Search if you don't know! Knowledge-Augmented Korean Grammatical Error Correction with Large Language Models

Seonmin Koo[1*], **Jinsung Kim**[1*], **Chanjun Park**[2†], **Heuiseok Lim**[1†]
[1]Department of Computer science and Engineering, Korea University
[2]Upstage AI
{fhdahd,jin62304,limhseok}@korea.ac.kr
chanjun.park@upstage.ai

Abstract

Grammatical error correction (GEC) system is a practical task used in the real world, showing high achievements alongside the development of large language models (LLMs). However, these achievements have been primarily obtained in English, and there is a relative lack of performance for non-English data, such as Korean. We hypothesize that this insufficiency occurs because relying solely on the parametric knowledge of LLMs makes it difficult to thoroughly understand the given context in the Korean GEC. Therefore, we propose a Knowledge-Augmented GEC (KAGEC) framework that incorporates evidential information from external sources into the prompt for the GEC task. KAGEC first extracts salient phrases from the given source and retrieves non-parametric knowledge based on these phrases, aiming to enhance the context-aware generation capabilities of LLMs. Furthermore, we conduct validations for fine-grained error types to identify those requiring a retrieval-augmented manner when LLMs perform Korean GEC. According to experimental results, most LLMs, including ChatGPT, demonstrate significant performance improvements when applying KAGEC.

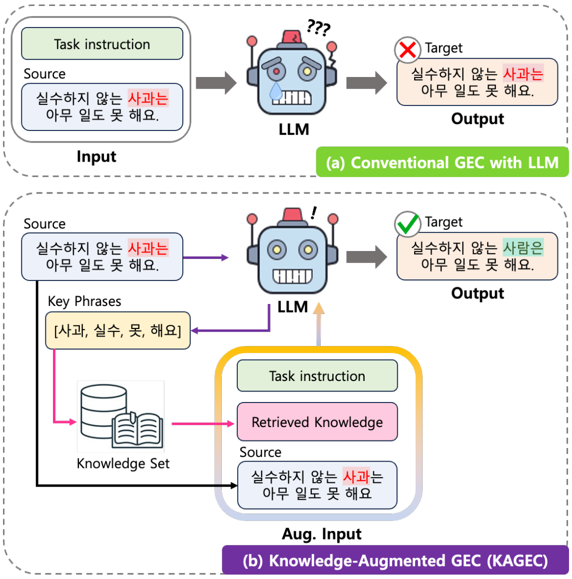

Figure 1: Overview of the proposed knowledge-augmented grammatical error correction (KAGEC) framework. The (a) part at the top shows the existing approach to perform the GEC task using LLMs and prompting. The (b) part at the bottom shows KAGEC's multi-step correction process. The purple arrows represent the first step, salient phrase extraction, and the pink arrows indicate the second step, knowledge retrieval. Aug. Input indicates the input prompt augmented with retrieved external information.

1 Introduction

Grammatical error correction (GEC) task aims to detect and correct textual errors in a given source sentence. In the real world, GEC is a practical and essential task used in applications such as writing assistance and language teaching (Rothe et al., 2021a; Bryant et al., 2023; Zhang et al., 2023). The development of large language models (LLMs) has led to the integration of various tasks with LLMs and has shown remarkable performance (Wu et al., 2023; Fang et al., 2023; Maeng et al., 2023). The recent GEC studies have also achieved decent outcomes by focusing on leveraging LLMs' inherent knowledge through prompting methods (Loem et al., 2023; Bryant et al., 2023).

However, these achievements have been mainly in English, which is heavily used in LLMs' training and have not shown the same level of superiority for non-English data including as Korean, compared to pre-trained language models (PLMs) (Kwon et al., 2023; Maeng et al., 2023). Figure 1 (a) shows the actual generation results of the conventional GEC approach with LLM [1]. The source sentence "실수하지 않는 사과는 아무 일도 못 해요. (An apple

[*] Equally contributed.
[†] Corresponding author.

[1]This is based on the actual results of ChatGPT (GPT-3.5).

Findings of the Association for Computational Linguistics: EACL 2024, pages 96–125
November 12-16, 2024 ©2024 Association for Computational Linguistics

that doesn't make mistakes can't do anything.)" exhibits an error where the subject is described as performing an action it practically cannot do. Despite the mistake of 'people' being replaced by 'apple' in the source sentence, the LLM fails to detect this error and outputs it as is in the target sentence, leading to an under-correction problem.

We hypothesize that *relying solely on the parametric knowledge of LLMs makes it challenging to understand the given context in the Korean GEC task thoroughly*. We find the clue to mitigate this issue by injecting linguistic information from external sources to enhance understanding of the source. Therefore, we propose a Knowledge-Augmented GEC (KAGEC) that refines information helpful for performing the GEC task from external sources and provides it in the prompt. We consider two aspects to provide relevant external knowledge for a given input: i) salient information (Su et al., 2022) and ii) relevant knowledge (Lee et al., 2019). By identifying key keywords in the input and retrieving information on these from external sources to provide alongside LLMs, we can enhance LLMs performance.

Figure 1 (b) represents the correction process of KAGEC. It involves a process of extracting salient key phrases from the source sentence, through which the errored word '사과는 (apple)' is identified. Afterward, Knowledge corresponding to the extracted keywords is retrieved, and an augmented input is built by including them. Through this method, LLMs enhance their linguistic understanding of the source sentence, recognizing the incorrect substitution of 'people' with 'apple' and precisely correcting it to "사람은 아무 일도 못 해요. (People who don't make mistakes can't do anything.)." This indicates that KAGEC effectively supports LLMs' inherent capabilities by augmenting retrieval from non-parametric knowledge.

Furthermore, we conduct a detailed validation for diverse error types in Korean GEC with LLMs to identify those specifically requiring knowledge augmentation. We provide additional experiments and extensive analysis, including investigation of qualitative results related to informativeness.

Our contributions are threefold: i) we categorize and validate fine-grained error types for the Korean GEC task, ii) we propose a Knowledge-Augmented GEC (KAGEC) framework that enhances the understanding of context (source) by furnishing non-parametric information as knowledge, and iii) we demonstrate the effectiveness of KAGEC through extensive additional experiments, including in-depth analysis and diversification of external knowledge types.

2 Related Works

2.1 Knowledge-grounded Prompting Studies

Recently, the effectiveness of various prompting methods such as zero-shot chain-of-thought (CoT) (Wei et al., 2022; Kojima et al., 2022) and task decomposition (Zhou et al., 2022; Wang et al., 2023; Zhang and Gao, 2023) has been demonstrated as strategies to actively elicit knowledge from LLMs. These methods have primarily focused on actively and directly drawing out the parametric knowledge of LLMs.

Conversely, there has been an emergence of research incorporating non-parametric knowledge to target knowledge-based generation in performing knowledge-intensive tasks (Guu et al., 2020; Karpukhin et al., 2020; Lee et al., 2019). This is achieved through the introduction and provision of external knowledge as evidence. The Retrieval-Augmented Generation (RAG) system is a representative non-parametric approach (Lewis et al., 2020; Mao et al., 2020; Shi et al., 2023). Studies integrating external and LLMs' inherent knowledge, particularly in knowledge-intensive tasks like open-ended question answering, have been pursued through search methods (Khalifa et al., 2023; Jiang et al., 2023; Ram et al., 2023).

However, there is a scarcity of research on enhancing GEC systems with knowledge from external sources. Therefore, we are the first to apply the non-parametric approach, which has shown exceptional achievements in knowledge-intensive tasks, to the Korean GEC task.

2.2 Grammatical Error Correction Studies

Many existing GEC studies have addressed the task by leveraging approaches from machine translation, considering the task as translating error sentences into correct sentences (Kaneko et al., 2022; Gan et al., 2021; Cao et al., 2021). These studies consider the GEC task as a process of translating error sentences into correct sentences. For instance, many methods exist that perform corrections in a black-box manner, utilizing sequence-to-sequence models without detailed differentiation of error types (Awasthi et al., 2019; Rothe et al., 2021b; Tarnavskyi et al., 2022).

Type	# Shots									
	0		1		4		8		16	
	BLEU	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**
Group 1: Types of sufficient performance										
Spacing	47.59	47.62	60.17	58.31	76.97	74.37	79.40	76.99	79.68	77.36
Punctuation	47.56	47.55	62.50	61.17	79.64	77.72	84.85	83.17	**88.67**	**87.14**
Addition	47.77	46.90	60.23	57.85	71.01	68.10	76.91	73.88	76.50	73.24
Rotation replace	51.83	49.48	59.10	56.37	67.49	64.25	74.46	71.80	76.04	72.31
Separation	**64.90**	**66.27**	**71.49**	**70.85**	**86.41**	**85.49**	**87.06**	**86.54**	85.39	85.24
Foreign and conversion	50.50	47.29	59.28	55.75	73.02	69.12	78.50	74.65	78.38	74.35
Consonant-vowel conversion	53.01	51.83	60.01	57.50	72.00	69.77	72.14	70.09	77.87	74.91
G2P	45.37	43.78	59.29	57.19	70.56	68.55	77.74	74.62	79.39	76.16
Postposition	53.93	50.19	62.74	59.25	68.45	65.45	78.32	75.73	77.11	73.97
Suffix	49.07	48.99	58.67	57.92	74.03	74.15	75.57	75.64	75.61	76.57
Group 2: Types of performance increase with setting adjustment										
Numerical	41.72	37.72	51.89	46.66	**63.97**	**58.33**	**68.53**	62.80	**72.53**	66.69
Remove	**42.37**	**41.27**	**53.75**	**50.90**	60.02	57.54	66.86	**63.85**	69.63	**67.03**
Tense	37.60	31.56	50.08	43.49	62.75	53.46	65.32	56.12	67.15	58.92
Neologism	36.62	37.00	51.18	48.78	59.57	57.24	65.27	62.13	69.02	66.46
Group 3: Types requiring new prompting										
Element	47.57	47.05	45.15	43.96	55.19	54.25	59.61	58.81	62.68	62.19
Auxiliary predicate	**51.44**	**51.80**	**56.74**	**57.78**	**64.63**	**64.32**	**72.43**	**74.76**	**69.64**	**70.49**
Behavioral	43.66	38.95	43.58	36.83	52.79	47.71	57.35	51.46	59.80	56.01
Avg.	46.58	45.34	55.45	53.19	67.51	65.31	72.02	69.79	73.63	71.52

Table 1: ChatGPT-3.5 results for different error types in the GEC task. Avg. denotes the average performance of all error types. G2P refers grapheme-to-phoneme. we set the temperature as 0.2. Table regarding temperature adjustments can be found in Appendix B.

With the advent of LLMs, research utilizing prompting methods to leverage parametric knowledge has shown remarkable performance in GEC (Wu et al., 2023; Fang et al., 2023; Loem et al., 2023). However, such achievements are still insufficient in non-English contexts (Kwon et al., 2023; Maeng et al., 2023), and the utilization of non-parametric knowledge remains limited. Therefore, we enhance the context-aware capabilities of LLMs through the augmentation of non-parametric knowledge.

