country, this natural cycle is perfectly represented. Every spring, the garden is full of blooming lotus, like pieces of emerald green jade plates quietly floating on the water, attracting countless visitors to stop and admire. And when autumn comes, it's time for the chrysanthemums to compete in beauty, with yellow, purple, and white chrysanthemums swaying in the autumn breeze, becoming a unique landscape of this ancient garden. All this seems to remind us that no matter how the world changes, the beauty of nature is always worth cherishing and protecting.
FAITHFUL COT	[Global News Network] As the seasons change, nature always has its unique way of showcasing the beauty of life. This natural cycle is perfectly embodied in an ancient garden in the southern part of our country. Every spring, the garden is filled with spring lotuses, which bloom like green jade discs quietly floating on the water, attracting countless visitors to stop and admire. When autumn arrives, it's the autumn chrysanthemums that compete in beauty, with yellow, purple, and white chrysanthemums swaying in the autumn breeze, forming a distinct landscape in this ancient garden. All of this seems to remind people that no matter how the world changes, the beauty of nature is always worth our appreciation and preservation.
CREATIVE COT	**Global News Network** - As seasons change, nature always finds its unique way to display the beauty of life. In an ancient garden in the south of our country, this natural cycle is perfectly exemplified. Every spring, the garden is adorned with lotus blossoms, like myriad verdant jade discs quietly floating on the water, drawing countless visitors to pause and admire. Come autumn, it's the chrysanthemums' turn to shine, with yellow, purple, and white blooms dancing in the autumn breeze, creating a distinct scenic line in this ancient garden. All these seem to remind us that, no matter how the world changes, the splendor of nature is always worth our appreciation and preservation.
THEME COT	[Global News Network] As the seasons change, nature always has its unique way of showcasing the beauty of life. In an ancient garden in the southern part of our country, this natural cycle is perfectly embodied. In spring, the garden is filled with blooming lotuses, like green jade plates quietly floating on the water, attracting countless visitors to stop and admire. In autumn, it's the chrysanthemums' turn to dazzle, with yellow, purple, and white blooms swaying in the autumn breeze, presenting a different kind of scenery in the ancient garden. All of this seems to remind us that no matter how the world changes, the beauty of nature is always worth cherishing and protecting.
FAITHFUL MULTI-TURN	[Global News Network] As the seasons change, nature always has its unique way of showcasing the beauty of life. In an ancient garden in the south of our country, this natural cycle is perfectly embodied. Each spring, the garden blooms with spring lotus, like pieces of emerald green jade discs, quietly floating on the water's surface, attracting countless visitors to stop and admire. In autumn, it's the turn of the autumn chrysanthemums to compete in beauty, with yellow, purple, and white chrysanthemums swaying in the autumn breeze, becoming a distinctive landscape in this ancient garden. All of this seems to remind us that no matter how the world changes, the beauty of nature is always worth cherishing and preserving.
CREATIVE MULTI-TURN	[Global News Network] With the seasonal shifts, nature unfurls the splendor of life in its unique expressions. Within the verdant confines of an ancient garden in our country's south, this rhythmic cadence of nature is flawlessly manifested. Come spring, the garden is a spectacle of spring lotuses, mirroring a collection of emerald platters, serenely adrift upon the waters, luring myriad visitors to pause and revel in their beauty. Autumn heralds the reign of chrysanthemums, weaving a tapestry of yellow, purple, and white against the autumnal air, presenting a contrasting spectacle to the observer. Through this endless cycle of spring lotus and autumn chrysanthemum, nature subtly reminds us - amidst the ever-changing tapestry of life, its inherent beauty remains a constant, beckoning our appreciation and preservation.
THEME MULTI-TURN	[Global News Network] As the calendar pages turn, nature invariably reveals the splendor of existence in its distinct manner. In a venerable garden in the south of our nation, this perpetual dance of nature is exquisitely showcased. When spring arrives, the garden is awash with spring lotuses, resembling emerald plates, serenely floating on the water, enticing legions of visitors to halt and admire. Come autumn, the garden transitions to showcase the rival beauty of autumn chrysanthemums, with blooms of yellow, purple, and white swaying in the autumn breeze, crafting an alternative picturesque scene in this ancient park. This eternal sequence of lotus blossoms in spring and chrysanthemums in fall gently nudges us to remember that, despite the shifts and turns of worldly affairs, the beauty encapsulated within nature is always there, worthy of our protection and esteem.

