

KatFishNet: Detecting LLM-Generated Korean Text through Linguistic Feature Analysis

Shinwoo Park Shubin Kim Do-Kyung Kim Yo-Sub Han[★]
Yonsei University, Seoul, Republic of Korea

pshkhh@yonsei.ac.kr, shubs919@yonsei.ac.kr, kdky95@yonsei.ac.kr, emmous@yonsei.ac.kr

Abstract

The rapid advancement of large language models (LLMs) increases the difficulty of distinguishing between human-written and LLM-generated text. Detecting LLM-generated text is crucial for upholding academic integrity, preventing plagiarism, protecting copyrights, and ensuring ethical research practices. Most prior studies on detecting LLM-generated text focus primarily on English text. However, languages with distinct morphological and syntactic characteristics require specialized detection approaches. Their unique structures and usage patterns hinder the direct application of methods primarily designed for English. Among such languages, we focus on Korean, which has relatively flexible spacing rules, a rich morphological system, and less frequent comma usage compared to English. We introduce KatFish, the first benchmark dataset for detecting LLM-generated Korean text. The dataset consists of text written by humans and generated by four LLMs across three genres. By examining spacing patterns, part-of-speech diversity, and comma usage, we illuminate the linguistic differences between human-written and LLM-generated Korean text. Building on these observations, we propose KatFishNet, a detection method specifically designed for the Korean language. KatFishNet achieves an average of 19.78% higher AUC-ROC compared to the best-performing existing detection method. Our code and data are available at <https://github.com/Shinwoo-Park/katfishnet>.

1 Introduction

The rise of LLMs has led to significant advancements in various writing tasks (Brown et al., 2020; Gómez-Rodríguez and Williams, 2023; Xiao et al., 2024). However, their ability to generate coherent texts also raises concerns about potential misuse, such as spreading misinformation (Pan et al., 2023;

[★] Corresponding author.

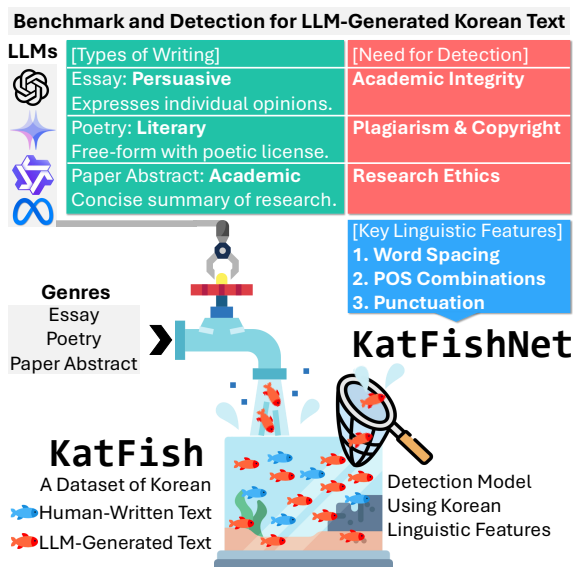


Figure 1: An illustration of our KatFish dataset (Sec. 2) and the detection method KatFishNet (Sec. 3).

Wang et al., 2024a) and facilitating academic dishonesty (Zellers et al., 2019; Perkins, 2023). Consequently, detecting LLM-generated text is paramount for safeguarding academic integrity, preventing plagiarism, protecting copyrights, and upholding research ethics (Guo et al., 2023; Su et al., 2023b; Wu et al., 2023; Orenstrakh et al., 2024).

Despite its significance, research on Korean text has been limited. The limitations in research on detecting LLM-generated Korean text include the lack of suitable benchmarks and the challenges arising from the unique linguistic characteristics of the Korean language (Park et al., 2020; Kim et al., 2022; Yoon et al., 2023; Choi et al., 2024; Kim et al., 2024). We present the first dataset for detecting LLM-generated Korean text, along with a detection method that utilizes Korean linguistic features. Figure 1 provides an illustration of our proposed dataset and detection approach.

We present KatFish (Korean LLM-generated text Benchmark For identifying Authorship), a

dataset developed specifically for detecting Korean text generated by LLMs. KatFish includes text samples from argumentative essays, poetry, and research paper abstracts, produced by both humans and four LLMs. Korean has distinct morphological and syntactic features, including flexible spacing rules and a rich system of postpositions and verb endings. These characteristics indicate that detection strategies designed for English may be less effective for Korean, highlighting the need for a language-specific approach. We explore the linguistic differences between human-written and LLM-generated Korean text by analyzing three key aspects: 1) word spacing; 2) part-of-speech combinations; and 3) punctuation. Our analysis uncovers distinct feature differences, revealing patterns that can be exploited for effective detection. Based on these findings, we propose KatFishNet, a machine learning-based detection method that incorporates the linguistic characteristics of the Korean language.