3 Preliminary: Validation of Error Types

In this section, we assess the GEC capabilities of each model across different error types and categorize these types based on specific criteria. To conduct this validation, we employ ChatGPT, known for its exceptional performance across a range of downstream tasks.

3.1 Error Types for Korean GEC

Previous studies on LLMs for GEC often overlook the variety of error types (Wu et al., 2023; Fang et al., 2023; Maeng et al., 2023). However, because different error types have their own characteristics, achieving high performance in specific types does not guarantee improvements across all types of errors. Therefore, it is necessary to investigate the types that require improvement in the GEC performance of LLMs (Koo et al., 2023; Zhao et al., 2023). To this end, we utilize the error types established in a prior study.

The K-NCT dataset, introduced by Koo et al. (2022), is a Korean gold-standard test dataset for the GEC task. Detailed descriptions of each error type are provided in Appendix A. This dataset first proposed fine-grained error types for Korean GEC with considerations for balance, diversity, and factuality, making it an appropriate choice for our validation dataset.

Our study encompasses error types that are identifiable within the context of a single sentence or that have sufficient examples for conducting few-shot experiments. Consequently, error types that are not discernible within a single sentence or have a scarcity of examples in the raw data are excluded from our work scope.

3.2 Error Type Clustering

We first define criteria for categorizing 17 error types and validate which individual error types can be included in each cluster. We group the error types into three categories as follows: 1) Types for which LLMs already exhibit sufficient performance (i.e., higher outcomes than average), 2) Types where adjustments in settings (e.g., the number of shots or temperature) are effective, and 3) Types for which neither of the previous two strategies is effective (requiring new prompting approach). The ineffectiveness of providing various exemplars or adjusting the hyperparameters indicates that there are limits to solving the GEC problem in these error types with the parametric knowledge of LLMs.

3.3 Validation Results by Error Type

Table 1 presents the GEC performance of ChatGPT across various segmented error types, including zero-shot and one-shot results, along with few-shot (4, 8, and 16) performances. Initially, error types such as spacing, punctuation, addition, rotation, replacement, separation, foreign and conversion, consonant-vowel conversion, grapheme-to-phoneme (G2P), postposition, and suffix fall under the category of types with sufficient performance (Group 1). These types demonstrate performance above the average in both zero-shot and 16-shot scenarios, indicating that LLMs possess adequate understanding and capacity related to these error types.

Error types, including numerical, removal, tense, and neologism, are identified as the group where adjustment of settings proves effective. (Group 2) The performance significantly improves with an increase in the number of provided examples. Notably, the numerical type shows a BLEU score improvement of 30.81 from zero-shot to 16-shots, suggesting these types can be adequately addressed by providing carefully crafted examples.

Lastly, element, auxiliary predicate, and behavioral types are classified as requiring new prompts. These types do not sufficiently improve with the LLM's parametric knowledge alone and necessitate support through external knowledge. This demonstrates that providing examples in a few-shot setting or enhancing diversity by adjusting the generation temperature is not universally effective across all error types. Consequently, we focus on improving the three identified types (element, auxiliary predicate, behavioral) that are categorized as requiring new prompts for improvement.

4 Knowledge-Augmented Grammatical Error Correction (KAGEC)

In this section, we propose an approach that enhances the effectiveness of the GEC task through a knowledge-augmented generation method (KAGEC) by adopting the retrieval of non-parametric knowledge.

In many knowledge-intensive tasks, such as open-ended question answering, building a valid pool of non-parametric knowledge and effective retrieval plays a crucial role (Ram et al., 2023; Shi et al., 2023). Due to the challenges of capturing relevance in a zero-shot setting, recent studies have demonstrated the efficiency and effectiveness of involving LLM in the search process for relevance modeling (Shao et al., 2023; Jiang et al., 2023).

Inspired by this, we first construct a set of knowledge, K, that can finally enhance the generation ability of LLMs. From this constructed knowledge set, we retrieve a subset of knowledge, K^s, that is relevant to the given source sentence, s, and reconstruct the input with the retrieved knowledge. Afterward, the input enriched with knowledge, s', supports the generative capabilities of LLMs, resulting in the generation of the target sentence, t.

4.1 Knowledge Retrieval Set Construction

This work defines textual descriptions of dictionary definitions for terminology registered in the Korean dictionary as the necessary knowledge for performing the GEC task. By injecting linguistic information about salient keywords in the context, we enable LLMs to serve context-aware GEC (Radford et al., 2019; Eisenschlos et al., 2023).

To construct a knowledge pool for retrieval, we scrap definitions for each terminology from the National Institute of Korean Language's Standard Korean Dictionary [2]. The National Institute of Korean Language is an authoritative institution that establishes norms for Korean linguistics [3]. Afterward, by refining these into a dictionary structure of term-definition pairs, terminology registered by the National Institute of Korean Language becomes the key k, and its corresponding definition description becomes the value v. Thus, the non-parametric knowledge set K is structured as $\{k_1 : v_1, k_2 : v_2, ..., k_m : v_m\}$.

[2] https://opendict.korean.go.kr/main
[3] https://www.korean.go.kr/

Models		Element		Auxiliary predicate		Behavioral	
		BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
ChatGPT	Baseline	47.57	47.05	51.44	51.80	43.66	38.95
	KAGEC (Ours)	**58.11**	**57.96**	**65.31**	**63.99**	**54.13**	**45.83**
	CoT	35.89	34.37	48.70	47.09	36.05	31.38
	Decomposed	43.51	41.05	54.58	54.76	49.92	43.91
Gemini	Baseline	56.22	55.24	**58.12**	58.41	50.42	41.83
	KAGEC (Ours)	**62.89**	**62.94**	57.97	**60.06**	**58.72**	**53.11**
	CoT	48.18	47.75	49.65	50.00	41.62	37.50
	Decomposed	50.05	50.99	56.24	58.07	51.29	46.90
Llama2-13B	Baseline	45.88	45.80	**52.13**	**51.80**	38.72	33.35
	KAGEC (Ours)	**52.11**	**49.37**	30.91	32.71	**38.85**	**33.60**
	CoT	40.43	40.45	41.82	44.20	30.53	27.33
	Decomposed	34.07	32.21	31.50	31.48	25.80	20.37
Llama2-70B	Baseline	42.47	41.75	52.17	53.22	52.78	39.90
	KAGEC (Ours)	**53.02**	**49.84**	**64.54**	**67.81**	**54.67**	**44.17**
	CoT	24.78	23.15	31.53	36.10	30.37	25.83
	Decomposed	39.39	36.03	38.57	38.03	38.87	29.67

Table 2: Main results of KAGEC and other reasoning-enhanced prompting methodologies in LLMs. CoT stands for zero-shot chain-of-thought, and Decomposed stands for task decomposition prompting method.

4.2 Input Reconstruction and Generation

Input Reconstruction. To construct the input s', which includes external knowledge for generating the final answer (i.e., the target sentence t), the given source sentence s is restructured in multiple steps.

Initially, the given source sentence s is used as a query to prompt the LLM to extract a set of top-n key phrases, P^s. Afterward, for each keyword $p_i \in P^s$ extracted, corresponding definitions are searched within the constructed knowledge set K. The search process $Ret(\cdot)$ can be formalized as follows:

$$\underset{p_i \in P^s}{Ret(p_i)} = \begin{cases} v_j, & \text{if } p_i == k_j \ (k_j \in K) \\ None, & \text{otherwise} \end{cases} \quad (1)$$

If k_i has homonyms, all retrieved descriptions are used as knowledge.

Through this search process, the knowledge subset K^s corresponding to the source sentence s is constructed, for example, as $\{v_{p_1}, ..., v_{p_i}, ..., v_{p_{|P^s|}}\}$. By concatenating K^s with the original given input s, the knowledge-augmented input s' is built.

Answer Generation. Finally, the input s', constructed with the addition of retrieved external knowledge, is fed into the LLM to generate the final corrected target sentence t.

5 Experiments

Detailed setups are described in Appendix C.

5.1 Prompting Methods

To evaluate the effectiveness of the method proposed in this study, KAGEC, we compare it against three other prompting methods: baseline (Wu et al., 2023), zero-shot chain-of-thought (CoT)(Kojima et al., 2022), and task decomposition(Zhou et al., 2022)[4].

The baseline prompt refers to the addition of responses in Korean to the existing English GEC prompt (Wu et al., 2023). Zero-shot CoT prompting introduces reasoning paths constructed with a trigger sentence such as "Let's think step-by-step," combined with the original task instructions and input. This method aims to draw out intrinsic knowledge to serve as evidence. Task decomposition prompting, inspired by the improved reasoning capabilities observed when LLMs dissect complex tasks or problems into smaller sub-claims, employs a multi-step approach. It involves task instructions that encourage the LLM to break down the GEC problem into more manageable segments for resolution.

5.2 Results

Table 2 shows the experimental results for each method across three error types. Notably, the KAGEC method uniformly demonstrates significant performance enhancements across all models for 'Element' and 'Behavioral' error types.

Specifically, the ChatGPT model shows the most

[4]For detailed examples of prompt configurations, please refer to Appendix D

Models		Element		Auxiliary predicate		Behavioral	
		BLEU	GLEU	BLEU	GLEU	BLEU	GLEU
ChatGPT	Baseline	47.57	47.05	51.44	51.80	43.66	38.95
	KAGEC (Ours)	**58.11**	**57.96**	**65.31**	**63.99**	**54.13**	**45.83**
	Random Knowledge	49.67	49.13	42.18	43.37	52.82	45.02
	Error Type Desc.	49.45	45.58	61.28	58.83	49.04	41.53
Gemini	Baseline	56.22	55.24	**58.12**	58.41	50.42	41.83
	KAGEC (Ours)	**62.89**	**62.94**	57.97	**60.06**	**58.72**	**53.11**
	Random Knowledge	56.50	56.08	42.33	42.81	56.68	49.53
	Error Type Desc.	33.36	32.61	33.24	34.52	41.29	33.74
Llama2-13B	Baseline	45.88	45.80	**52.13**	**51.80**	38.72	33.35
	KAGEC (Ours)	**52.11**	**49.37**	30.91	32.71	**38.85**	**33.60**
	Random Knowledge	9.79	10.74	20.77	22.87	12.96	11.19
	Error Type Desc.	16.77	16.77	32.56	33.59	20.67	16.35
Llama2-70B	Baseline	42.47	41.75	52.17	53.22	52.78	39.90
	KAGEC (Ours)	**53.02**	**49.84**	**64.54**	**67.81**	**54.67**	**44.17**
	Random Knowledge	5.36	7.62	2.70	3.58	3.69	3.89
	Error Type Desc.	11.61	10.92	2.99	4.45	17.27	16.05

Table 3: GEC performance results of LLMs according to the type of knowledge provided as evidence. Random Knowledge is when dictionary definition information of arbitrary terminology is provided. Additionally, Error Type Desc. is a method of providing descriptions for all fine-grained error types.

considerable improvement when KAGEC is applied for the 'Element' type, achieving an uplift of 10.54 %p in BLEU and 10.91 %p in GLEU compared to the baseline approach. Except for the Llama2-13B model, the KAGEC method also exhibits the most robust performance for the Auxiliary predicate type. The performance decline of KAGEC in addressing 'Auxiliary predicate' errors in the Llama2-13B model is detailed in section 6.3.