Table 18: The example paragraph translations for plausible Chinese.

Method	Size	Chinese		Japanese		Korean		Plausible Chinese	
		Faithfulness	Creativity	Faithfulness	Creativity	Faithfulness	Creativity	Faithfulness	Creativity
GOOGLE	500	4.05 ± 0.76	3.43 ± 0.52	3.77 ± 0.96	3.43 ± 0.56	3.14 ± 1.00	3.15 ± 0.64	3.74 ± 0.86	3.59 ± 0.48
DEEPL	500	3.77 ± 1.00	3.46 ± 0.58	3.41 ± 1.06	3.40 ± 0.58	3.13 ± 1.01	3.38 ± 0.60	3.45 ± 1.00	3.66 ± 0.48
BASELINE	500	4.26 ± 0.59	3.70 ± 0.43	4.11 ± 0.73	3.63 ± 0.49	3.62 ± 0.83	3.46 ± 0.52	4.09 ± 0.67	3.84 ± 0.37
DIVERSITY EXPLICIT	2500	4.22 ± 0.53	3.86 ± 0.33	4.06 ± 0.63	3.78 ± 0.39	3.62 ± 0.77	3.68 ± 0.45	4.08 ± 0.57	3.95 ± 0.29
DIVERSITY DIALOG	2500	4.15 ± 0.51	3.94 ± 0.27	4.00 ± 0.59	3.87 ± 0.34	3.60 ± 0.75	3.79 ± 0.39	4.02 ± 0.55	4.02 ± 0.25
ZERO-SHOT CREATIVELY	500	4.30 ± 0.50	4.09 ± 0.23	4.16 ± 0.58	3.99 ± 0.27	3.76 ± 0.68	3.97 ± 0.33	4.21 ± 0.55	4.10 ± 0.28
CONTEXT EXPLICIT	2000	4.29 ± 0.58	3.73 ± 0.42	4.11 ± 0.70	3.62 ± 0.49	3.61 ± 0.81	3.48 ± 0.55	4.12 ± 0.65	3.86 ± 0.36
ANALOGY NATURAL	500	3.95 ± 0.70	4.06 ± 0.27	3.70 ± 0.74	4.03 ± 0.27	3.55 ± 0.72	3.97 ± 0.28	3.79 ± 0.78	4.06 ± 0.25
ANALOGY CREATIVE	500	3.19 ± 0.91	4.25 ± 0.31	3.07 ± 0.88	4.29 ± 0.32	2.80 ± 0.91	4.23 ± 0.35	3.10 ± 0.91	4.22 ± 0.29
SHUFFLE ORDER	500	4.29 ± 0.56	3.74 ± 0.40	4.16 ± 0.63	3.65 ± 0.47	3.65 ± 0.77	3.52 ± 0.49	4.14 ± 0.63	3.85 ± 0.36
SUCCINCT	500	4.21 ± 0.61	3.71 ± 0.44	4.10 ± 0.67	3.61 ± 0.49	3.66 ± 0.77	3.55 ± 0.48	4.05 ± 0.68	3.81 ± 0.38
TWO-STEP	500	3.89 ± 0.70	3.36 ± 0.57	3.68 ± 0.85	3.27 ± 0.57	3.26 ± 0.90	3.07 ± 0.61	3.71 ± 0.82	3.53 ± 0.52
DISCONTINUOUS 1	500	4.25 ± 0.60	3.74 ± 0.39	4.09 ± 0.67	3.63 ± 0.47	3.56 ± 0.85	3.59 ± 0.50	4.05 ± 0.68	3.84 ± 0.34
DISCONTINUOUS 2	500	4.13 ± 0.65	3.58 ± 0.49	3.95 ± 0.78	3.47 ± 0.53	3.44 ± 0.89	3.37 ± 0.60	3.90 ± 0.80	3.72 ± 0.44
FEW-SHOT	500	4.34 ± 0.53	3.86 ± 0.33	4.21 ± 0.65	3.79 ± 0.40	3.70 ± 0.77	3.72 ± 0.44	4.16 ± 0.59	3.93 ± 0.33
FEW-SHOT CREATIVELY	500	4.22 ± 0.61	4.08 ± 0.27	4.14 ± 0.64	4.05 ± 0.36	3.72 ± 0.71	4.01 ± 0.37	4.18 ± 0.56	4.08 ± 0.28

Table 19: The faithfulness and creativity scores for all strategies and all languages.