2 Dataset Construction: 🐟 KatFish

The KatFish consists of three types of text:

- **Essay:** Argumentative essays aim to persuade readers of a specific viewpoint or claim. They include a thesis statement, a logically organized structure with supporting evidence and counterarguments, and a concise conclusion.
- **Poetry:** Poetry is a creative form of writing that focuses on expressing emotions and artistic ideas. It often features metaphor, symbolism, and rhythm, breaking traditional linguistic norms to achieve distinctive artistic effects.
- **Paper Abstract:** Paper abstracts are concise summaries of academic research. They use precise language and technical terminology to clearly communicate the purpose, methodology, and key findings of a study.

We select these three genres to build a dataset that captures a diverse range of real-life scenarios and linguistic features while addressing practical challenges in detecting LLM-generated Korean text. Each genre highlights the importance of accurate detection in real-world applications. For a detailed rationale behind our genre selection, see Appendix A.

Linguistic and Structural Diversity Essays typically follow a logical and coherent structure, incorporating argumentation and supporting evidence.

Poetry stands out with its use of metaphor, rhythm, and stylistic innovation, often pushing the boundaries of traditional linguistic norms. In contrast, paper abstracts are compact and information-dense, marked by the frequent use of technical terms and discipline-specific language.

Practical Importance of Detection Detecting LLM-generated text in these genres is critical due to potential misuse. Essays generated by LLMs could facilitate academic dishonesty, eroding the value of original thinking. LLM-generated poetry raises concerns about plagiarism and the loss of authenticity in creative expression. Similarly, LLM-generated paper abstracts could threaten the integrity of academic research by introducing inaccuracies.

2.1 Human-Written Text Collection

We collect human-written Korean text from different sources depending on the type of writing.

Essay We collect writings from the essay corpus provided by AIHub¹, which includes argumentative essays written by elementary, middle, and high school students. The collected essays cover a total of 11 topics: 4 topics for elementary school students, 4 topics for middle school students, and 3 topics for high school students. There are no overlapping topics among the essays written by elementary, middle, and high school students. The statistics of the essays written by humans included in our final dataset are as follows: 1) Elementary school: 69 essays; 2) Middle school: 78 essays; and 3) High school: 34 essays. Descriptions of each essay topic are provided in Table 5.

Poetry We collect free verse poems from the poetry corpus provided by the National Institute of Korean Language². The collected poems are written by individuals under 10, those aged 10 to 19, those in their 20s and 30s, and those aged 40 and above. The statistics of the poems are as follows:

- Individuals under 10: 19 poems
- Individuals aged 10 to 19: 116 poems
- Individuals in their 20s and 30s: 44 poems
- Individuals aged 40 and above: 10 poems

¹<https://aihub.or.kr/>

²<https://kcorpus.korean.go.kr/index/goMain.do>

Paper Abstract We randomly select 100 papers related to language engineering from those published by the Korean Institute of Information Scientists and Engineers³ in 2016–2018.

2.2 LLM-Generated Text Collection

We generate LLM-generated Korean text using two commercial LLMs and two open-source LLMs. Specifically, we use the following LLMs: 1) **GPT-4o**: GPT-4o is a commercial LLM capable of understanding and processing all forms of input, including text, images, and speech. 2) **Solar**: Solar is a commercial LLM developed by Upstage, a Korean AI startup. 3) **Qwen2 72B**: Qwen2 is an open-source LLM developed by Alibaba, capable of understanding and processing around 30 languages, including Korean. 4) **Llama3.1 70B**: Llama3.1 is an open-source LLM developed by Meta, showing outstanding performance across various tasks. Table 6 shows the prompts used for text generation.

Essay When generating essays using LLMs, we design instructions based on education levels, essay topics, and prompts. The same essay prompts used by human writers serve as inputs for LLMs. LLMs receive instructions to write essays following the given topic and prompt while maintaining a writing style suitable for the specified education level. This approach helps minimize the influence of writing proficiency differences across education levels when distinguishing between human-written and LLM-generated essays.

Poetry When generating poems with LLMs, we provide the model with a human-written poem along with the age group of the poet and instruct it to create a new poem that matches the style and content suitable for that age group. The model takes the full human-written poem as input and generates a new poem based on it, mimicking a realistic scenario where a person may draw inspiration from existing works to produce something original. Additionally, the model composes poems tailored to a given age group, which helps reduce the impact of age-related differences in writing style when distinguishing between human-written and LLM-generated poems.

Paper Abstract A paper abstract summarizes the overall content of a study and highlights its key contributions. Therefore, we have the LLM read

³https://www.kiise.or.kr/academy/main/main_fa/

the entire paper excluding the abstract and generate a new abstract from the remaining content.

Data Cleaning We perform a manual analysis of the LLM-generated text and remove those that fall into the following three categories: 1) texts that do not follow the instructions and simply output the given prompt; 2) texts that repeatedly produce meaningless content (*e.g.*, AI assistant); 3) text generated in languages other than Korean.