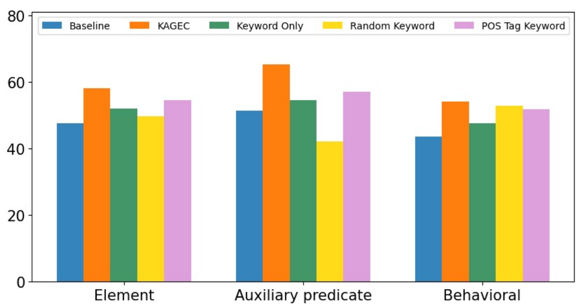

Figure 2: Results for additional knowledge provision methods in ChatGPT, i.e., 'Keyword Only', 'POS Tag Keyword.'

For reasoning-enhanced prompting methods, namely zero-shot CoT and task decomposition, there is a trend of performance reduction or inconsistent gains relative to the baseline prompting. CoT, in particular, tends to underperform across all error types and models, with a notable average performance decrease of approximately 20%p observed in the Llama2-70B model. Moreover, while the decomposed method indicates a minor score increase for the ChatGPT and Gemini models in Auxiliary predicate and 'Behavioral' types, it exhibits inconsistent outcomes for the Llama2 models, resulting in decreased scores.

These results can be interpreted as demonstrating the effectiveness of the KAGEC method in enhancing the GEC task performance through the incorporation of external knowledge into the prompting approach. To gain a deeper understanding of the outcomes between the KAGEC method and other reasoning-enhanced methods, a qualitative analysis comparing the results is conducted in section 6.2.

6 Analysis

6.1 Effectiveness of Retrieved Knowledge

We conduct additional experiments to ascertain whether the provision of deliberately retrieved knowledge actually guarantees effectiveness in the GEC task. KAGEC enables LLMs to extract key phrases from a source sentence and retrieve corresponding non-parametric knowledge. Table 3 demonstrates a performance comparison with two other cases of providing knowledge, aiming to prove the informativeness of knowledge integrated through KAGEC as evidence.

'Random Knowledge' refers to replacing the deliberately extracted knowledge based on key phrases in the KAGEC method with random knowledge (i.e., dictionary definitions of arbitrary terms). 'Error Type Desc.' involves providing descriptions for each detailed error type along with the input.

Source Sentence	
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 싶어. (I want to stay cool this summer, so I want to new clothes.)	
Target Sentence	
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 사고 싶어. (I want to stay cool this summer, so I want to buy new clothes.	
(a) Predicted Results According to Prompting Method	
Baseline	~~이번 여름에는 시원하게 보내고 싶어서~~ 옷을 새로 사고 싶어요. (~~I want to stay cool this summer, so~~ I would like to buy new clothes.)
CoT	이번 여름에는 시원하게 보내고 싶어서 옷을 새로 살고 싶어요. - "옷을 새로 싶어" → "옷을 새로 살고 싶어요" : "싶어" 는 동사 "새로 살다"와 함께 사용되어야 합니다. 왜 옷을 새로 살고 싶을까요? - 시원한 옷을 입어서 더 쾌적하게 여름을 보내고 싶어서예요. (I want to stay cool this summer, so I want to live anew clothes. - "want new clothes" → "want to live anew clothes.": "want" should be used in conjunction with the verb "to live anew". Why want to live anew clothes?: It is because I want to spend the summer more comfortably by wearing cool clothes.)
Decomposed	– *Step 1* 새로 싶어 → 새로 샀어 (want to newly→ bought new) – *Step 2* 이번 여름에는 시원하게 보내고 싶어서 옷을 새로 샀어. (I want to stay cool this summer, so I bought new clothes.
KAGEC	– *Top-n*: [‘이번’, ‘여름’, ‘시원하게’, ‘보내다’, ‘옷’] ([‘this’, ‘summer’, ‘cool’, ‘stay’, ‘clothes.’]) {{retrieved definitions}} 이번 여름에는 시원하게 보내고 싶어서 옷을 새로 사고 싶어. (I want to stay cool this summer, so I want to buy new clothes.)
(b) Predicted Results by External Information Types	
Random	– *Top-n*: [‘원함수’, ‘바래움’, ‘복화합물’, ‘베어링’, ‘명감’] ([‘this’, ‘summer’, ‘cool’, ‘stay’, ‘clothes.’]) {{random definitions}} 이번 여름에는 시원하게 보내고 싶어서 옷을 새로 살고 싶어. (I want to stay cool this summer, so I want to live anew clothes.)
Type Desc.	~~이번 여름에는 시원하게 보내고 싶어서~~ 옷을 새로 사고 싶어요. (~~I want to stay cool this summer, so~~ I would like to buy new clothes.)

Table 4: ChatGPT prediction examples (a) according to prompting methods and (b) based on external information types for element error type. Type Desc. indicates the provision of the error type descriptions.

Since it is unknown which error type the given source sentence falls into during inference, all detailed error types are provided. For example, for the auxiliary predicate error type, a description like "Using auxiliary verbs that do not conform to grammar. This occurs..." is contained.

Providing random knowledge results in performance degradation for all models for the auxiliary predicate type and shows only slight performance increases or decreases for the other error types across models. This indicates the high informativeness of providing knowledge that is deliberately relevant to the given source sentence. Moreover, providing descriptions of error type classifications instead of knowledge about dictionary definitions generally results in poor performance across most models and error types. For instance, the Gemini model exhibits a performance decline of over 20%p for the 'Element' type compared to the baseline method, which can be interpreted as the provided descriptions acting as noise.

Furthermore, Figure 2 demonstrates the results for additional knowledge provision methods in ChatGPT. 'Keyword Only' refers to cases where only key phrases are provided in the input prompt as additional clues, without corresponding knowledge in the dictionary. In the 'POS Tag Keyword' method, keywords are extracted through the POS tagger (Park and Cho, 2014) instead of LLMs. In comparison with various knowledge provision methods, KAGEC continues to exhibit the most superior performance. Therefore, the effectiveness of KAGEC in enhancing understanding of the context through the provision of relevant semantic information can be observed.

6.2 Qualitative Analysis

Table 4 shows the generated results according to each method for the 'Element' error type in the ChatGPT model[5]. According to Table 4 (a), the baseline method, which provides only the task instruction alongside the source sentence, arbitrarily deletes a dependent clause. Moreover, existing multi-step reasoning-based prompting methods are observed to have errors in intermediate steps that negatively affect the final response generation. The zero-shot CoT method exhibits errors in creating reasoning paths that propagate to the final correction result. Similarly, the Task decomposition method, which breaks down the correction task into several sub-tasks for multi-step reasoning, shows comparable errors. An error made in Step 1, replacing '싶어 (want)' with '샀어 (bought)', leads

[5]For qualitative analysis of other error types and models, refer to Appendix E

to an incorrect modification of the sentence's verb, and this mistake from Step 1 is directly applied when generating the entire target sentence in Step 2. In contrast, when applying KAGEC, an appropriate contextually fitting correction is performed from '싶어 (want to)' to '사고 싶어 (want to buy)' based on salient key phrases.

According to Table 4 (b), providing random external knowledge generates contextually inappropriate results. For example, it is observed that the verb '살다 (live)', which does not match the noun '옷 (clothes)' at all, is newly added as the main verb. Furthermore, unlike when dictionary definitions are provided as knowledge, feeding descriptions for GEC error types alongside the input, similar to baseline prompting, results in the arbitrary deletion of the original sentence's dependent clause. This implies that input and relevant external knowledge, which is deliberately retrieved as in KAGEC, facilitates the model's understanding for generation.

6.3 Generalization for Various LLMs

In this work, experiments were conducted on various LLMs. Based on the experimental results, we discuss the generalizability of KAGEC across different LLMs. Particularly, according to Table 2, while most models show performance improvement when applying KAGEC, the performance of the Llama2-13B model in the 'Auxiliary predicate' type is an exception, displaying a decline.

Firstly, for the 'Auxiliary predicate' type, the Llama2-13B model exhibits the highest performance with the baseline method. The application of other methods, including KAGEC, tends to result in a minimum decrease of 10 points in BLEU and 7 points in GLEU score. In contrast, in the 70B model, while the performance drops compared to the baseline when applying CoT and Decomposed, similar to the 13B, the application of KAGEC results in a significant increase of 12.37%p in BLEU and 14.59%p in GLEU.

Hence, these outcomes can be interpreted as stemming from the differences in the number of parameters between the models. This is fundamental because the 70B model's capability to follow instructions is significantly superior to that of the 13B model, as indicated in Table 5. Although both the 13B and 70B models were provided with the same knowledge subset by extracting the same key phrases, the final generated outcome was incorrect for the 13B model, whereas the 70B model produced the correct answer.

Source Sentence	
그런데 제가 환자 모르게 비밀로 하고 싶엇서요. (But I would have wanted to keep it a secret without the patient knowing.)	
Target Sentence	
그런데 제가 환자 모르게 비밀로 하고 싶어서요. (But I want to keep it a secret without the patient knowing.	
Llama2-13B Prediction	
Baseline	제가 환자 모르게 비밀로 하고 싶습니까? (correction: 제가 instead of 그런데, 싶습니까 instead of 싶엇서요)
KAGEC	– *Top-n*: ['환자', '모르게', '비밀로', '하고', '싶었어요'] (['patient', 'without knowing', 'secret', 'want to', 'do']) {{retrieved definitions}} Output: 그런데 제가 환자 치료를 받고 싶엇서요. (Corrected sentence using the keywords and their explanations) Explanation: * 환자 (hwan-ja) means "patient" in Korean. * 모르게 (mo-r-ge) means "secretly" or "privately" in Korean. * 비밀 (bi-mil) means "secret" or "private" in Korean. * 하고 (ha-go) is a filler word used to indicate a desire or intention. Therefore, the corrected sentence should be: "그런데 제가 환자 치료를 받고 싶엇서요" (I want to receive treatment as a patient privately).
Llama2-70B Prediction	
Baseline	그런데 제가 환자 모르게 비밀로 하고 싶엇니다. (Correction: 싶엇서요 should be 싶엇니다)
KAGEC	– *Top-n*: ['환자', '모르게', '비밀로', '하고', '싶었어요'] (['patient', 'without knowing', 'secret', 'want to', 'do']) {{retrieved definitions}} Correction: 그런데 제가 환자 모르게 비밀로 하고 싶어서요. Explanation: 싶어서요 is more natural way to express the past tense in informal speech.

Table 5: Prediction examples of Llama2 models (13B and 70B) for auxiliary predicate error type.

7 Conclusion

LLMs have demonstrated exceptional generative capabilities in GEC systems. However, their performance in non-English languages, such as Korean, remains relatively lacking. We hypothesize that relying solely on the parametric knowledge of LLMs makes it challenging to understand the given context in the Korean GEC task thoroughly. To address this, we propose the Knowledge-Augmented GEC (KAGEC) method, which consists as follows: it extracts the top-n key phrases, retrieves definitions for these terms, and reconstructs the input for generation. Afterward, this augmented input is fed into LLMs to enrich the generated answer (target sentence). Moreover, through detailed validation of LLMs, we investigate the error types requiring knowledge augmentation for the Korean GEC task. Experimental results indicate that the proposed KAGEC method consistently leads to performance improvements across various LLMs and the target error types. Through KAGEC, we can effectively support the inherent capabilities of LLMs by augmenting non-parametric knowledge.