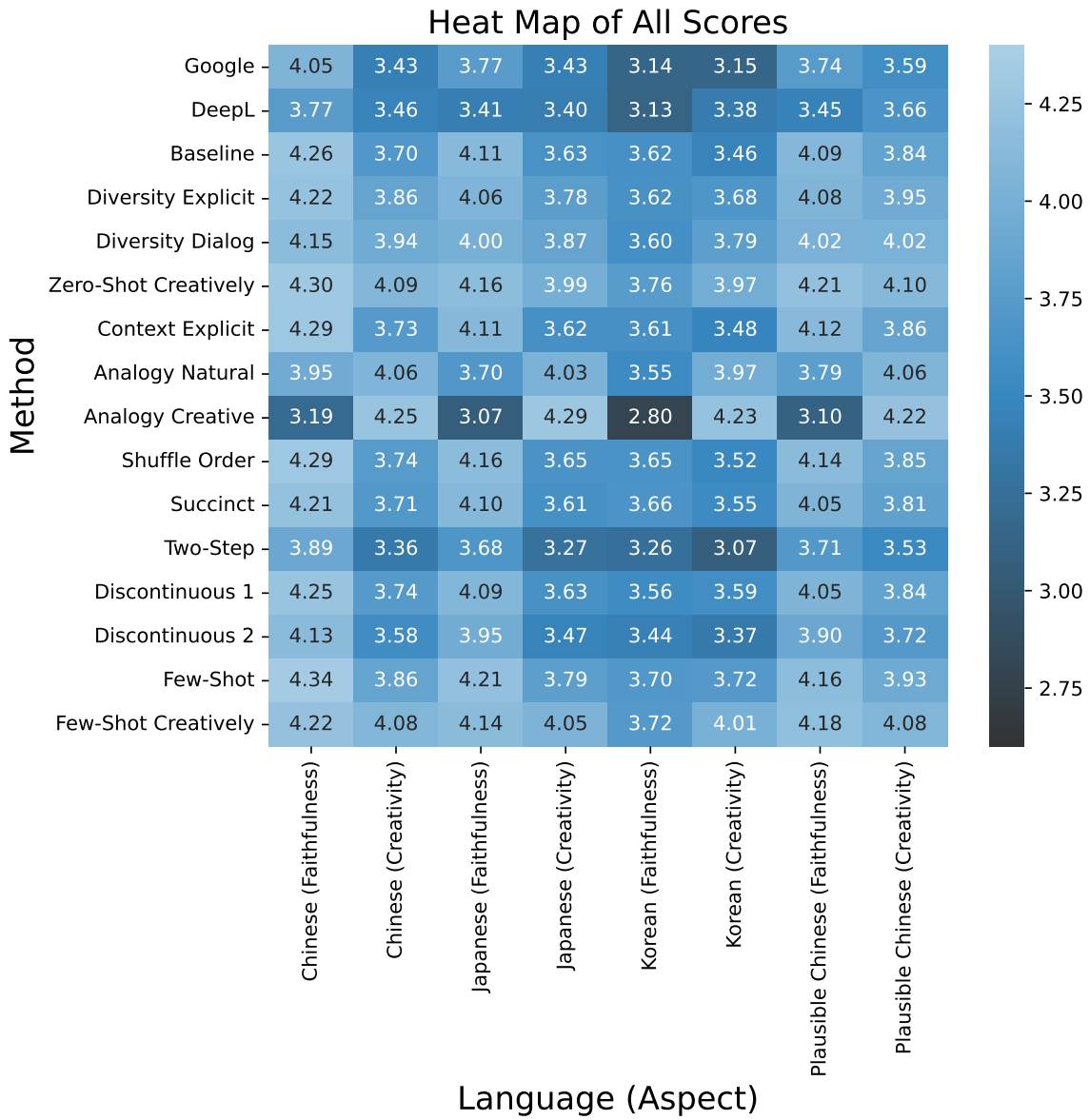

Figure 3: The heat map for all faithfulness and creativity scores.