	Essay	Poetry	Paper Abstract	Total
# Human	181	189	100	470
# GPT-4o	181	189	100	470
# Solar	140	189	100	429
# Qwen2	181	189	17	387
# Llama3.1	88	189	61	338
Total	771	945	378	2,094

Table 1: Data statistics of the KatFish dataset.

Dataset Statistic Table 1 presents the data statistics of the KatFish dataset. The KatFish includes 470 human-written Korean text and 1,624 LLM-generated Korean text. Each text undergoes a careful manual review to ensure it does not contain any sensitive personal information. We demonstrate that KatFish provides a sufficiently large benchmark for the task of distinguishing between human-written and LLM-generated Korean text. In comparison, recent studies by Mitchell et al. (2023) and Su et al. (2023a) conduct experiments on similar tasks involving human-written and LLM-generated English texts using 150 to 500 examples.

Table 7 presents the mean and standard deviation of Eojeol counts in texts by each author for each genre. An Eojeol is the smallest unit in a Korean sentence, separated by spaces.

3 Detection Method: 🐟KatFishNet

We compare and analyze the linguistic features of human-written Korean text and LLM-generated Korean text, and design KatFishNet based on these findings. Specifically, we focus on spacing patterns, part-of-speech n-gram diversity, and comma usage patterns. These are closely related to writing habits, grammatical structures, and textual coherence.

3.1 Word Spacing Patterns

Unlike English, Korean has many exceptions and flexibilities in its spacing rules, making it one of the most variable and challenging aspects of writing. This makes it a valuable feature for examining stylistic and grammatical differences. In line with

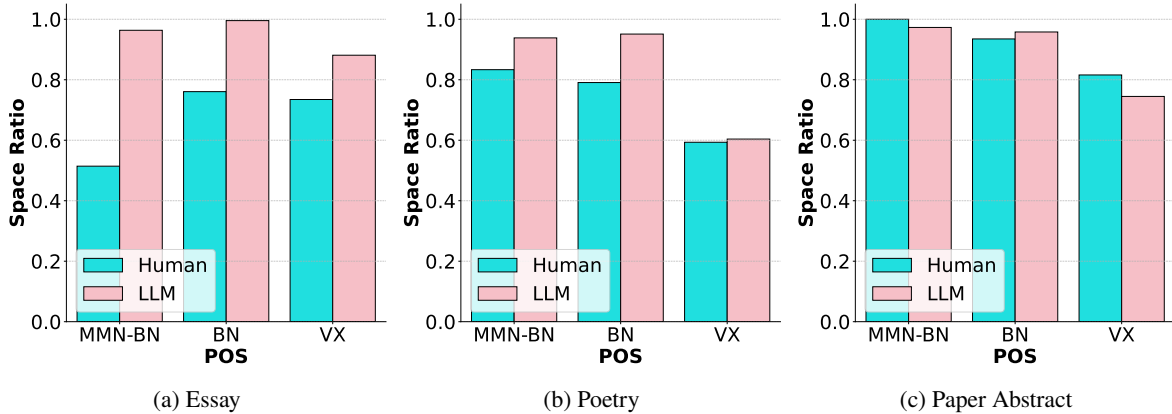


Figure 2: Comparison of space occurrence ratios between MMN+BN, prior to BN, and prior to VX.

the Korean orthography word spacing guidelines⁴, we examine bound nouns (BN) and auxiliary predicate elements (VX), which are the key part-of-speech (POS) categories related to word spacing. We also investigate **Eojeol POS Diversity** and **Unspaced VX Diversity** in Appendix E. We use Bareun POS tagger.

Bound nouns must be used in conjunction with another word but require word spacing as they function as nouns. Their dependent nature leads to frequent mistakes, accounting for 70% of total word spacing errors (Shin et al., 2015). We examine two metrics: **Numeral Determiner–Bound Noun (MMN-BN) Space Ratio** and **Bound Noun (BN) Space Ratio**. The MMN-BN Space Ratio measures the frequency of word spacing between a numeral determiner (MMN) and a bound noun (BN). BN Space Ratio quantifies the frequency of word spacing before a BN. To focus on cases with greater variation, we eliminate trivial cases where spaces may be omitted.

Auxiliary predicate elements attach to the main predicate to complement its meaning. In principle, VXs should have a space preceding them, but the guidelines allow flexibility in exceptional cases. When used correctly, omitting the space may enhance readability. **VX Space Ratio** measures the frequency of word spacing before a VX, excluding the specific case of "-아"/"-어" (ENDING) + "지" (VX), where spacing is strictly prohibited. This metric indicates how strictly the author adheres to the principle while also considering flexibility in applying exceptions.