Limitations

In conducting our experiments, we adhered to a rigorous and transparent process to ensure a fair comparison across different methods and models. We meticulously designed our experimental setup to minimize biases and external influences that could affect the outcomes. Despite our efforts to maintain the highest standards of experimental integrity, it is important to acknowledge the limitations inherent in our study.

Firstly, the variability in results due to the dependency on the version of the models used cannot be overlooked. Given the dynamic nature of Large Language Models (LLMs), the outputs generated can vary significantly with each execution, making it challenging to guarantee consistent results across runs.

Furthermore, our validation experiments faced constraints due to the limited availability of dataset resources. Specific segmented error types, such as 'Typing Language Error', were excluded from our validation targets due to the difficulty in acquiring a sufficient number of examples for few-shot validation. This limitation highlights the need for more comprehensive datasets that cover a broader range of error types to enhance the robustness of GEC systems.

The use of APIs introduced another limitation related to cost, which restricted our ability to employ a diverse range of models for validation, particularly for detailed error types. Our study primarily relied on ChatGPT, and while this model provided valuable insights, the inclusion of additional models could have enriched our findings.

Lastly, our experiments focused exclusively on the Korean language. While this focus allowed for an in-depth exploration of GEC in a non-English context, it also underscores the necessity for future research to extend to multi-lingual settings or to investigate aspects related to language transfer. Expanding the scope of research to include multiple languages would not only broaden the applicability of our findings but also contribute to the development of more versatile and effective GEC systems.

Ethics Statement

In this section, we discuss the key ethical considerations related to the approach proposed in our work.

Firstly, the dataset K-nct (Koo et al., 2022) utilized for validating LLMs and experimenting with prompting methods in this work is a linguistic resource that includes sentences with grammatical errors and their corrected versions within the Korean domain. As such, it does not entail privacy issues, since it consists of constructed examples specifically designed for research purposes and does not contain personally identifiable information or sensitive data.

Next, our validation and experimental work involves generating responses using LLMs. While careful attention has been paid to prompt engineering to mitigate risks, there still remains the potential for issues related to the generated content. Given the nature of prompt engineering, which is heavily influenced by the biases and preferences of human engineers, not all intended effects may be fully controlled. Moreover, with updates and versions of LLMs evolving, there is a possibility that inappropriate results could be generated. This underscores the importance of ongoing monitoring and evaluation to ensure that the outputs remain relevant and appropriate, adhering to ethical standards in natural language processing and AI research.

We acknowledge these ethical considerations and strive to conduct our research responsibly, with an awareness of the implications of our work on broader societal and ethical norms.

Acknowledgements

This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425). This work was supported by Institute for Information communications Technology Planning Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220369, 2022-0-

00369 (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge)

References

Abhijeet Awasthi, Sunita Sarawagi, Rasna Goyal, Sabyasachi Ghosh, and Vihari Piratla. 2019. Parallel iterative edit models for local sequence transduction. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4260–4270, Hong Kong, China. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical error correction: A survey of the state of the art. Computational Linguistics, 49(3):643–701.

Hannan Cao, Wenmian Yang, and Hwee Tou Ng. 2021. Grammatical error correction with contrastive learning in low error density domains. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4867–4874, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Tao Fang, Shu Yang, Kaixin Lan, Derek F Wong, Jinpeng Hu, Lidia S Chao, and Yue Zhang. 2023. Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. arXiv preprint arXiv:2304.01746.

Zifa Gan, Hongfei Xu, and Hongying Zan. 2021. Self-supervised curriculum learning for spelling error correction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3487–3494.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In International conference on machine learning, pages 3929–3938. PMLR.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7969–7992, Singapore. Association for Computational Linguistics.

Masahiro Kaneko, Sho Takase, Ayana Niwa, and Naoaki Okazaki. 2022. Interpretability for language learners using example-based grammatical error correction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7176–7187, Dublin, Ireland. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.

Muhammad Khalifa, Lajanugen Logeswaran, Moontae Lee, Honglak Lee, and Lu Wang. 2023. Few-shot reranking for multi-hop qa via language model prompting. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15882–15897.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. Advances in neural information processing systems, 35:22199–22213.

Seonmin Koo, Chanjun Park, Jinsung Kim, Jaehyung Seo, Sugyeong Eo, Hyeonseok Moon, and Heuiseok Lim. 2023. KEBAP: Korean error explainable benchmark dataset for ASR and post-processing. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4798–4815, Singapore. Association for Computational Linguistics.

Seonmin Koo, Chanjun Park, Jaehyung Seo, Seungjun Lee, Hyeonseok Moon, Jungseob Lee, and Heuiseok Lim. 2022. K-nct: Korean neural grammatical error correction gold-standard test set using novel error type classification criteria. IEEE Access, 10:118167–118175.

Sang Yun Kwon, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Beyond english: Evaluating llms for arabic grammatical error correction. arXiv preprint arXiv:2312.08400.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. arXiv preprint arXiv:1906.00300.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems, 33:9459–9474.

Mengsay Loem, Masahiro Kaneko, Sho Takase, and Naoaki Okazaki. 2023. Exploring effectiveness of gpt-3 in grammatical error correction: A study on performance and controllability in prompt-based methods. arXiv preprint arXiv:2305.18156.

Junghwan Maeng, Jinghang Gu, and Sun-A Kim. 2023. Effectiveness of chatgpt in korean grammatical error correction. In Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation, pages 464–472.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2020. Generation-augmented retrieval for open-domain question answering. arXiv preprint arXiv:2009.08553.

Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2015. Ground truth for grammatical error correction metrics. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 588–593.

OpenAI-Blog. 2022. Chatgpt: Optimizing language models for dialogue.

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.

Eunjeong L. Park and Sungzoon Cho. 2014. Konlpy: Korean natural language processing in python. In Proceedings of the 26th Annual Conference on Human Cognitive Language Technology, Chuncheon, Korea.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8):9.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics, 11:1316–1331.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021a. A simple recipe for multilingual grammatical error correction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 702–707, Online. Association for Computational Linguistics.

Sascha Rothe, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2021b. A simple recipe for multilingual grammatical error correction. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 702–707.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. arXiv preprint arXiv:2305.15294.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. arXiv preprint arXiv:2301.12652.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. In Findings of the Association for Computational Linguistics: ACL 2022, pages 744–756, Dublin, Ireland. Association for Computational Linguistics.

Maksym Tarnavskyi, Artem Chernodub, and Kostiantyn Omelianchuk. 2022. Ensembling and knowledge distilling of large sequence taggers for grammatical error correction. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3842–3852.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. arXiv preprint arXiv:2305.04091.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35:24824–24837.

Haoran Wu, Wenxuan Wang, Yuxuan Wan, Wenxiang Jiao, and Michael Lyu. 2023. Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark. arXiv preprint arXiv:2303.13648.

Xuan Zhang and Wei Gao. 2023. Towards llm-based fact verification on news claims with a hierarchical step-by-step prompting method. arXiv preprint arXiv:2310.00305.

Yu Zhang, Yue Zhang, Leyang Cui, and Guohong Fu. 2023. Non-autoregressive text editing with copy-aware latent alignments. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 7075–7085, Singapore. Association for Computational Linguistics.

Huan Zhao, Qian Ling, Yi Pan, Tianyang Zhong, Jin-Yu Hu, Junjie Yao, Fengqian Xiao, Zhenxiang Xiao, Yutong Zhang, San-Hua Xu, et al. 2023. Ophtha-llama2: A large language model for ophthalmology. arXiv preprint arXiv:2312.04906.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. arXiv preprint arXiv:2205.10625.

A Error Type Descriptions

The following are descriptions of some of the segmented error types for Grammatical Error Correction (GEC) in the K-NCT dataset (Koo et al., 2022), used for validation in section 3 and experiments in section 5.

- Spacing: This violates the Korean spacing rules.
- Punctuation: Punctuation marks are not attached in Korean sentences or are attached in the wrong position.
- Numerical: Cardinal number indicating quantity and the ordinal number indicating the order are in error.
- Remove: Some words are not recognized, or endings or suffixes are omitted.
- Addition: Same word is repeated, or an unused postposition or ending is added.
- Rotation replace: Order of syllables changes within a one phrase.
- Separation: Separating consonants and vowels in characters.
- Foreign and conversion: Writing differently from the standard foreign language pronunciation.
- Consonant vowel conversion: Spelling error in non-speaking alphabet units.
- Grapheme-to-phoneme (G2P): Writing spellings according to pronunciation.
- Element: The Korean sentence components are not equipped or the word order is not correct.
- Tense: Using a verb that does not match the tense.
- Postposition: Probing that does not fit the grammar.
- Suffix: Using an ending that is not grammatically correct.
- Auxiliary predicate: Using an auxiliary verb that is not grammatically correct.
- Behavioral: Expressions that the subject cannot perform.
- Neologism: Using grammar or new words that are not included in the existing grammar system.

B Additional Validation of Error Types

Type	# shots									
	0		1		4		8		16	
	BLEU	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**
Spacing	45.93	45.84	63.32	61.40	76.31	74.07	78.89	76.18	80.22	77.71
punctuation	48.97	49.14	63.11	60.69	80.59	78.75	84.59	82.71	86.64	85.14
Numerical	42.92	38.98	52.61	46.95	64.16	58.34	68.68	63.85	73.45	67.63
Remove	45.05	44.33	53.10	51.39	57.24	54.85	64.98	61.93	67.74	64.98
Addition	48.01	47.37	59.14	56.99	68.10	65.59	76.28	73.18	76.31	73.33
Rotation replace	51.43	50.93	58.72	55.92	67.19	64.21	77.97	74.20	76.61	73.66
Separation	57.01	58.17	75.31	74.58	85.85	85.57	87.55	87.22	85.06	84.89
Foreign and conversion	47.68	45.16	58.23	56.03	75.20	71.59	77.18	72.88	80.34	76.43
Consonant-vowel conversion	52.55	50.96	60.50	57.91	70.98	68.88	73.87	71.69	78.15	75.53
G2P	47.40	45.44	59.93	57.17	71.93	69.98	75.36	73.12	77.64	74.77
Element	38.59	38.85	47.83	46.45	56.84	56.39	63.99	63.11	64.13	63.52
Tense	45.42	39.39	50.66	44.04	60.24	49.33	71.31	61.02	66.75	58.96
Postposition	49.44	47.25	62.86	59.10	67.93	65.32	73.09	71.01	77.49	73.85
Suffix	48.69	48.04	62.86	62.26	73.79	73.65	74.94	75.05	76.76	76.86
Auxiliary predicate	51.64	52.07	52.18	53.04	66.73	67.38	72.29	74.05	70.37	71.11
Behavioral	38.90	35.31	51.19	45.74	53.03	47.33	57.51	51.32	57.54	51.81
Neologism	39.32	39.12	47.32	46.42	61.56	59.54	66.68	64.15	64.53	62.06
Avg.	47.05	46.06	57.07	54.97	67.60	65.39	72.66	70.47	73.53	71.34

Table 6: ChatGPT-3.5 results for different error types in the GEC task. Avg. denotes the average performance of all error types. G2P refers grapheme-to-phoneme. we set the temperature as 0.5.

