Translation with context in LLMs: A study in Korean Idioms

Anonymous ACL submission

Abstract

Despite advancements in Large Language Models (LLMs), translation quality of generated outputs still remains inconsistent, particularly due to the misalignment in corresponding expressions across source and target languages. In this paper, we study the behavior of LLMs, focusing on the translational strategies of non-compositional expressions or idiomatic expressions. While LLMs are capable of translating non-compositional expressions as shown by the high average COMET score of 0.7969, a high inconsistent corresponding idiomatic translation accuracy across multiple context sentences for the same idiom indicate a lack of deeper understanding of the idiom and its surrounding context. Our results provide a starting point to understand how LLMs process and handle non-compositional expressions.

1 Introduction

As newer and more advanced large language models (LLMs) are developed, LLMs are becoming more proficient across multiple languages (Zhu et al., 2024; Xue et al., 2021; Brown et al., 2020), including low resource languages (Cahyawijaya et al., 2024). In addition, proprietary LLMs such as HyperCLOVAX have been observed to perform well on FLORES+ benchmark (Yoo et al., 2024). However, on the other hand, translation still remains a challenge, particularly due to the generation of less natural-sounding output in other applications such as literacy (Shafayat et al., 2024) and text-to-image (Saxon et al., 2024). This inconsistency in translation performance across various benchmarks thus motivates us to answer the following question: *Are LLMs able to understand what they are translating?*

As such, in this study, our objective is to understand the behavior of LLMs during the translation of non-compositional expressions, which are defined by Dankers et al. (2022) as expressions with definitions that cannot be derived from individual entities. In particular, we use Korean idioms as an approximate for non-compositional expressions. We first compiled a Korean-English idiomatic dataset (section 3.1) and prompted 3 Korean LLMs (section 3.2) to translate multiple *context sentences* or sentences that contain the use of idioms and studied the translational behavior through the quality of the translation output using two measurements: **overall translational accuracy** and **corresponding idiomatic translation accuracy** (section 3.3). Our results show that though LLMs generally perform well from the perspective of overall translation accuracy with a high COMET score of 0.7969 (section 4.1), there is still a lack of understanding towards the context surrounding the idiom as seen from the high inconsistency rate of 70% to 80% *corresponding idiomatic translation* (section 4.2). We believe that these results can be beneficial towards a wider NLP community, particularly on how LLMs handle hallucinations and unseen instances.

2 Related work

Translation in idioms Due to their non-compositional nature, idioms have been frequently studied in translation tasks (Dankers et al., 2022; Hwang and Hidey, 2019).[1] In particular, Dankers et al. (2022) observed that idioms are generally processed as compositional terms and thus tend to be directly translated. However, translating such terms still remain a challenge. Shafayat et al. (2024) noted several translation errors such as mismatches in honorifics and use of inappropriate phrases that do not sound natural in the context.

Contextual knowledge in LLMs Contrary to the expectation that understanding context is important in generating appropriate translation, context

[1] Though there are other studies in linguistics (Kim, 2018) which have postulated idioms being compositional phrases, we will assume that idioms are generally non-compositional in nature.

can hinder rather than enhance understanding of idioms in LLMs. Mi et al. (2024) observed that LLMs were not able to differentiate the type of context (e.g. figurative or literal) when the idiom was used. Similarly, LLMs that are not trained to be aware of idioms experienced a degradation in performance (Cheng and Bhat (2024) as cited in Mi et al. (2024)). However, these studies focused on how LLMs interpret idiomatic expressions rather than the quality of translated outputs across multiple contextual sentences.

Evaluating idiomatic processing in LLMs A variety of methods have been used in previous studies, which include attention heads (Dankers et al., 2022), similarity score (He et al., 2024) and a mixture of automated metrics and analysis of quality of translated output through lexical, honorifics and syntax (Shafayat et al., 2024). However, we consider these methods out of our scope as we are focusing on translational strategies across multiple context sentences translated by auto-regressive LLMs.

Memorization in idiomatic LLMs Idioms has also been widely used to demonstrate memorization in LLMs and are generally treated as "stored expressions" when processed by pre-trained language (Mi et al., 2024). In particular, Haviv et al. (2023); Li et al. (2024) defined memorization of idioms based on the ability of LLMs to correctly predict the final token However, due to the nature of Korean idioms in which the last token refers to the infinitive form (or 다), this makes it difficult to adapt the aforementioned tests into this study.