Figure 2 reveals that in essays and poetry, human-

written text exhibits a lower space ratio across all metrics. Notably, LLM-generated essays display a highly consistent BN Space Ratio, with a standard deviation of 0.02. While LLMs rigidly enforce spacing rules, humans often omit spaces, influenced by various stylistic and grammatical factors. These factors include prioritizing readability and convenience over adherence to principles, poetic license, and a lack of understanding of spacing rules.

The differences are the most evident in essays and the least pronounced in paper abstracts. This illustrates that spacing behavior is influenced by context-dependent stylistic tendencies. In domains where humans adhere to highly structured formats and conventions, word spacing patterns may be less significant. However, they remain useful in domains with a wider range of authors and writing styles.

Finding. LLMs strictly follow spacing rules, while human writers omit spaces due to stylistic and grammatical factors.

3.2 Part-of-Speech N-gram Diversity

We analyze POS n-gram diversity to examine structural differences between human-written and LLM-generated Korean text. Using the Kkma POS tagger (Park and Cho, 2014), we extract POS sequences from each text and compute the **POS N-gram Diversity Score** by dividing the number of unique POS n-grams by the total number of POS n-grams in the text. After calculating the average diversity score for all human-written and LLM-generated text, we compare the results to identify differences. We consider n-grams ranging from unigrams (1-gram) to pentagrams (5-gram) to capture linguistic patterns at different levels. Unigrams reflect basic lexical choices, while higher-order n-grams capture more complex syntactic structures and dependen-

⁴https://korean.go.kr/kornorms/regltn/regltnView.do?regltn_code=0001®ltn_no=182#a182

cies. By analyzing diversity across these varying n-gram lengths, we aim to gain a comprehensive understanding of how humans and LLMs construct text differently.

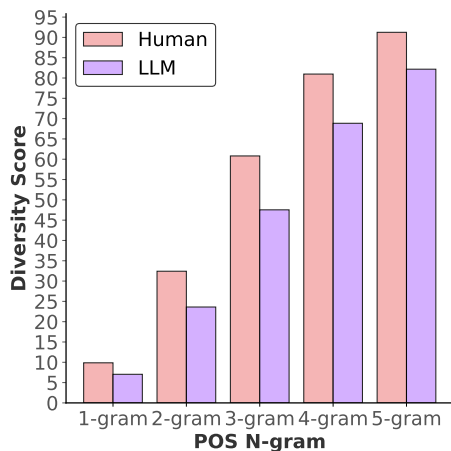


Figure 3: Comparison of POS n-gram diversity between human-written and LLM-generated Korean essays.

Figure 3 compares the average POS n-gram diversity scores between human-written and LLM-generated essays. We include the analysis results for poetry and paper abstracts in the Appendix F due to space limitations. Human-written essays exhibit a higher diversity score than LLM-generated essays. The analysis results highlight that humans use a wider range of grammatical structures and construct sentences more flexibly than LLMs. Since LLMs generate text by selecting the most probable word combinations based on training data, they tend to repeat commonly used structures more frequently.

Finding. Humans tend to use a more diverse range of POS combinations in their writing compared to LLMs.

3.3 Comma Usage Patterns

In Korean, commas help improve readability and clarify meaning within sentences. We analyze differences in comma usage to compare how humans and LLMs structure sentences and manage their flow. Specifically, we investigate the proportion of commas within sentences, their placement, structural changes in sentences, and linguistic diversity around commas. We compute the following five metrics: 1) **Comma Inclusion Rate:** The proportion of sentences containing at least one comma out of all sentences in a text. 2) **Average Comma Usage Rate:** The number of commas in a sentence divided by the total number of morphemes in that

sentence. 3) **Average Relative Position of Comma:** The position of each comma (counting the number of morphemes before it) divided by the total number of morphemes in the sentence. If multiple commas appear in a sentence, the average relative position is calculated. 4) **Average Segment Length:** The average length of sentence segments split by commas. 5) **POS Diversity Score Before and After Comma:** The diversity of part-of-speech pairs appearing before and after a comma. This score is calculated by dividing the number of unique POS pairs by the total number of POS pairs. We segment each text into individual sentences, compute these metrics for each sentence, and use the average values of these sentence-level metrics as the representative values for the entire text. Appendix G further explores the specific POS patterns surrounding commas.

Genre	Metric	Human	LLM
Essay	Inclusion Rate (%)	26.31	61.03
	Usage Rate (%)	1.13	2.56
	Avg. Relative Position	0.09	0.18
	Avg. Segment Length	4.35	8.56
	POS Diversity Score	24.38	59.39
Poetry	Inclusion Rate (%)	27.01	42.90
	Usage Rate (%)	2.61	4.84
	Avg. Relative Position	0.14	0.28
	Avg. Segment Length	1.96	2.13
	POS Diversity Score	23.13	23.86
Paper	Inclusion Rate (%)	47.48	65.21
	Usage Rate (%)	1.73	2.40
	Avg. Relative Position	0.20	0.25
Abstract	Avg. Segment Length	9.07	11.55
	POS Diversity Score	42.85	61.95

Table 2: Comparison of comma usage patterns between human-written and LLM-generated Korean text.