Type	# shots									
	0		1		4		8		16	
	BLEU	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**	**BLEU**	**GLEU**
Spacing	44.92	44.73	63.36	61.68	76.93	74.69	79.84	77.39	80.13	77.50
Punctuation	45.56	45.36	64.73	63.18	80.01	78.34	83.66	81.94	88.60	87.12
Numerical	42.73	39.10	51.99	46.43	64.14	57.95	68.72	63.73	72.20	66.47
Remove	43.34	42.50	52.79	50.71	57.63	55.56	65.14	63.06	67.59	64.70
Addition	50.90	48.75	59.00	56.89	71.32	68.45	76.88	74.15	76.53	73.61
Rotation replace	46.89	45.79	62.89	59.35	70.81	66.73	76.71	73.22	77.82	74.64
Separation	59.41	61.52	77.39	76.83	84.46	84.31	86.49	86.31	86.33	85.78
Foreign and conversion	45.36	43.96	59.78	57.35	74.87	71.93	77.86	74.21	80.34	76.86
Consonant-vowel conversion	52.20	50.91	60.84	57.45	72.11	70.07	73.66	71.75	77.74	75.16
G2P	46.70	44.75	58.45	55.62	71.74	70.47	79.01	76.23	79.11	76.27
Element	39.82	40.17	46.21	46.43	54.56	53.89	58.73	57.24	58.17	57.43
Tense	44.35	37.55	50.76	44.61	61.49	51.29	65.87	54.93	68.02	59.35
Postposition	41.25	39.24	65.25	61.75	72.38	70.61	76.11	72.90	75.71	73.52
Suffix	45.92	45.43	64.13	63.58	74.10	74.97	75.66	75.58	80.13	80.61
Auxiliary predicate	48.12	50.23	54.65	56.04	66.50	66.98	72.46	74.45	75.15	76.73
Behavioral	39.52	36.11	47.01	42.80	55.91	51.35	59.80	55.11	54.89	49.94
Neologism	37.92	37.40	52.26	49.51	58.96	56.74	66.04	62.67	65.61	63.21
Avg.	45.52	44.55	57.76	55.78	68.07	66.03	72.34	70.11	73.67	71.61

Table 7: ChatGPT-3.5 results for different error types in the GEC task. Avg. denotes the average performance of all error types. G2P refers grapheme-to-phoneme. we set the temperature as 0.8.

C Experimental Setup

Dataset. For the experiments, we utilize the K-NCT dataset mentioned in section 3, which was also used for validation earlier (Koo et al., 2022). This dataset serves as a gold-standard test set for the Korean GEC task, providing a newly defined set of various error type classification guidelines along with resources. The K-NCT dataset is publicly available for research purposes, enabling anyone to utilize it. In the released dataset, the authors propose error type classification standards for Korean GEC research. These proposed types are based on four significant criteria and are divided into 23 sub-categories, considering aspects of factuality and diversity. The statistical information for the text is provided in Table 8

K-NCT	Test	
	Error sentence	**Correct sentence**
# of sents	3,000	3,000
# of tokens	129,798	129,886
# of words	31,183	31,700
avg of SL △	43.27	43.29
avg of WS	10.39	10.57
avg of SS	9.39	9.57

Table 8: Statistics of our K-NCT dataset. # of sents/tokens/words: number of sentences/tokens/words; △ avg of SL/WS/SS: average of sentence length/words/spaces per sentence.

To construct a knowledge pool for retrieval, we select definitions for each term from the National Institute of Korean Language's Standard Korean Dictionary [6]. The National Institute of Korean Language is an authoritative institution that establishes norms for Korean linguistics [7]. The National Institute of the Korean Language is a CC BY-SA 2.0 KR license. This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, even for commercial purposes. If others remix, adapt, or build upon the material, they must license the modified material under identical terms. We acknowledge and utilize the licensing information solely for academic and research purposes. Table 9 shows National Institute of Korean Language's Standard Korean Dictionary dataset

	Error sentence	**Correct sentence**
# of sents	568,128	568,128
# of tokens	1737,497	25,119,115
# of words	568,129	5,960,778
avg of SL △	3.06	44.21
avg of WS	1	10.49
avg of SS	1.76	9.49

Table 9: Statistics of our National Institute of Korean Language's Standard Korean Dictionary dataset. # of sents/tokens/words: number of sentences/tokens/words; △ avg of SL/WS/SS: average of sentence length/words/spaces per sentence.

Models. For the experiments, the LLMs adopted include ChatGPT (OpenAI-Blog, 2022), Llama2 (Touvron et al., 2023), and Gemini (Team et al., 2023). Specifically, for Llama2, to investigate the differences according to the parameter size within the same LLM family, experiments were conducted on both the 13B and 70B models. The experiments with ChatGPT and Llama2 models were carried out using the OpenAI API and Llama2-api [8], respectively. For ChatGPT, the 1106 version was used in the experiments. The Gemini model results were generated through the Google AI studio API [9].

Evaluation Metrics. For evaluation, the output answers generated by the models are measured against reference sentences using BLEU (Papineni et al., 2002) and GLEU (Napoles et al., 2015) scores, which are commonly utilized as evaluation indicators in various deep learning-based grammatical correction studies.

Since BLEU score evaluates word-based similarity against a reference text, its applicability as an assessment tool relies on the presumption that it aligns with and anticipates the practical effectiveness of such systems, gauged either through task performance or user contentment. Also, as GLEU score is an n-gram-based metric, it penalizes hypothesis n-grams that overlap with the source but not the reference and rewards hypothesis n-grams that overlap with the reference but not the source.

[6] https://opendict.korean.go.kr/main
[7] https://www.korean.go.kr/
[8] https://github.com/llamaapi/llamaapi-python
[9] https://ai.google.dev

D Prompt Templates

Task Instruction
zero-shot `Do grammatical error correction on all the following sentences I type in the conversation. Always answer in Korean.` —– # few-shots `Referring to the example, do grammatical error correction that fit the given sentences. An example of doing grammatical error correction is as follows: {{examples}}`
Input (error sentence)
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 싶어. `(I want to stay cool this summer, so I want to new clothes.)`

Table 10: Prompt template for validation and baseline prompting. Under the # zero-shot is utilized as both a verification and baseline prompt. In the verification process under the few-shot setting, add the # few-shots prompt.

Task Instruction
Phase 1 `Do grammatical error correction on all the following sentences I type in the conversation. Let's think step-by-step. Always answer in Korean.` # Phase 2 `{{reasoning path}}` —– `Do grammatical error correction that fit the given sentences. Let's think step-by-step. Always answer in Korean.`
Input (error sentence)
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 싶어. `(I want to stay cool this summer, so I want to new clothes.)`

Table 11: Prompt template for chain-of-thought (CoT) prompting.

Task Instruction
Phase 1 `Please detect word with any grammatical errors in the given sentence. Always answer in Korean.` # Phase 2 `{{reasoning path}}` `---` `Based on detected errors, do grammatical error correction that fit the given sentences.`
Input (error sentence)
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 싶어. `(I want to stay cool this summer, so I want to new clothes.)`

Table 12: Prompt template for task decomposition prompting.

Task Instruction
Phase 1 `Please extract the top 5 most significant keywords from the given sentence. Always answer in Korean.` # Phase 2 `Referring to the keywords and their explanations, do grammatical error correction that fit the given sentences.` `The keywords and their explanations are as follows.: {{examples}}` `Always answer in Korean.`
Input (error sentence)
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 싶어. `(I want to stay cool this summer, so I want to new clothes.)`

Table 13: Prompt template for KAGEC prompting.

Task Instruction
`Referring to the keywords and their explanations, do grammatical error correction that fit the given sentences.` `The keywords and their explanations are as follows.: {{random examples}}` `Always answer in Korean.`
Input (error sentence)
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 싶어. `(I want to stay cool this summer, so I want to new clothes.)`

Table 14: Prompt template for random knowledge prompting.

Task Instruction

```
Referring to the error definition, do grammatical error correction that fit the given sentences.
The error definition is as follows:
- spacing: This violates the Korean spacing rules.
- punctuation: This occurs when punctuation marks are not attached in the sentence or misplaced.
- numerical: This occurs when a cardinal number indicates quantity and an ordinal number indicates
order.

(...)

- neologism: Using grammar or new words that are not included in the existing grammar system.
Always answer in Korean.
```

Input (error sentence)

```
이번 여름에는 시원하게 보내고 싶어서 옷을 새로 싶어.
(I want to stay cool this summer, so I want to new clothes.)
```

Table 15: Prompt template for error type description prompting.

E Additional Qualitative Analysis

Following from section 6.2, we provide additional generated results for various Large Language Models (LLMs) and error types according to each method.

Source Sentence	
줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠졌다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and are distributed it to the people standing in line.)	
Target Sentence	
줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠줬다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and distributed it to the people standing in line.)	
ChatGPT Prediction	
Baseline	줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠 주었다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and divided and provided it to the people standing in line.)
CoT	줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠 주었다. "미리 맡겨둔 돈을 환전한 듯한" 이 부분에 반말로 되어 있고, "나눠졌다" 동사는 주어에 맡게 되었으므로 "나눠 주었다" 로 변경하였습니다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and divided and provided it to the people standing in line. The part where it said "seeming to have exchanged the money in advance" was written in casual language, and the verb "distributed" was adjusted according to the subject, so it was changed to "divided and provided.")
Decomposed	– *Step 1* "미리 맡겨둔 돈을 환전한 듯한" 부분에 문법 오류가 있습니다. (There is a grammatical error in the part "seeming to have exchanged the money in advance".) – *Step 2* 줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 미리 맡겨둔 돈을 환전한 듯한 5만원짜리 지폐뭉치를 꺼내 나눠주었다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and divided and provided it to the people standing in line.)
KAGEC	– *Top-n*: ['중국인', '가이드', '돈', '환전', '지폐'] (['Chinese', 'guide', 'money', 'exchange', 'bill']) {{retrieved definitions}} 줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠줬다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and distributed it to the people standing in line.)

Table 16: ChatGPT prediction examples by prompting methodology for auxiliary predicate error type

Random	– *Top-n*: [‘꿈길’, ‘교가’, ‘조드’, ‘양지군’, ‘방음’] ([‘dream road’, ‘school song’, ‘dzud’, ‘Yangjigun’, ‘soundproof’]) {{random definitions}} 줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠졌다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and are distributed it to the people standing in line.)
Type Desc.	줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠졌다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and are distributed it to the people standing in line.)
KAGEC	– *Top-n*: [‘중국인’, ‘가이드’, ‘돈’, ‘환전’, ‘지폐’] ([‘Chinese’, ‘guide’, ‘money’, ‘exchange’, ‘bill’]) {{retrieved definitions}} 줄을 서 있는 그들에게 중국인 가이드가 큰 가방에서 (미리 맡겨둔 돈을 환전한 듯한) 5만원짜리 지폐뭉치를 꺼내 나눠줬다. (Chinese guide pulled out a bundle of 50,000-won bills from a large bag (seeming to have exchanged the money in advance) and distributed it to the people standing in line.)