3 Methodology

3.1 Dataset

We first collated a list of Korean idioms with their corresponding definitions from two published books – New Style TOPIK Idiom (**source A**) [2] and TOPIK II Pass Recipe (**source B**),[3] which we have obtained permission from the publishers.[4] We extracted a total of 330 idioms, of which 123 and 187 idioms are from sources A and B respectively. Based on the definitions in the sources, we filtered for idioms with only 1 definition.

Next, we included context sentences to each idiom, which we define as sentences that combine idioms with various grammatical structures. We excluded sentences that contained modifications to the idioms such as addition of adverbs to avoid any potential loss in idiomatic meaning, in a manner similar to Mi et al. (2024). We sourced 5 sentences from source A for idioms present in Source A and 2-3 sentences from 4 electronic dictionaries for idioms present in source B. These 4 dictionaries[5] are: Naver dictionary, [6] Daum dictionary, [7] Urimalsaem (우리말샘) dictionary,[8],and Korean basic dictionary (한국어기초사전).[9]

After compilation of the context sentences, we conducted another round of check and replaced the sentences using source A and electronic dictionaries when required. This gives a total of **161 idioms and 651 context sentences**. A sample of the sentences can be seen in Figure 11. More details can also be found in Appendix A. Although the idioms and context sentences used in this study is of a small scale, we consider this to be acceptable for evaluation of the LLMs as we are neither fine-tuning nor training.

3.2 Models

We used `EXAONE-3.0-7.8B-Instruct` (Research et al., 2024), `Ko-Gemma-2-9B-IT`,[10] `Llama-VARCO-8B-Instruct`[11] in our study. Based on our understanding, these LLMs are not specifically trained to translate Korean idioms.

3.3 Experiment

We prompted the LLMs to translate the context sentences into English using greedy search. Since our study focus more on translational behavior in LLMs rather than reasoning, we chose a simple prompt technique over prompt engineering techniques such as Chain-of-Thought (Wei et al., 2023). More details can be found in appendix B.

3.4 Evaluation

We analyzed the generated translations through *overall translation accuracy* and *corresponding idiomatic translation*. We conducted two separate

[2]Original Korean title: New 스타일 TOPIK 관용표현

[3]Original Korean title: TOPIK II 합격 레시피

[4]TOPIK (Test of Proficiency in Korean) is a proficiency test in Korean language and targets Korean language learners.

[5]Note: Several dictionaries also cite several other dictionaries, including the 4 listed here.

[6]https://ko.dict.naver.com.

[7]https://dic.daum.net/index.do?dic=kor

[8]https://opendict.korean.go.kr/main

[9]https://krdict.korean.go.kr/

[10]huggingface.co/rtzr/ko-gemma-2-9b-it

[11]huggingface.co/NCSOFT/Llama-VARCO-8B-Instruct

analysis as there are no available metrics that evaluate both overall translation and idiomatic translation to our best understanding.

Overall translation accuracy We compared the LLM-generated outputs with source context sentences using a reference-free `wmt22-cometkiwi-da`,[12] which estimates the quality of the expression of interest from a scale of 0 to 1 (Rei et al., 2022). This provides an understanding of the overall translation quality, regardless of the accuracy of idiomatic translation.

Corresponding idiomatic translation We then define the corresponding idiomatic translation as the English translations of the Korean idioms in the context sentences. Inspired by Dankers et al. (2022); Mi et al. (2024), we annotated and categorized these translations into three strategies: **Direct translation** or word-for-word translation; **Indirect translation** or figurative translation, idiomatic translation, similar translations that captures the essence of the idiom; and **Mistranslation**. We then calculate the accuracy and precision using the formulas below.

$$accuracy(\%) = \frac{n_{indirect\ translation}}{n_{idiomatic\ translations}} * 100 \quad (1)$$

$$precision(\%) = \frac{n_{idioms\ indirect}}{n_{total\ idioms}} * 100 \quad (2)$$

where n refers to the count and $n_{idioms\ indirect}$ refers to the number of idioms with all translations marked as indirect translation. We use precision as an approximate of **model understandability** as we hypothesize that LLMs will be able to translate across all context sentences containing the idiom after acquiring relevant knowledge on the said idiom. Additional information is included in Appendix C.

4 Results

4.1 Overall translation accuracy

Table 1 shows the average COMET scores of all LLMs tested. All LLMs achieved a score above 0.7 with an average of 0.7969. These indicate a high overall accuracy in the generated translations. The score is further supported in Figure 1, where most LLM-generated translations were observed to be between the scores of 0.8-0.9. The result is in line with our expectation as the context in the context sentences can be generally translated through direct translation compared to idioms. As such, the tendency for LLMs to stay faithful to the source context sentence will be higher.