Table 2 presents the analysis results. 1) LLMs include commas in more sentences and uses them more frequently: LLM-generated text contains a higher proportion of sentences with commas compared to human-written text. The frequency of comma usage within sentences is also higher in LLM-generated text, leading to more frequent segmentation. 2) LLMs tend to place commas later in a sentence than humans: While both humans and LLMs often place commas near the beginning of a sentence, analysis shows that LLMs tend to insert them slightly later. This can be attributed to the fact that LLMs are trained on large multilingual datasets, particularly those with substantial amounts of English, where comma usage patterns can differ from those commonly found in Korean. Additionally, in LLM-generated text, the segments of sentences separated by commas tend to be relatively longer compared to human-written text. 3)

LLM-generated text shows greater diversity in part-of-speech combinations around commas: A higher diversity of POS combinations before and after commas indicates that LLMs apply a wider range of grammatical patterns when constructing sentences. Poetry shows almost no difference between human and LLM-generated text. This similarity likely arises because poetry naturally consists of shorter and simpler sentence structures.

Finding. LLMs use commas more often, place them later, and show greater POS diversity than humans.

3.4 Design of KatFishNet

Our detection method, KatFishNet, is a machine learning-based model that leverages Korean linguistic features. Table 8 in Appendix H presents the Korean linguistic features used by KatFishNet. KatFishNet offers several advantages: First, it enables analysis of which features play a crucial role in detecting LLM-generated text, making the model highly interpretable. Second, compared to deep learning-based models such as Transformer-based approaches, it does not require training text embeddings or large-scale datasets, allowing for a lightweight and efficient detection system. Lastly, traditional machine learning models like logistic regression, random forests, and support vector machines provide a practical and efficient solution by enabling training on CPUs, avoiding the costs associated with GPU resources.

We construct input feature vectors based on quantitative values obtained from word spacing, POS combinations, or punctuation analysis, and perform machine learning on these vectors. In Section 5, we investigate which types of quantitative metrics are most effective in detecting LLM-generated Korean text by comparing the performance of models trained with different feature sets.

4 Experimental Settings

4.1 Task Definition

We address a binary classification task where, given a Korean text, the goal is to classify whether the text is generated by a human or an LLM. We evaluate performance using the Area Under the Receiver Operating Characteristic curve (AUC-ROC). AUC-ROC measures the area under the curve that plots the True Positive Rate against the False Positive Rate at different threshold levels. This metric measures

overall detection performance across all potential classification thresholds.

4.2 Baselines

We select the following five types of baselines.

Confidence-based Methods Confidence-based methods leverage a pre-trained language model such as GPT-2 to analyze text and extract distinctive features, such as the rank or entropy of each word based on its preceding context. We employ the following methods: 1) Log-Likelihood (Solaiman et al., 2019): A language model calculates the log probability of each word in a given text, and the average of these values serves as the score. Higher average log probabilities indicate a greater likelihood that the text was generated by an LLM. 2) Entropy (Gehrmann et al., 2019): The entropy of each word is measured based on its preceding context, and the average entropy across the text is used as the score. LLM-generated text typically have lower entropy values. 3) Log-Rank (Mitchell et al., 2023): The absolute rank of each word is determined based on the preceding context, and the text-level score is obtained by averaging the logarithm of these values. 4) LRR (Log-Likelihood Log-Rank Ratio) (Su et al., 2023a): The LRR is computed by dividing log-likelihood values by log-rank values. Generally, LLM-generated text tend to have higher LRR values than human-written text.

Perturbation-based Methods Perturbation-based methods evaluate changes in the log probability of a model when making slight modifications to the original text. These methods use a pre-trained language model such as T5 to generate multiple perturbed versions of the text. By calculating log probabilities for both the original and perturbed text, they determine whether a text is machine-generated. We use the following methods: 1) DetectGPT (Mitchell et al., 2023): DetectGPT measures how the log probability of a model changes when making small modifications to the original text. This method is based on the idea that text generated by an LLM typically reaches a local optimum of the model log probability function. As a result, minor alterations to LLM-generated text tend to lower the log probability compared to the original version. 2) NPR (Normalized Log-Rank Perturbation) (Su et al., 2023a): NPR examines how the log-rank score responds to slight perturbations. When small modifications are applied, the log-rank

score of LLM-generated text increases more than that of human-written text.