Table 17: ChatGPT prediction examples for auxiliary predicate error type based on External Information Types. Source/Target/Baseline results are consistent with Table 16

Source Sentence	
실수하지 않는 사과는 아무 일도 못 해요. (An apple that doesn't make mistakes can't do anything.)	
Target Sentence	
실수하지 않는 사람은 아무 일도 못 해요. (People who don't make mistakes can't do anything.)	
ChatGPT Prediction	
Baseline	실수하지 않는 사과는 아무 일도 못 해요. (An apple that doesn't make mistakes can't do anything.)
CoT	"실수하지 않는 사과는 아무 일도 못 해요" 는 올바른 문장이 아니에요. "실수하지 않는 사람은 없어요" 가 더 올바른 표현이에요. "사과" 대신에 "사람"을 써야 해요. ("An apple that doesn't make mistakes can't do anything." is not a correct sentence. "No one makes mistakes" is a more correct expression. Instead of "apple" you should use "person".)
Decomposed	– *Step 1* 해당 단어에 문법적인 오류는 없습니다. (There are no grammatical errors in that word.) – *Step 2* 실수하지 않는 사과는 아무 일도 못 해요. (An apple that doesn't make mistakes can't do anything.)
KAGEC	– *Top-n*: ['실수', '사과', '못', '일', '해요'] (['mistake', 'apple', 'don't', 'anything', 'do']) {{retrieved definitions}} 실수하지 않는 사람은 아무 일도 못 해요. (People who don't make mistakes can't do anything.)

Table 18: ChatGPT prediction examples by prompting methodology for behavior error type.

Random	– *Top-n*: ['기호론자', '준', '생디이', '신정', '잔디'] (['symbolist', 'equivalent', 'saengdiei', 'new year', 'grass']) {{random definitions}} 실수하지 않는 기호론자는 아무 일도 못 해요. (Symbolist who don't make mistakes can't do anything.)
Type Desc.	실수하지 않는 사과는 아무 일도 못 해요. (An apple that doesn't make mistakes can't do anything.)
KAGEC	– *Top-n*: ['실수', '사과', '못', '일', '해요'] (['mistake', 'apple', 'don't', 'anything', 'do']) {{retrieved definitions}} 실수하지 않는 사람은 아무 일도 못 해요. (People who don't make mistakes can't do anything.)

Table 19: ChatGPT prediction examples for behavior error type based on External Information Types. Source/Target/Baseline results are consistent with Table 18

Source Sentence	
올스타전 전야제는 NBA 중계 중 코피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 전한다 생중계로. (The eve of the All-Star Game will be locally broadcast live by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during NBA broadcasts.)	
Target Sentence	
올스타전 전야제는 NBA 중계 중 코피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 생중계로 전한다. (The eve of the All-Star Game will be broadcast live locally by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during NBA broadcasts.)	
Gemini Prediction	
Baseline	올스타전 전야제는 NBA 중계 중 코피를 쏟아 유명세를 타게 된 조현일 해설위원과 박찬웅 캐스터가 현지에서 생중계로 전한다. (The eve of the All-Star Game will be broadcast live locally by commentator Hyun-il Cho and caster Chan-woong Park, who are became famous for suffering nosebleeds during NBA broadcasts.)
CoT	올스타전 전야제는 NBA 중계 중 커피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 생중계로. "전한다 생중계로"는 오류입니다. "전한다"와 "생중계로"는 둘 다 동사이므로 하나만 사용해야 합니다. "전한다"를 생략하고 "생중계로"만 사용하면 됩니다. (The eve of the All-Star Game will be broadcast live locally by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during NBA broadcasts. "broadcast live locally" is an error. are both verbs, so only one should be used. You can omit "locally" and just use "broadcast live.")
Decomposed	– *Step 1* '전한다 생중계로' (locally broadcast live) – *Step 2* 올스타전 전야제는 NBA 중계 중 코피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 생중계로 전한다. (The eve of the All-Star Game will be broadcast live locally by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during NBA broadcasts.)
KAGEC	– *Top-n*: ['올스타전', '전야제', '중계', '조현일'] (['All-Star Game', 'eve', 'broadcasts', 'Hyun-il Cho']) {{retrieved definitions}} 올스타전 전야제는 NBA 중계 중 코피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 생중계로 전한다. (The eve of the All-Star Game will be broadcast live locally by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during NBA broadcasts.)

Table 20: Gemini prediction examples by prompting methodology for element error type.

Random	– *Top-n*: ['매치리', '당월', '호남', '배미', '재정비'] (['Matchiri', 'current month', 'Honam', 'Baemi', 'Reorganization']) {{random definitions}} 올스타전 전야제는 NBA 중계 중 커피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 재정비로 생중계한다. (The eve of the All-Star Game will be broadcast live by reorganization by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during NBA broadcasts.)
Type Desc.	spacing: ~~올스타전~~ 전야제는 NBA 중계 중에 코피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 생중계로 전한다. (spacing: : The eve of ~~the All-Star Game~~ will be broadcast live locally by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during the NBA broadcasts.)
KAGEC	– *Top-n*: ['올스타전', '전야제', '중계', '조현일'] (['All-Star Game', 'eve', 'broadcasts', 'Hyun-il Cho']) {{retrieved definitions}} 올스타전 전야제는 NBA 중계 중 코피를 쏟아 유명세를 탄 조현일 해설위원과 박찬웅 캐스터가 현지에서 생중계로 전한다. (The eve of the All-Star Game will be broadcast live locally by commentator Hyun-il Cho and caster Chan-woong Park, who became famous for suffering nosebleeds during NBA broadcasts.)

Table 21: Gemini prediction examples based on External Information Types. Source/Target/Baseline results are consistent with Table 20

Source Sentence	
현재까지 확보한 업체들부터 빨리 광고를 시작하 보죠. (Let's commence advertising quickly starts from the vendors we have secured thus far.)	
Target Sentence	
현재까지 확보한 업체들부터 빨리 광고를 시작해 보죠. (Let's commence advertising quickly starting from the vendors we have secured thus far.)	
Gemini Prediction	
Baseline	빠르게 광고를 시작하도록 하겠습니다. (I will start advertising quickly.)
CoT	현재까지 확보한 업체들부터 빠르게 광고를 시작해 보죠. 1. "확보한" should be "확보한" because it is a past participle of the verb "확보하다." 2. "빨리" should be "빠르게" because it is an adverb and needs to be in the adverbial form. 3. "시작하 보죠" should be "시작해 보죠" because it is the polite form of the verb "시작하다.
Decomposed	– *Step 1* 시작하 보죠 (starts) – *Step 2* 현재까지 확보한 업체부터 적극적으로 광고를 진행해 보시기 바랍니다. (We recommend that you actively advertise with the companies you have secured so far.)
KAGEC	– *Top-n*: ['광고', '확보', '업체', '시작', '현재'] (['advertising', 'secure', 'vendor', 'start', 'thus far']) {{retrieved definitions}} 현재까지 확보한 업체들부터 빨리 광고를 시작해 보죠. (Let's commence advertising quickly starting from the vendors we have secured thus far.)

Table 22: Gemini prediction examples by prompting methodology for auxiliary predicate error type.

Random	– *Top-n*: ['제남', '골통대', '종단면', '비의', '담합'] (['Jinam', 'Goltongdae', 'longitudinal section', 'Biui', 'collusion']) {{random definitions}} 현재까지 확보한 업체들부터 빨리 광고를 시작하세요. (Start advertising quickly with the vendors you have secured so far.)
Type Desc.	spacing: 현재까지 확보한 업체들부터 빨리 광고를 시작하보죠. (spacing: Let's commence advertising quickly starts from the vendors we have secured thus far.)
KAGEC	– *Top-n*: ['광고', '확보', '업체', '시작', '현재'] (['advertising', 'secure', 'vendor', 'start', 'thus far']) {{retrieved definitions}} 현재까지 확보한 업체들부터 빨리 광고를 시작해 보죠. (Let's commence advertising quickly starting from the vendors we have secured thus far.)

Table 23: Gemini prediction examples for auxiliary predicate based on external information types. Source/Target/Baseline results are consistent with Table 22

Source Sentence	
상자가 포장하면 따로 돈을 내야 합니다. (When the box does the packaging, an additional fee must be paid separately.)	
Target Sentence	
상자에 포장하면 따로 돈을 내야 합니다. (When packaging in a box, an additional fee must be paid separately.)	
Gemini Prediction	
Baseline	문장을 수정하였습니다: 상자가 포장되면 따로 돈을 내야 합니다. (When the box is packed, an additional fee must be paid separately.)
CoT	상자가 포장하면 따로 돈을 내야 합니다. 포장지 서비스는 개별적으로 제공되며, 상자가 포장되면 따로 돈을 내야 합니다. 해당 문장은 "포장지 서비스는 개별적으로 제공되며"에 해당하는 내용을 기술하고자 하는 문장이며, "개별적으로 제공되며"에 해당하는 부분은 문장의 주어가 되지 못하므로 "포장지 서비스는 개별적으로 제공되며" 대신 "포장지 서비스는 개별로 제공되므로"으로 수정하였습니다. (When the box does the packaging, an additional fee must be paid separately. Packaging services are provided individually, and when packaging is done, an additional fee must be paid. This sentence aims to describe the part ‘packaging services are provided individually,’ and as it cannot serve as the subject of the sentence, ‘Packaging services are provided individually’ has been replaced with ’Packaging services are provided individually, so’ for clarification.)
Decomposed	– *Step 1* 포장하면 (When packaging) – *Step 2* 상자가 포장되면 따로 돈을 내야 합니다. (When the box is packed, an additional fee must be paid separately.)
KAGEC	– *Top-n*: [‘상자’, ‘포장’, ‘돈’, ‘별도’, ‘지불’] ([‘box’, ‘packaging’, ‘fee’, ‘separately’, ‘pay’]) {{retrieved definitions}} 상자에 포장하면 따로 돈을 내야 합니다. (When packaging in a box, an additional fee must be paid separately.)

Table 24: Gemini prediction examples by prompting methodology for behavior error type.

Random	– *Top-n*: [‘방수’, ‘사양토’, ‘그림엽서’, ‘조교유’, ‘이불’] ([‘Waterproofing’, ‘sandy soil’, ‘picture postcard’, ‘Jogyoyu’, ‘blankets’]) {{random definitions}} 상자가 포장하면 따로 돈을 내야 합니다. 문장에서 ’포장하면’은 오류입니다. ’포장하면’은 비주격 활용이며, 주어가 아니어서 문장의 시작에 올 수 없습니다. ’포장하면’을 주격 활용인 ’포장해도’로 바꾸어야 합니다. 수정된 문장: 상자를 포장해도 따로 돈을 내야 합니다. (When the box does the packaging, an additional fee must be paid separately. The phrase ‘packaging’ is incorrect. ‘packaging is a non-subject form, and thus cannot be used at the beginning of a sentence. It should be changed to the subject form ‘Even if packaging’ Revised sentence: Even if the box is packaged, an additional fee must be paid.)
Type Desc.	상자가 포장하면 따로 돈을 내야 합니다. (spacing) (When the box does the packaging, an additional fee must be paid separately.) (spacing)
KAGEC	– *Top-n*: [‘상자’, ‘포장’, ‘돈’, ‘별도’, ‘지불’] ([‘box’, ‘packaging’, ‘fee’, ‘separately’, ‘pay’]) {{retrieved definitions}} 상자에 포장하면 따로 돈을 내야 합니다. (When packaging in a box, an additional fee must be paid separately.)