[12]/huggingface.co/Unbabel/wmt22-cometkiwi-da

LLM	Average COMET (4dp)
VARCO	0.7883
EXAONE	0.8035
Gemma	0.7990
Overall	0.7969

Table 1: Average COMET scores for all LLMs.

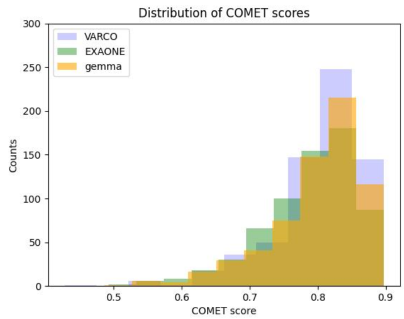

Figure 1: Distribution of COMET scores across all LLMs. Note that the yellow portion is due to the overlap between the different distributions.

4.2 Corresponding idiomatic translations

We present the distribution of corresponding idiomatic translations and understandability in Table 2. In general, the average idiomatic translation accuracy across all LLMs is observed to be 43%, which showed **that Though idiomatic translation accuracy is relatively lower than mistranslation due to the experimental design**

In summary, based on the results for corresponding idiomatic translation, we can infer that LLMs do have some form of "idiomatic understanding", but are less precise when translating across various context sentences.

5 Analysis and discussion

Based on our experimental findings and previous studies, we include additional elaboration in the following areas.

Why are LLMs generating inconsistent translations? We are inclined to think that these could be related to the difficulties LLMs experience in detecting the type of context used (Mi et al., 2024).

Our experimental results show a high inconsistency recorded. This could signify that a lack of awareness in context (i.e. direct or indirect), which

LLMs	idiomatic translations (accuracy)			idiomatic understandability (precision)		
	direct	figurative	mistranslation	consistency (f)	consistency (f+d)	inconsistency (1-(f+d))
VARCO	27.0	32.0	41.0	10.3	19.9	80.1
EXAONE	24.6	50.1	25.3	23.1	10.3	66.7
Gemma	24.9	38.4	36.7	10.9	9.6	79.5

Table 2: Breakdown of the *corresponding idiomatic translations* and rate of understandability. LLM with the highest percentage in each category are marked in blue and red.

LLM	EM	Some matches	Varied
VARCO	7.1	35.3	57.7
EXAONE	8.3	50.0	41.7
Gemma	7.1	41.7	51.3

Table 3: Percentage of exact matches (EM), some matches and varied translations across then 3 LLMs tested. Varied refers to instances where LLMs generated varied corresponding idiomatic translation, whereas some matches refers to cases where some corresponding idiomatic translations are identical but not all.

further impacted the type of translations generated. In addition, we theorize that instances with missing translations (i.e. marked as part of mistranslation) could also be related to the observation made by De Luca Fornaciari et al. (2024) where LLMs misidentify idioms, though further investigation is required, which is out of scope for this study.

Does acquiring idiomatic knowledge indicate any form of memorization? We examined the translated phrases and hypothesize that the tendency is higher when LLM generates identical corresponding idiomatic translation or *exact match* across all context sentences tested, regardless of the type of translational strategies annotated in section 3.3. [13] We present the results on Table 3. Based on the results, we observe that that the EM rate is generally low, indicating some form of memorization. Interestingly, upon further examination of the translations in Table 10 , we note that the LLMs tend to express corresponding idiomatic translations using phrases with similar meanings rather than using the exact words.

Generic behavior of LLMs in unseen cases For instances where LLMs are not aware of the translation, direct translation seems to be the default translation strategy, which is further supported by previous studies (Dankers et al., 2022). Though LLMs generate varied expression, we also note a higher tendency to generate hallucinated translations. We believe that this deviation can be explained by the lack of cultural knowledge and nuance, as supported by previous studies on Korean cultural benchmarks (Lee et al., 2024; Kim et al., 2024) and the inability to transfer such learning across multiple context sentences (Chhikara et al., 2025).

Impact to the wider NLP community We believe that our findings can help in understanding how LLMs approach unseen instances, particularly for non-compositional expressions. This is because LLMs may be required to understand from a semantic, cultural and bilingual level. Though prompting may enhance cultural alignment Pawar et al. (2024), further study is still required. Another application is on how LLMs interpret and handle hallucination as LLMs seem experience difficulties in differentiating between "acceptable" and "hallucinated" outputs. Ideally, LLMs should be able to rationalize their outputs, similar to the thought process (Cha, 1997) proposed. However, reasoning methods such as CoT was found to hinder performance in LLMs (Shafayat et al., 2024). This discrepancy highlights the need to further drive awareness in LLMs.