LLM Paraphrasing This method determines whether a given text is human-written or LLM-generated by paraphrasing the original text using an LLM and measuring the similarity between the paraphrased and original versions. The method operates on the intuition that LLM-generated text undergo fewer changes because they align more closely with the generation patterns of LLMs (Zhu et al., 2023). In other words, if the similarity between the original and paraphrased text is high, the original text is considered to be LLM-generated. We measure this similarity using BARTScore (Yuan et al., 2021). We use Exaone 3.5 (Research et al., 2024) 32B, an LLM released by the Korean company LG AI Research, to perform paraphrasing.

LLM Prompting We provide a Korean text to an LLM and ask it to output 1 if the text is LLM-generated and 0 if it is human-written. For this baseline, we use Exaone 3.5.

Fine-tuning We fine-tune the encoder of a pre-trained language model using the KatFish dataset. Specifically, we build on a RoBERTa-base model initially trained on the HC3 dataset (Guo et al., 2023), which consists of English and Chinese text written by both humans and ChatGPT. By further training this model with the KatFish dataset, we enhance its ability to distinguish between human-written and LLM-generated Korean text.

The implementation details of KatFishNet and the baselines are provided in the Appendix I.

4.3 OOD Evaluation: Unseen LLMs

We perform out-of-distribution (OOD) evaluation to assess how well the detection methods generalize. Specifically, we test whether the model can accurately distinguish between human-written and LLM-generated Korean text even when faced with texts from an unseen LLM. This evaluation design is essential given the frequent emergence of LLMs with distinct text generation patterns. If a detection system relies only on data from familiar LLMs, it may struggle to maintain performance when confronted with a previously unseen model. By testing on LLMs not encountered during training, we can better approximate real-world conditions and gain deeper insights into how the detection system adapts without retraining for every new LLM.

We split the human-written text into an 8:2 ratio, using 80% of it along with text generated by GPT-4o—one of the most representative LLMs—to create the training dataset. For evaluation, we use text generated by Solar, Qwen2, and Llama3.1. Specifically, we construct three separate test sets by combining the text from each of these LLMs with the remaining 20% of human-written text. The detection methods requiring training include our proposed approach along with the fine-tuning baseline. These methods undergo training on the training dataset and are evaluated using the test sets. Despite the training process, the models are always tested on text generated by unseen LLMs. This ensures that all detection methods operate in a zero-shot classification setting.

5 Experimental Results

Table 3 presents the experimental results. We analyze the results from two perspectives: 1) Which baseline method performs best? 2) Which type of linguistic features contributes most to performance?

Best Performing Baseline Method The experimental results show that among the baseline methods, LLM paraphrasing achieves the highest performance for essays and abstracts, while DetectGPT performs best for poetry. In terms of average performance across the three text genres, LLM paraphrasing outperforms the other baselines. This may be because LLM paraphrasing directly exploits the characteristics of LLM-generated text. The results provide experimental support for the hypothesis that when an LLM modifies text, it introduces fewer changes to LLM-generated text than to human-written text.

Most Effective Linguistic Features We use logistic regression as the backbone model for KatFishNet and provide additional experimental results with random forest and support vector machine models in Appendix J. The results show that KatFishNet achieves the highest performance when leveraging comma usage patterns, compared to spacing patterns and POS n-gram diversity. KatFishNet with comma usage patterns outperforms all other methods across all three text genres. It achieves a 16.74% performance improvement over LLM paraphrasing for essays, a 10.72% improvement over DetectGPT for poetry, and a 31.90% improvement over LLM paraphrasing for paper abstracts. Meanwhile, KatFishNet with POS

Genre	Detection Methods		→ Solar	→ Qwen2	→ Llama3.1	Average
Essay	Confidence	Log-Likelihood	83.84	23.89	66.20	57.97
		Entropy	31.25	84.53	44.12	53.30
		Log-Rank	78.84	20.66	61.92	53.80
		LRR	45.08	80.56	53.15	59.59
	Perturbation	DetectGPT	52.78	37.45	47.18	45.80
		NPR	55.22	19.90	44.71	39.94
	LLM Paraphrasing	Exaone 3.5	92.08	79.74	72.00	81.27
	LLM Prompting	Exaone 3.5	50.42	49.74	50.07	50.07
	Fine-tuning	RoBERTa	66.77	66.65	64.37	65.93
	KatFishNet (Ours)	Word Spacing	86.00	80.63	71.91	79.51
POS Combinations		92.26	83.10	73.63	82.99	
Punctuation		97.57	94.63	92.45	94.88	
Poetry	Confidence	Log-Likelihood	77.06	47.34	59.99	61.46
		Entropy	30.90	68.28	47.68	48.95
		Log-Rank	75.76	45.67	60.54	60.65
		LRR	34.40	55.86	39.79	43.35
	Perturbation	DetectGPT	67.04	64.00	67.02	66.02
		NPR	63.75	41.21	62.92	55.96
	LLM Paraphrasing	Exaone 3.5	71.32	58.79	61.51	63.87
	LLM Prompting	Exaone 3.5	50.53	50.16	49.42	50.03
	Fine-tuning	RoBERTa	60.35	69.61	55.96	61.97
	KatFishNet (Ours)	Word Spacing	71.85	65.56	43.81	60.40
POS Combinations		39.41	79.17	53.32	57.30	
Punctuation		62.65	93.45	63.22	73.10	
Paper Abstract	Confidence	Log-Likelihood	58.52	42.41	47.86	49.59
		Entropy	36.13	72.64	51.85	53.54
		Log-Rank	57.08	45.05	47.57	49.90
		LRR	49.39	47.82	54.80	50.67
	Perturbation	DetectGPT	55.81	51.70	51.11	52.87
		NPR	63.14	46.76	60.98	56.96
	LLM Paraphrasing	Exaone 3.5	70.80	36.47	64.72	57.33
	LLM Prompting	Exaone 3.5	48.60	46.41	47.18	47.39
	Fine-tuning	RoBERTa	50.70	49.73	50.02	50.15
	KatFishNet (Ours)	Word Spacing	57.73	66.91	49.36	58.00
POS Combinations		47.47	70.05	42.47	53.33	
Punctuation		78.99	77.47	70.41	75.62	