Table 25: Gemini prediction examples for behavior error type based on external information types. Source/Target/Baseline results are consistent with Table 24

Source Sentence	
맞아요, 딱 그렇게 조정되는 높이가 것이 필요합니다. (That's right, I need the thing that height adjusts the exactly like that.)	
Target Sentence	
맞아요, 딱 그렇게 높이가 조정되는 것이 필요합니다. (That's right, I need the thing that adjusts the height exactly like that.)	
Llama2-13B Prediction	
Baseline	맞아요, 딱 그러한 조정이 필요합니다. (That's right, I need adjustments exactly like that.)
CoT	맞아요, 딱 그렇게 조정되는 높이가 것이 필요합니다. 즉, 높이가 가장 필요한 부분을 특정하는 데 사용되는 표현입니다. Correction: * 딱 should be 단 (dan) * 그렇게 should be 그렇게 (geureoke) * 조정되는 should be 조정되는 (jogeum-eun) * 높이 should be height (height) * 것이 should be 이iga (ee-ga) * 필요합니다 should be 필요합니다 (pil-yoo-ham-ni-da) So the corrected sentence would be: 맞아요, 단 그렇게 조정되는 height이 필요합니다.
Decomposed	– *Step 1* failed the perform – *Step 2* 좋은 선택을 했어요! 이제 문장을 grammarCheck해 보입니다. 높이를 조정하려면 어떤 형식의 잘못 Also, do you have any specific word or part of the sentence you want to check?</p> <p>Here are some possible grammar errors and suggestions for correction:</p> 1. 딱 그렇게 조정되는 높이가 (incorrect)</p> Corrected version: 딱 그렇게 조정되는 높이에 (using the correct particle 에 to indicate the object of the sentence)</p> 2. 높이를 (incorrect)</p> Corrected version: 높이를 조정하고 (using the verb 조정; the correct verb for "adjust")</p> 3. 높이가 (incorrect)</p> Corrected version: 높이가 조정되는 (using the correct form 조정되는 to indicate the progressive aspect of the verb)</p> Please let me know if you would like me to check any other parts of the sentence or if you have any other questions!
KAGEC	– *Top-n*: ['맞아요', '딱', '그렇게', '조정', '높이'] (['That's right', 'exactly', 'like that', 'adjusts', 'height']) {{retrieved definitions}} 괜찮아요! 딱('닥'의 방언) 그렇게 조정(가서 정벌함)되는 높이(높은 정도)를 것이 필요합니다. (It's okay! thing needs to be adjusted to just that height.).
Llama2-70B Prediction	
Baseline	맞아요, 딱 그러한 조정이 필요합니다. (That's right, I need adjustments exactly like that.)
CoT	맞아요, 딱 그렇게 조정되는 높이가 필요합니다. (Corrected sentence) Explanation: The word "것" at the end of the sentence is not necessary and can be omitted. It's a minor error, but it's always good to avoid unnecessary words in Korean sentences. Also, the word order in the sentence is already clear and easy to understand, so there's no need to add "것" at the end. Alternatively, the sentence could be revised to: 맞아요, 딱 그렇게 조정되는 높이가 필요하다. (More natural way of saying it) In this revised sentence, the word " 필요하다" is added to make the sentence more natural and concise.
Decomposed	– *Step 1* "딱 그렇게"는 문장의 뜻을 분명하게 하지 않습니다. 대신 "그렇게"나 "그렇게 되는"를 사용하세요. ("like that" does not make the meaning of the sentence clear. Instead, use "so" or "so that".) – *Step 2* 문장 중 "딱 그렇게"가 문법적으로 올바르지 않습니다. "딱"은 부정적인 의미를 가지므로 "그렇게" 나 "그렇게 되는"를 사용하여 문장을 수정하세요. 수정된 문장: 맞아요, 그렇게 조정되는 높이가 것이 필요합니다. 또한, "높이"는 숫자가 아닌 것으로 보이므로, "높이가 것이 필요합니다"는 문장이 올바르지 않습니다. "높이"를 "높이를"로 교체하여 문장을 수정하세요. 수정된 문장: 맞아요, 그렇게 조정되는 높이를 것이 필요합니다. (That's right, I need the thing that height adjusts the exactly like that.)
KAGEC	– *Top-n*: ['맞아요', '딱', '그렇게', '조정', '높이'] (['That's right', 'exactly', 'like that', 'adjusts', 'height']) {{retrieved definitions}} 맞아요, 딱 그렇게 높이가 조정되는 것이 필요합니다. (That's right, I need the thing that adjusts the height exactly like that.).

Table 26: Llama2 prediction examples by prompting methodology for element error type.

Llama2-13B Prediction	
Random	– *Top-n*: ['마웃바가지', '글리시닌', '연조', '달찰', '달천리'] (['Mautbagaji', 'glycinin', 'Yeonjo', 'Dalchal', 'Dalcheonri']) {{random definitions}} 글리시닌이 맞춤되어 있습니까? 달찰이 있습니까? 달천리에 있습니까? (Is glycinin tailored? Is there a dalchal? Are you in Dalcheon-ri?).
Type Desc.	맞아요, 딱 그렇게 조정되는 높이가 것이 필요합니다. Error: * spacing: This violates the Korean spacing rules. Correction: * 맞아요, 딱 그렇게 조정되는 높이가 것이 필요합니다. (corrected) Explanation: In Korean, it is important to maintain proper spacing between words and phrases. In the original sentence, there is no space between "맞아요" and "딱 그렇게 조정되는 높이가 것이 필요합니다." To correct this, we added a space between the two words.
KAGEC	– *Top-n*: ['맞아요', '딱', '그렇게', '조정', '높이'] (['That's right', 'exactly', 'like that', 'adjusts', 'height']) {{retrieved definitions}} 괜찮아요! 딱('닥'의 방언) 그렇게 조정(가서 정벌함)되는 높이(높은 정도)를 것이 필요합니다. (It's okay! thing needs to be adjusted to just that height.).
Llama2-70B Prediction	
Random	– *Top-n*: ['농업근로자절', '쇼코', '누렁치', '단도', '초회'] (['Agricultural Workers' Day', 'Shoko', 'Nurungchi', 'Dando', 'Chohoe']) {{random definitions}} 농업근로자절 날짜를 달리는 것이 필요합니다. (Correcting the date of the Farmers' Day holiday)
Type Desc.	There is a mistake in your sentence. Here's a corrected version: 맞아요, 딱 그렇게 조정되는 높이가 필요합니다. Error: * spacing: There should be no space between 맞아요 and 딱. * punctuation: The sentence should end with a period. Corrected sentence: 맞아요, 딱 그렇게 조정되는 높이가 필요합니다. Additional feedback: * The word 높이 (Height) should be capitalized because it is a noun. * Consider using a more polite expression instead of 맞아요 (Yes) at the beginning of the sentence. For example, 네 (Yes) or 예 (Yes) would be more appropriate.
KAGEC	– *Top-n*: ['맞아요', '딱', '그렇게', '조정', '높이'] (['That's right', 'exactly', 'like that', 'adjusts', 'height']) {{retrieved definitions}} 맞아요, 딱 그렇게 높이가 조정되는 것이 필요합니다. (That's right, I need the thing that adjusts the height exactly like that.).

Table 27: Llama2 prediction examples for element error type. based on external information types. Source/Target/Baseline results are consistent with Table 26

Source Sentence	
그런데 제가 환자 모르게 비밀로 하고 싶엇서요. (But I would have wanted to keep it a secret without the patient knowing.)	
Target Sentence	
그런데 제가 환자 모르게 비밀로 하고 싶어서요. (But I want to keep it a secret without the patient knowing.	
Llama2-13B Prediction	
Baseline	제가 환자 모르게 비밀로 하고 싶습니까? (correction: 제가 instead of 그런데, 싶습니까 instead of 싶엇서요)
CoT	그런데 제가 환자 모르게 비밀로 하고 싶어요. Explanation: * 환자 (patient) should be 의사 (doctor) to match the context of the sentence. * 모르게 (don't know) should be 密레 (secretly) to convey the intended meaning of doing something secretly. * 하고 싶엇서요 (want to do) should be 하고 싶어요 (want to do) to correct the verb ending. So, the corrected sentence would be: 그런데 제가 의사 모르게 비밀로 하고 싶어요.
Decomposed	– *Step 1* In this sentence, the word "환자" is missing an article, so it should be "환자를" instead. Additionally, the word "모르게" is not a word in Korean, so it should be replaced with "몰래" or "몰라" to convey the same meaning. – *Step 2* 제가 환자를 몰래 비밀로 하고 싶엇서요. Explanation: * 그런데 (geureon-de) is an awkward phrase and can be simplified to 데 (de) * 모르게 (mori-ge) should be replaced with 몰래 (moll-a) or 몰라 (moll-a) to convey the same meaning * 비밀로 (bim-il-ro) should be replaced with 비밀 (bim-il) to fit the context better Corrected sentence: 제가 환자를 몰래 비밀로 하고 싶엇서요.
KAGEC	– *Top-n*: ['환자', '모르게', '비밀로', '하고', '싶었어요'] (['patient', 'without knowing', 'secret', 'want to', 'do']) {{retrieved definitions}} Output: 그런데 제가 환자 치료를 받고 싶엇서요. (Corrected sentence using the keywords and their explanations) Explanation: * 환자 (hwan-ja) means "patient" in Korean. * 모르게 (mo-r-ge) means "secretly" or "privately" in Korean. * 비밀 (bi-mil) means "secret" or "private" in Korean. * 하고 (ha-go) is a filler word used to indicate a desire or intention. Therefore, the corrected sentence should be: "그런데 제가 환자 치료를 받고 싶엇서요" (I want to receive treatment as a patient privately).
Llama2-70B Prediction	
Baseline	그런데 제가 환자 모르게 비밀로 하고 싶엇니다. (Correction: 싶엇서요 should be 싶엇니다)
CoT	그런데 제가 환자 모르게 비밀로 하고 싶어요. Explanation: * 환자 (patient) should be 의사 (doctor) to match the context of the sentence. * 모르게 (don't know) should be 密레 (secretly) to convey the intended meaning of doing something secretly. * 하고 싶엇서요 (want to do) should be 하고 싶어요 (want to do) to correct the verb ending. So, the corrected sentence would be: 그런데 제가 의사 모르게 비밀로 하고 싶어요.
Decomposed	– *Step 1* grammatical error, '환자' should be '환자입니다' to match the honorific verb ending. – *Step 2* 그런데 제가 환자입니다 모르게 비밀로 하고 싶엇서요. (Corrected sentence) Explanation: In Korean, when using the honorific verb ending (-습니다, -입니다, etc.), the subject of the sentence must be in the honorific form as well. In this case, "환자" should be "환자입니다" to match the honorific verb ending "모르게" (which is also in the honorific form).
KAGEC	– *Top-n*: ['환자', '모르게', '비밀로', '하고', '싶었어요'] (['patient', 'without knowing', 'secret', 'want to', 'do']) {{retrieved definitions}} Correction: 그런데 제가 환자 모르게 비밀로 하고 싶어서요. Explanation: 싶어서요 is more natural way to express the past tense in informal speech.