6 Conclusion

In this study, we aim to understand the behavior of LLMs during translation of non-compositional expressions. Using Koreans idioms as a proxy, we collated context sentences containing idioms and evaluated using overall translational accuracy and corresponding idiomatic translation accuracy. We observe that LLMs are generally able to translate context sentences containing idioms appropriately. However, LLMs seem to lack a deeper understanding across various context sentences, as seen from the inconsistent and hallucinated translations. We hope that the findings can be beneficial in instances where LLMs are exposed to unseen idioms or hallucination-related issues.

[13] We included translations with slight variations in the tenses, articles and pronouns.

Limitations

The study assumes pre-existing knowledge present in the fine-tuned and instruction-tuned LLMs. As such, we focus more on how LLMs can utilize their existing knowledge from CPT and fine-tuning to adapt into translation task and less on the specific type of dataset used. Though the mistranslation and inconsistent translation could be explained by the lack of exposure to specific knowledge or information during training, we do not consider this aspect and leave it for further studies to examine it.

In addition, we conducted our analysis through a combination of heuristic examination and automated metrics. Due to resource constraint, we did not conduct any large-scaled human annotation and focused on small-scaled annotation as a starting point for the study. As such, certain annotations and results can be subjective and may vary from person to person. We intend to automate and expand our annotation to include large-scaled human annotation as part of future work.

As this study is small-scaled in nature, we leave it to future work to expand into other similar bilingual LLMs, context sentences (e.g. both literal and figurative form, various formalities) and idioms for a better understanding of the overall behavior in LLMs. Scalability related experiments were not conducted as the focus was towards 8B-10B LLMs, though it would be interesting to see if LLMs can acquire and understand context better upon scaling.

Ethics Statement

There are no known ethical issues. However, the presence of slang, profanities or sexual/violence-related words in idioms and context sentences used during translation may indicate some form of ethical concern as the LLMs may have already been exposed to these words during training and fine-tuning.

Given the subjective nature of the heuristic approach in categorizing of the corresponding idiomatic translation, there might be some form of human bias present. In addition, in the course of annotation, we have sought guidance and translations from `GPT4o-mini-0718`, `ChatGPT` and `CLOVAX` (Yoo et al., 2024), on top with dictionaries, forums and commercial translators such as DeepL[14], which may further lead to more bias being present. As such, we advise that the results should be considered as an approximate before large-scale annotation is conducted.

[14] https://www.deepl.com/ko/translator

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Afina Putri, Emmanuel Dave, Jhonson Lee, Nuur Shadieq, Wawan Cenggoro, Salsabil Maulana Akbar, Muhammad Ihza Mahendra, Dea Annisayanti Putri, Bryan Wilie, Genta Indra Winata, Alham Fikri Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for indonesian languages.

Kil Cha. 1997. *Comparative Linguistic Studies of Korean and English Idioms*. Ph.D. thesis. Copyright - Copyright ProQuest Dissertations Publishing 1997; Last updated - 2023-06-21.

Kellen Cheng and Suma Bhat. 2024. No context needed: Contextual quandary in idiomatic reasoning with pre-trained language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4863–4880, Mexico City, Mexico. Association for Computational Linguistics.

Garima Chhikara, Abhishek Kumar, and Abhijnan Chakraborty. 2025. Through the prism of culture: Evaluating llms' understanding of indian subcultures and traditions.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2024. Investigating idiomaticity in word representations.

Alyssa Hwang and Christopher Hidey. 2019. Confirming the non-compositionality of idioms for sentiment analysis. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 125–129, Florence, Italy. Association for Computational Linguistics.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean.

Lan Kim. 2018. A note on aspectual properties of korean idioms. *Studies in Generative Grammar*, (3):433 – 447.

Jiyoung Lee, Minwoo Kim, Seungho Kim, Junghwan Kim, Seunghyun Won, Hwaran Lee, and Edward Choi. 2024. Kornat: Llm alignment benchmark for korean social values and common knowledge.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2024. Rolling the dice on idiomaticity: How llms fail to grasp context.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2024. Survey of cultural awareness in language models: Text and beyond.

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.

LG AI Research, :, Soyoung An, Kyunghoon Bae, Eunbi Choi, Stanley Jungkyu Choi, Yemuk Choi, Seokhee Hong, Yeonjung Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, Hyunjik Jo, Jiyeon Jung, Yountae Jung, Euisoon Kim, Hyosang Kim, Joonkee Kim, Seonghwan Kim, Soyeon Kim, Sunkyoung Kim, Yireun Kim, Youchul Kim, Edward Hwayoung Lee, Haeju Lee, Honglak Lee, Jinsik Lee, Kyungmin Lee, Moontae Lee, Seungjun Lee, Woohyung Lim, Sangha Park, Sooyoun Park, Yongmin Park, Boseong Seo, Sihoon Yang, Heuiyeen Yeen, Kyungjae Yoo, and Hyeongu Yun. 2024. Exaone 3.0 7.8b instruction tuned language model.