Table 3: Performance of detecting LLM-generated Korean text. We report the average performance (AUC-ROC) over five experiments. We separately report the performance of the detection model for the task of distinguishing between human-written text and text generated by a specific LLM. For example, → Solar indicates that the detection model is evaluated on the task of classifying human-written text and text generated by Solar.

n-gram diversity as features ranks second among all methods for essay, while KatFishNet with spacing patterns as features ranks second for abstract. We hypothesize that comma usage patterns are more useful for detecting LLM-generated Korean text than spacing patterns or POS n-gram diversity, as LLMs tend to have more difficulty learning comma usage than word spacing or POS combinations. Spacing follows relatively clear patterns in training data, and POS sequences can be learned as probabilistic patterns. In contrast, comma usage reflects contextual and stylistic factors, making it highly variable depending on the intent of the writer.

We analyze the detection performance of KatFishNet in Appendix K by combining three categories of linguistic features using an ensemble approach.

Finding. Comma usage patterns serve as a key feature for distinguishing between human-written and LLM-generated Korean text.

6 Related Work

Zellers et al. (2019) developed the GROVER dataset, which includes human-written and AI-generated news articles to support research on detecting machine-generated disinformation. Similarly, Fagni et al. (2021) created TweepFake, a dataset of tweets authored by both humans and bots, facilitating studies on social media content authenticity. Guo et al. (2023) introduced the HC3 dataset, which contains questions and answers generated by both human experts and ChatGPT in English and Chinese. Wang et al. (2024c) advanced this direction by introducing the M4 dataset, a multi-generator, multi-domain, and multilingual corpus.

Wang et al. (2024b) introduced M4GT-Bench, a multilingual, multi-domain corpus covering various LLMs and proposes more nuanced tasks, such as pinpointing the specific LLM behind the generated text (multi-way detection) and identifying human-machine boundary points in mixed-content texts. Similarly, Macko et al. (2023) developed MULTITuDE, a large-scale multilingual dataset spanning eleven languages, to examine the cross-lingual and cross-generator generalization capabilities of detectors, comparing multilingual detectors to monolingual detectors. By systematically varying the training and evaluation languages, MULTITuDE demonstrates that English-focused models often do not seamlessly transfer to other languages without explicit multilingual training.

Although M4GT-Bench and MULTITuDE both address multilingual machine-generated text detection, neither benchmark focuses on language-specific approaches. In M4GT-Bench, the monolingual setup is trained exclusively on English data, while the multilingual setup uses a mixture of languages excluding the target language. As a result, there is no training that captures the morphological or syntactic characteristics of any single language in particular. MULTITuDE trains a separate model for each language and applies the same detection method across all of them. Its primary goal is to explore language transfer—that is, to examine how a model trained on certain languages generalizes to others—rather than to enhance performance for individual languages.

In contrast, our work highlights Korean-specific features—such as flexible spacing rules, part-of-speech n-gram patterns, and punctuation usage—in the design of the detection method. By focusing on language-specific characteristics, we demonstrate how linguistically tailored strategies can enhance the performance of LLM-generated text detection.

7 Conclusion

We address the challenge of distinguishing between human-written and LLM-generated Korean text by introducing KatFish, the first benchmark dataset for this task. Building on this foundation, we propose KatFishNet, a detection method that leverages linguistic features of the Korean language, including word spacing, POS combinations, and punctuation. Experimental results show that KatFishNet, particularly its use of comma usage patterns, sets a new state-of-the-art in detection performance. Notably,

the strong performance of KatFishNet demonstrates the effectiveness of designing language-specific detection methods. Our research demonstrates the potential of designing detection methods based on linguistic features, providing a foundation for developing similar approaches for other languages in the future.