Table 28: Llama2 prediction examples by prompting methodology for auxiliary predicate error type

Llama2-13B Prediction	
Random	– *Top-n*: ['필헬레니즘', '싱크대', '태대형', '공공전', '융성'] (['Philhellenism', 'sink', 'Taedaehyung', 'Public War', 'Prosperity']) {{random definitions}} 친나라에 있는 것처럼 친그리스주의를 잡아두고 싶을 때, 싱크대를 이용하여 실천하는 것이 적절합니다. 태대형에 따라 십사 관등 가운데 둘째 등급에 해당되는 싱크대를 설치하여 필요한 재료를 다듬거나 씻거나 조리할 수 있습니다. 또한, 공공전을 통해 사람들이 함께 실천하고 융성을 일으킬 수도 있습니다. (When you want to maintain pro-Greekism as if you were in a pro-country, it is appropriate to practice using the sink. Depending on the type, a sink corresponding to the second rank among the fourteen can be installed to prepare, wash, or cook necessary ingredients. Additionally, through public war, people can practice together and create prosperity.)
Type Desc.	Error: * 제가 (I) should be 저가 (I) * 환자 (patient) should be 환자 (sick person) * 모르게 (secretly) should be 모르게 (secret) * 비밀로 (secretly) should be 비밀로 (secret) Corrected sentence: 저가 환자 모르게 비밀로 하고 싶었어요. (I wanted to keep the sick person a secret.)
KAGEC	– *Top-n*: ['환자', '모르게', '비밀로', '하고', '싶었어요'] (['patient', 'without knowing', 'secret', 'want to', 'do']) {{retrieved definitions}} Output: 그런데 제가 환자 치료를 받고 싶엇서요. (Corrected sentence using the keywords and their explanations) Explanation: * 환자 (hwan-ja) means "patient" in Korean. * 모르게 (mo-r-ge) means "secretly" or "privately" in Korean. * 비밀 (bi-mil) means "secret" or "private" in Korean. * 하고 (ha-go) is a filler word used to indicate a desire or intention. Therefore, the corrected sentence should be: "그런데 제가 환자 치료를 받고 싶엇서요" (I want to receive treatment as a patient privately).
Llama2-70B Prediction	
Random	– *Top-n*: ['회수처', '교건만상', '풀솜나물', '블록화', '흰주둥이바다지렁이'] (['collection' 'Gyogunmansang', 'Flossy greens', 'Blockhwa', 'White-snouted sea worm']) {{random definitions}} 회수처에서 환자를 모르게 비밀로 하고 싶은 분들에게 적합한 옵션으로 교건만상이 있어요. 이 서비스를 이용하면 환자의 치료와 동시에 거추장을 받을 수 있습니다. (Gyogunmansang is a suitable option for those who want to keep the patient secret from the collection location. Using this service, you can receive cumbersome at the same time as the patient.)
Type Desc.	그렇지 않아요. 그런 환자는 당신의 의견을 존중하지 않으므로 비밀로 하지 않으면 됩니다. Error definition: * spacing: 당신의 의견을 존중하지 않으므로 (two spaces between 존중 and 하지) * punctuation: 환자는 당신의 의견을 존중하지 않으므로 (no space between 환자 and 는) Corrected sentence: 그렇지 않아요. 그런 환자는 당신의 의견을 존중하지 않으므로 비밀로 하지 않으면 됩니다. Explanation: * Spacing error: The sentence contains two spaces between 존중 and 하지, which is not allowed in Korean. * Punctuation error: The sentence lacks a space between 환자 and 는, which is necessary in Korean. The corrected sentence follows the proper spacing and punctuation rules in Korean.
KAGEC	– *Top-n*: ['환자', '모르게', '비밀로', '하고', '싶었어요'] (['patient', 'without knowing', 'secret', 'want to', 'do']) {{retrieved definitions}} Correction: 그런데 제가 환자 모르게 비밀로 하고 싶어서요. Explanation: 싶어서요 is more natural way to express the past tense in informal speech.

Table 29: Llama2 prediction examples for auxiliary predicate error type. based on external information types. Source/Target/Baseline results are consistent with Table 28

Source Sentence	
전화가 연락하면 시간이 꽤 오래 걸릴 것 같은데요? (I think it might take quite a while if the phone try to contact them, don't you think?)	
Target Sentence	
전화로 연락하면 시간이 꽤 오래 걸릴 것 같은데요? (I think it might take quite a while if we try to contact them by phone, don't you think?)	
Llama2-13B Prediction	
Baseline	전화가 연락합니까? (Has the phone contacted?)
CoT	저는 실제로 전화가 연락을 하면 시간이 꽤 오래 걸릴 것 같은데요? (jeoneun jigeum-eun yeonhwa-ga yeon-nyeok-il go-mun-eun oh-sa-ham-ni-da?) Corrected sentence: Assistant: 저는 실제로 전화가 연락을 하면 시간이 꽤 오래 걸릴 것 같습니까? (jeoneun jigeum-eun yeonhwa-ga yeon-nyeok-il go-mun-eun oh-sa-ham-ni-da?) Explanation: 1. 저를 replacement for 제안 (je-an) to make the sentence more polite and formal. 2. 실제로 (jigeum-eun) is an adverb that means "actually" or "in reality". 3. 전화 (yeonhwa) is a noun that means "phone call". 4. 연락 (yeon-nyeok) is a noun that means "contact" or "reach". 5. 시간 (si-gian) is a noun that means "time". 6. 꽤 (go-mun) is an adverb that means "quite" or "rather". 7. 오래 (o-ra) is an adverb that means "long" or "for a long time". 8. 걸릴 (geol-rin) is a verb form that means "to take" or "to spend". So, the corrected sentence is more polite and formal, and it uses the correct adverbs and verb forms to convey the intended meaning.
Decomposed	– *Step 1* 해당 문장에는 grammatical error가 있습니다. "전화가 연락하면"는 올바른 방식으로 " phone calls (을/를) connect(s) when" 으로 수정해야 합니다. – *Step 2* 텔레파 Hijacks 연락이 되면 시간이 꽤 오래 걸릴 것 같습니다.
KAGEC	– *Top-n*: ['전화', '연락', '시간', '꽤', '오래'] (['phone', 'communication', 'time', 'quite', 'take']) {{retrieved definitions}} 전화로 연락하면 시간이 꽤 오래 걸릴 것 같은데요? (I think it might take quite a while if we try to contact them by phone, don't you think?)
Llama2-70B Prediction	
Baseline	전화가 연락하면 시간이 꽤 오래 걸릴 것 같습니다. (Corrected: "걸릴" should be "걸리는" in this sentence.) (I think it will take quite some time for phone to contact me.)
CoT	전화가 연락하면 시간이 꽤 오래 걸릴 것 같습니다. (Correct) Explanation: The sentence "전화가 연락하면 시간이 꽤 오래 걸릴 것 같은데요?" contains a grammatical error. The verb "걸릴" is in the wrong tense. It should be in the present tense to match the present tense of the sentence. Therefore, the corrected sentence should be "전화가 연락하면 시간이 꽤 오래 걸릴 것 같습니다." Do you have any questions?
Decomposed	– *Step 1* The sentence you provided contains a grammatical error. The verb "걸릴" is in the wrong tense. It should be in the present tense to match the present tense of the sentence. – *Step 2* 전화가 연락하면 시간이 꽤 오래 걸리는 것 같습니다. (When the phone contacts, it seems to take quite a while.)
KAGEC	– *Top-n*: ['전화', '연락', '시간', '걸릴', '걸리는'] (['phone call', 'communication', 'time', 'to take', 'to last']) {{retrieved definitions}} Answer: 전화로 연락하면 시간이 꽤 오래 걸릴 것 같은데요? Explanation: The sentence is grammatically correct, but the word "전화가" is not suitable for the context. Alternative answer: 연락하면 시간이 꽤 오래 걸릴 것 같은데요? Keywords: 전화 (phone call), 연락 (communication), 시간 (time), 걸릴 (to take), 걸리는 (to last)

Table 30: Llama2 prediction examples by prompting methodology for behavior error type.

Llama2-13B Prediction	
Random	– *Top-n*: ['거지', '교생', '환증', '광정', '셰이드'] (['the poor', 'teacher', 'disease', 'Gwangjeong', 'Shade']) {{random definitions}} 셰이드가 있습니까? (Do you have a shade?)
Type Desc.	전화가 연락하면 시간이 꽤 오래 걸릴 것 같은데요? (I think it might take quite a while if the phone try to contact them, don't you think?)
KAGEC	– *Top-n*: ['전화', '연락', '시간', '꽤', '오래'] (['phone', 'communication', 'time', 'quite', 'take']) {{retrieved definitions}} 전화로 연락하면 시간이 꽤 오래 걸릴 것 같은데요? (I think it might take quite a while if we try to contact them by phone, don't you think?)
Llama2-70B Prediction	
Random	– *Top-n*: ['드린제', '쌍림열반', '개장', '황화방', '오른편'] (['Drinje', 'Ssanglim Nirvana', 'opening', 'Hwanghwabang', 'right Side']) {{random definitions}} 드린제, 저는 알드린 없어요. 저는 엔드린 입니다. 저는 석가모니의 숲에서 죽음을 맞이하였습니다. 저는 과장에서 과거를 보기 시작할 때 황화방에 있었습니다. 저는 끈목, 담배, 쌈지, 바늘, 실 따위의 자질구레한 일용 잡화를 벌여 놓고 팔았습니다. 저는 오른편에 있었습니다. (Drinje, I don't have Aldrin. My name is Endrin. I met death in the forest of Buddha. I was in Hwanghwabang when I started looking at the past in the section. I sold small daily necessities such as braid, cigarettes, wrapping paper, needles, and thread. I was on the right.)
Type Desc.	전화가 연락하면 시간이 꽤 오래 걸릴 것 같습니다. (I think it will take quite some time for phone to contact me.)
KAGEC	– *Top-n*: ['전화', '연락', '시간', '걸릴', '걸리는'] (['phone call', 'communication', 'time', 'to take', 'to last']) {{retrieved definitions}} Answer: 전화로 연락하면 시간이 꽤 오래 걸릴 것 같은데요? Explanation: The sentence is grammatically correct, but the word "전화가" is not suitable for the context. Alternative answer: 연락하면 시간이 꽤 오래 걸릴 것 같은데요? Keywords: 전화 (phone call), 연락 (communication), 시간 (time), 걸릴 (to take), 걸리는 (to last)

Table 31: Llama2 prediction examples for behavior error type. based on external information types. Source/Target/Baseline results are consistent with Table 30