Michael Saxon, Yiran Luo, Sharon Levy, Chitta Baral, Yezhou Yang, and William Yang Wang. 2024. Lost in translation? translation errors and challenges for fair assessment of text-to-image models on multilingual concepts.

Sheikh Shafayat, Dongkeun Yoon, Woori Jang, Jiwoo Choi, Alice Oh, and Seohyon Jung. 2024. A 2-step framework for automated literary translation evaluation: Its promises and pitfalls.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, Donghyun Kwak, Hanock Kwak, Se Jung Kwon, Bado Lee, Dongsoo Lee, Gichang Lee, Jooho Lee, Baeseong Park, Seongjin Shin, Joonsang Yu, Seolki Baek, Sumin Byeon, Eungsup Cho, Dooseok Choe, Jeesung Han, Youngkyun Jin, Hyein Jun, Jaeseung Jung, Chanwoong Kim, Jinhong Kim, Jinuk Kim, Dokyeong Lee, Dongwook Park, Jeong Min Sohn, Sujung Han, Jiae Heo, Sungju Hong, Mina Jeon, Hyunhoon Jung, Jungeun Jung, Wangkyo Jung, Chungjoon Kim, Hyeri Kim, Jonghyun Kim, Min Young Kim, Soeun Lee, Joonhee Park, Jieun Shin, Sojin Yang, Jungsoon Yoon, Hwaran Lee, Sanghwan Bae, Jeehwan Cha, Karl Gylleus, Donghoon Ham, Mihak Hong, Youngki Hong, Yunki Hong, Dahyun Jang, Hyojun Jeon, Yujin Jeon, Yeji Jeong, Myunggeun Ji, Yeguk Jin, Chansong Jo, Shinyoung Joo, Seunghwan Jung, Adrian Jungmyung Kim, Byoung Hoon Kim, Hyomin Kim, Jungwhan Kim, Minkyoung Kim, Minseung Kim, Sungdong Kim, Yonghee Kim, Youngjun Kim, Youngkwan Kim, Donghyeon Ko, Dughyun Lee, Ha Young Lee, Jaehong Lee, Jieun Lee, Jonghyun Lee, Jongjin Lee, Min Young Lee, Yehbin Lee, Taehong Min, Yuri Min, Kiyoon Moon, Hyangnam Oh, Jaesun Park, Kyuyon Park, Younghun Park, Hanbae Seo, Seunghyun Seo, Mihyun Sim, Gyubin Son, Matt Yeo, Kyung Hoon Yeom, Wonjoon Yoo, Myungin You, Doheon Ahn, Homin Ahn, Joohee Ahn, Seongmin Ahn, Chanwoo An, Hyeryun An, Junho An, Sang-Min An, Boram Byun, Eunbin Byun, Jongho Cha, Minji Chang, Seunggyu Chang, Haesong Cho, Youngdo Cho, Dalnim Choi, Daseul Choi, Hyoseok Choi, Minseong Choi, Sangho Choi, Seongjae Choi, Wooyong Choi, Sewhan Chun, Dong Young Go, Chiheon Ham, Danbi Han, Jaemin Han, Moonyoung Hong, Sung Bum Hong, Dong-Hyun Hwang, Seongchan Hwang, Jinbae Im, Hyuk Jin Jang, Jaehyung Jang, Jaeni Jang, Sihyeon Jang, Sungwon Jang, Joonha Jeon, Daun Jeong, Joonhyun Jeong, Kyeongseok Jeong, Mini