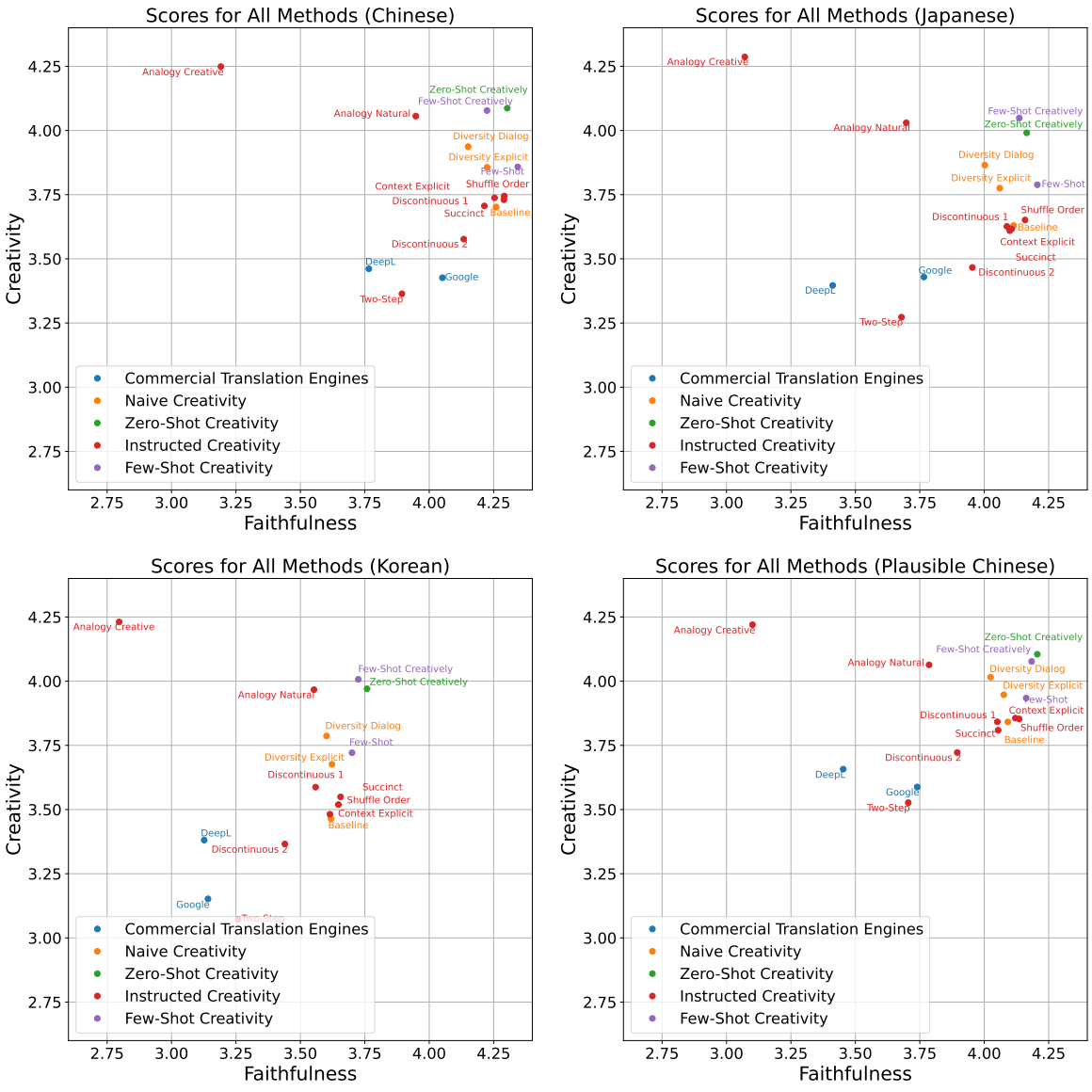

Figure 4: The scatter plots for all faithfulness and creativity scores.