Jeong, Sol Jin, Hanbyeol Jo, Hanju Jo, Minjung Jo, Chaeyoon Jung, Hyungsik Jung, Jaeuk Jung, Ju Hwan Jung, Kwangsun Jung, Seungjae Jung, Soonwon Ka, Donghan Kang, Soyoung Kang, Taeho Kil, Areum Kim, Beomyoung Kim, Byeongwook Kim, Daehee Kim, Dong-Gyun Kim, Donggook Kim, Donghyun Kim, Euna Kim, Eunchul Kim, Geewook Kim, Gyu Ri Kim, Hanbyul Kim, Heesu Kim, Isaac Kim, Jeonghoon Kim, Jihye Kim, Joonghoon Kim, Minjae Kim, Minsub Kim, Pil Hwan Kim, Sammy Kim, Seokhun Kim, Seonghyeon Kim, Soojin Kim, Soong Kim, Soyoon Kim, Sunyoung Kim, Taeho Kim, Wonho Kim, Yoonsik Kim, You Jin Kim, Yuri Kim, Beomseok Kwon, Ohsung Kwon, Yoo-Hwan Kwon, Anna Lee, Byungwook Lee, Changho Lee, Daun Lee, Dongjae Lee, Ha-Ram Lee, Hodong Lee, Hwiyeong Lee, Hyunmi Lee, Injae Lee, Jaeung Lee, Jeongsang Lee, Jisoo Lee, Jongsoo Lee, Joongjae Lee, Juhan Lee, Jung Hyun Lee, Junghoon Lee, Junwoo Lee, Se Yun Lee, Sujin Lee, Sungjae Lee, Sungwoo Lee, Wonjae Lee, Zoo Hyun Lee, Jong Kun Lim, Kun Lim, Taemin Lim, Nuri Na, Jeongyeon Nam, Kyeong-Min Nam, Yeonseog Noh, Biro Oh, Jung-Sik Oh, Solgil Oh, Yeontaek Oh, Boyoun Park, Cheonbok Park, Dongju Park, Hyeonjin Park, Hyun Tae Park, Hyunjung Park, Jihye Park, Jooseok Park, Junghwan Park, Jungsoo Park, Miru Park, Sang Hee Park, Seunghyun Park, Soyoung Park, Taerim Park, Wonkyeong Park, Hyunjoon Ryu, Jeonghun Ryu, Nahyeon Ryu, Soonshin Seo, Suk Min Seo, Yoonjeong Shim, Kyuyong Shin, Wonkwang Shin, Hyun Sim, Woongseob Sim, Hyejin Soh, Bokyong Son, Hyunjun Son, Seulah Son, Chi-Yun Song, Chiyoung Song, Ka Yeon Song, Minchul Song, Seungmin Song, Jisung Wang, Yonggoo Yeo, Myeong Yeon Yi, Moon Bin Yim, Taehwan Yoo, Youngjoon Yoo, Sungmin Yoon, Young Jin Yoon, Hangyeol Yu, Ui Seon Yu, Xingdong Zuo, Jeongin Bae, Joungeun Bae, Hyunsoo Cho, Seonghyun Cho, Yongjin Cho, Taekyoon Choi, Yera Choi, Jiwan Chung, Zhenghui Han, Byeongho Heo, Euisuk Hong, Taebaek Hwang, Seonyeol Im, Sumin Jegal, Sumin Jeon, Yelim Jeong, Yonghyun Jeong, Can Jiang, Juyong Jiang, Jiho Jin, Ara Jo, Younghyun Jo, Hoyoun Jung, Juyoung Jung, Seunghyeong Kang, Dae Hee Kim, Ginam Kim, Hangyeol Kim, Heeseung Kim, Hyojin Kim, Hyojun Kim, Hyun-Ah Kim, Jeehye Kim, Jin-Hwa Kim, Jiseon Kim, Jonghak Kim, Jung Yoon Kim, Rak Yeong Kim, Seongjin Kim, Seoyoon Kim, Sewon Kim, Sooyoung Kim, Sukyoung Kim, Taeyong Kim, Naeun Ko, Bonseung Koo, Heeyoung Kwak, Haena Kwon, Youngjin Kwon, Boram Lee, Bruce W. Lee, Dagyeong Lee, Erin Lee, Euijin Lee, Ha Gyeong Lee, Hyojin Lee, Hyunjeong Lee, Jeeyoon Lee, Jeonghyun Lee, Jongheok Lee, Joonhyung Lee, Junhyuk Lee, Mingu Lee, Nayeon Lee, Sangkyu Lee, Se Young Lee, Seulgi Lee, Seung Jin Lee, Suhyeon Lee, Yeonjae Lee, Yesol Lee, Youngbeom Lee, Yujin Lee, Shaodong Li, Tianyu Liu, Seong-Eun Moon, Taehong Moon, Max-Lasse Nihlenramstroem, Wonseok Oh, Yuri Oh, Hongbeen Park, Hyekyung Park, Jaeho Park, Nohil Park, Sangjin Park, Jiwon Ryu, Miru Ryu, Simo Ryu, Ahreum Seo, Hee Seo, Kangdeok Seo, Jamin Shin, Seungyoun Shin, Heetae Sin, Jiangping Wang, Lei Wang, Ning Xiang, Longxiang Xiao, Jing Xu, Seonyeong Yi, Haanju Yoo, Haneul Yoo, Hwanhee Yoo, Liang Yu, Youngjae Yu, Weijie Yuan, Bo Zeng, Qian Zhou, Kyunghyun Cho, Jung-Woo Ha, Joonsuk Park, Jihyun Hwang, Hyoung Jo Kwon, Soonyong Kwon, Jungyeon Lee, Seungho Lee, Seonghyeon Lim, Hyunkyung Noh, Seungho Choi, Sang-Woo Lee, Jung Hwa Lim, and Nako Sung. 2024. Hyperclova x technical report.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis.

Source	Idioms	Sentences	Sentence /idiom
A	98	485	5 (93); 4 (5)
B	58	166	2 (8); 3 (50)
Total	**156**	**651**	

Table 4: Number of idioms and context sentences in Sources A and B. Due to resource constraint, idioms with only 1 context sentence are excluded from this study. Final count of idioms and sentences are **bolded**.

A Idiomatic dataset collection

This section contains additional information on the selection of the context sentences.

To ensure consistency across all context sentences, we excluded sentences that are of conversational setting or contain idiomatic expressions with other grammatical structures (e.g. passive forms, addition of adverb or adjective).

We replace the sentences with alternative sentences from Source A (for idioms present in **source A**) or electronic dictionaries (for idioms present in **source B**). To ensure a consistency in the type of source used, for idioms in source A where only 4 context sentences were valid instead of 5, we kept the count of the sentence as it is.

Idioms with only 1 context sentence were also excluded from the study. This gives a final count of **156** idioms and **651** context sentences. A breakdown of the count is shown in Table 4. We also list some criteria in defining duplicates and unique idiom in Table 9.

We selected context sentences with varied context and position of idioms to the best of our ability, though we note of certain sentences with similar context or with idioms of similar positions. As we are interested in the consistency behavior of LLMs during translation, these instances will still be included in the analysis.

B Model

This section includes additional details on the settings used for implementation. We list the prompts used in Table 5.

LLM	Prompt
VARCO	"role": "system", "content": "You are a helpful assistant Varco. Respond accurately and diligently according to the user's instructions.", "role": "user", "content": Translate into English:context sentences
EXAONE	"role": "system", "content": "You are EXAONE model from LG AI Research, a helpful assistant.", "role": "user", "content": prompt prompt= f"""Translate into English:context sentence
GEMMA	"role": "user", "content": prompt prompt=영어로 번역해줘:context sentence

Table 5: Prompts used for the LLMs

We used a Korean prompt for GEMMA instead of English as we observe that that the generated translations are more desirable. For the implementation of the models, we used RTX A4000 GPUs.

C Categorization for corresponding idiomatic translation

The following section includes additional details on the categorization of corresponding idiomatic translation. We compared the corresponding idiomatic translations based on the LLM-generated outputs with the direct and figurative translations of the Korean idioms and categorized into the following translational strategies as seen in 3.3.

For instances that contains both direct and indirect translation, we categorized as direct translation. An example can be seen in Table 11. Additional examples for direct, indirect and mistranslations can also be seen in Tables 6-8. We also provide translation generated by `GPT-4o-mini-0718` for reference.

In order to minimize variation that might arise, we define similar meaning as words/phrases that are synonyms. For cases that are ambiguous or do not reflect fully the meaning of the idiom, we will consider under "mistranslation".

Direct translation		
Context sentence: 형이 언론사 채용 시험에 응시했다가 그만 미역국을 먹었다. **Idiom**: 미역국을 먹다 **Direct translation**: To eat seaweed soup; To eat miyeokguk[15] **Indirect translation**: To fail an examination		
LLM	**Translation**	**Categorization**
VARCO	He took the job exam at the newspaper, but gave up and ate miyeokguk instead.	direct
EXAONE	His older brother took the media company's hiring test but ended up eating seaweed soup.	direct
GEMMA	My older brother took a journalism job interview and then just gave up, like he ate a whole bowl of seaweed soup.	direct
GPT (reference)	My older brother took the hiring exam for a media company, but unfortunately, he failed.	indirect

Table 6: Examples of direct corresponding idiomatic translations across all 3 LLMs tested. All LLMs tested managed to provide an appropriate idiomatic translation.

Indirect translation		
Context sentence: 프리랜서로 일하기 시작한 뒤로 수입이 불안정해서 일거리가 없을 때는 입에 풀칠하기도 어렵다. **Idiom**: 입에 풀칠하다 **Direct translation**: put glue on one's mouth **Indirect translation**: To live in severe poverty. make ends meet; eke out a living		
LLM	**Translation**	**Categorization**
VARCO	It was hard for us to make ends meet with father's salary of eight dollars a month.	indirect
EXAONE	With the father's salary, it was difficult for the four family members to make ends meet.	indirect
GEMMA	His salary wasn't even enough to feed his family.	indirect
GPT (reference)	With my father's salary, it was difficult for our family of four to make ends meet.	indirect

Table 7: Examples of indirect corresponding idiomatic translations across all 3 LLMs tested. All LLMs tested managed to provide an appropriate idiomatic translation.

Mistranslations		
Context sentence: 부모의 행동이 변덕이 죽 끓듯 하면 아이들은 항상 눈치를 보게 된다. **Idiom**: 변덕이 죽 끓듯 하다 **Direct translation**: Fickle like a boiling porridge **Indirect translation**: For words, behavior or feelings to be fickle and frequently change.		
LLM	**Translation**	**Categorization**
VARCO	When parents' behavior changes erratically like a boiling pot, children always have to watch their step.	mistranslation
EXAONE	If parents' behavior is as unpredictable as boiling water, children will always be on guard.	mistranslation
GEMMA	When parents' behavior is as unpredictable as boiling water, children are always on edge, trying to read the situation.	mistranslation
GPT (reference)	When parents' behavior is as unpredictable as boiling water, children are always on edge.	direct

Table 8: Example of mistranslated corresponding idiomatic translations across all 3 LLMs tested. All LLMs, including GPT tested translated "porridge" as "water".

Excluded from dataset			
Description	**Examples**	**Definitions**	**Translations**
Synonyms	1. 가닥을 잡다 2. 실마리를 찾다/잡다	문제 해결 방법을 찾아내다.	Find a solution to the problem.
Polysemy	고개를 숙이다	1. 잘못이나 패배를 인정하다. 2. 부끄럽다.	1.Admit fault or defeat 2. Ashamed/embarrased.
Alternative idioms	1. 긴 말이 필요 없다 2. 긴말할 것 없다	길게 여러 말로 설명할 필요가 없다.	No need for long, wordy explanations.
Passive forms	1. 속을 태우다 2. 속이 타다	1. 너무 걱정이 되다. 2. 너무 걱정을 하다.	1. (being) so worried 2. Worrying too much.

Table 9: List of examples excluded from the collection and processing of the idiomatic dataset. Note: Examples are non-exhaustive. In cases where other external sources (e.g. dictionaries) show differences in definition or alternative expressions, we will use Sources A and B as the base. Translations of the meanings were obtained from DeepL().

LLM	Examples of mistranslation
VARCO	the fire burns on my toes fire broke out on my ankles
GEMMA	a burn on his foot like having a fire under their feet working until my feet burn bolt of lightning struck my foot

Table 10: An example of varied corresponding idiomatic translation generated by VARCO and GEMMA. The idiom of interest is 발등에 불이 떨어지다, which is literally translated to "a fire drops on top of one's foot".

Context sentence	Translation (VARCO)	Reference translation (GPT-4o-mini-0718)	Result
어릴 때는 하루가 멀다 하게 동생들에게 골탕을 먹었다.	When I was young, I used to give my siblings a hard time, making them feel like they were in for a long day.	When I was young I often played tricks on my siblings.	indirect
경쟁사의 공격적인 마케팅에 우리 회사는 골탕을 먹고 말았다.	The company has been taken aback by the aggressive marketing of our competitor.	The aggressive marketing of our competitor has caused our company a lot of trouble.	mistranslation
사람을 너무 잘 믿던 최 대리는 한번 크게 골탕을 먹고 정신을 차렸다.	Having trusted people too much, Choi Daeryeong was thoroughly humiliated and woke up to reality.	Choi who used to trust people too much got badly hurt once and came to his senses.	indirect
우리는 상대 팀의 예상치 못한 공격에 골탕을 먹고 경기에서 크게 패했다.	We were caught off guard by the opposing team's unexpected attack and suffered a crushing defeat in the match.	We suffered greatly in the match due to the unexpected attack from the opposing team.	indirect
마당에 자주 놀러오던 고양이가 우리 집 개한테 물려 잔뜩 골탕을 먹고 달아나 버렸다.	A cat that often played in the yard ate a large amount of dog food from our dog and ran away.	The cat that often came to play in the yard got bitten by our dog and ended up getting a good scare before running away.	mistranslation

Table 11: An example of a set of translation of context sentences with varying results. The idiom of interest is 골탕을 먹다, which is figuratively translated to "cause serious trouble or damage to someone". Words marked in Red refers to *corresponding idiomatic translations* with mismatches. In other words, there is a mistranslation.