Limitations

This study has several limitations that should be acknowledged. First, the scope of the KatFish benchmark is limited to three specific genres: essays, poetry, and paper abstracts. While these genres provide a useful foundation, they do not encompass all text types where LLM-generated content could present risks, such as news articles, social media posts, and legal documents. Expanding the dataset to include a more diverse range of text types would enhance the generalizability of the findings.

Second, this study focuses on distinguishing between fully human-written and fully LLM-generated text. However, real-world scenarios often involve hybrid content, where human- and LLM-generated text are interwoven. Future research should investigate detection methods capable of effectively handling such mixed cases.

Finally, enhancing the performance of detection methods that rely on linguistic features requires advancements in Korean morphological analysis. Current morphological analyzers still face challenges, which can impact the accuracy and reliability of extracted features. Further improvements to these tools could improve linguistic feature extraction and detection performance.

Ethical Considerations

We ensure that the data collection process for KatFish respects privacy and intellectual property rights by using publicly available texts and generating AI content within ethical guidelines.

Acknowledgments

This research was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (RS-2025-02222626 and RS-2025-00562134), and by the AI Graduate School Program (RS-2020-II201361).

References

- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*.
- ChangSu Choi, Yongbin Jeong, Seoyoon Park, Inho Won, HyeonSeok Lim, SangMin Kim, Yejee Kang, Chanhyuk Yoon, Jaewan Park, Yiseul Lee, HyeJin Lee, Younggyun Hahm, Hansaem Kim, and Kyung-Tae Lim. 2024. Optimizing language augmentation for multilingual large language models: A case study on Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.
- Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweep-Fake: About Detecting Deepfake Tweets. *Plos one*.
- Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv:2301.07597*.
- Yuxuan Guo, Zhiliang Tian, Yiping Song, Tianlun Liu, Liang Ding, and Dongsheng Li. 2024. Context-aware watermark with semantic balanced green-red lists for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. MGTBench: Benchmarking Machine-Generated Text Detection. *arXiv preprint arXiv:2303.14822*.
- HS Heaps. 1978. Information Retrieval: Computational and Theoretical Aspects.
- Hyeondey Kim, Seonhoon Kim, Inho Kang, Nojun Kwak, and Pascale Fung. 2022. Korean language modeling via syntactic guide. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*.
- Jong Myoung Kim, Young-Jun Lee, Yong-Jin Han, Ho-Jin Choi, and Sangkeun Jung. 2024. Does incomplete syntax influence korean language model? focusing on word order and case markers. In *First Conference on Language Modeling*.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *International Conference on Machine Learning (ICML)*.
- Dominik Macko, Robert Moro, Adaku Uchendu, Jason Lucas, Michiharu Yamashita, Matúš Pikuliak, Ivan Srba, Thai Le, Dongwon Lee, Jakub Simko, and Maria Bielikova. 2023. MULTITuDE: Large-scale multilingual machine-generated text detection benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- Niloofer Mireshghallah, Justus Mattern, Sicun Gao, Reza Shokri, and Taylor Berg-Kirkpatrick. 2024. Smaller Language Models are Better Zero-shot Machine-Generated Text Detectors. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*.
- Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. DetectGPT: Zero-shot Machine-Generated Text Detection using Probability Curvature. In *International Conference on Machine Learning*.
- Michael Sheinman Orenstrakh, Oscar Karnalim, Carlos Anibal Suarez, and Michael Liut. 2024. Detecting LLM-Generated Text in Computing Education: Comparative Study for ChatGPT Cases. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE.
- Yikang Pan, Liangming Pan, Wenhui Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Eunjeong L. Park and Sungzoon Cho. 2014. KoNLPy: Korean Natural Language Processing in Python. In *Proceedings of the 26th Annual Conference on Human and Cognitive Language Technology, Chuncheon, Korea*.
- Kyubyong Park, Joohong Lee, Seongbo Jang, and Da-woon Jung. 2020. An Empirical Study of Tokenization Strategies for Various Korean NLP Tasks. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*.
- Mike Perkins. 2023. Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*.

- LG Research, Soyoung An, Kyunghoon Bae, Eunbi Choi, Kibong Choi, Stanley Jungkyu Choi, Seokhee Hong, Junwon Hwang, Hyojin Jeon, Gerrard Jeongwon Jo, et al. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.
- Hocheol Shin, Buyeon Kim, and Kyubum Lee. 2015. A study on the hangeul orthography error status. *Grammar Education*, 23:63–94.
- Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askill, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release Strategies and the Social Impacts of Language Models. *arXiv preprint arXiv:1908.09203*.
- Jinyan Su, Terry Zhuo, Di Wang, and Preslav Nakov. 2023a. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.
- Zhenpeng Su, Xing Wu, Wei Zhou, Guangyuan Ma, and Songlin Hu. 2023b. HC3 Plus: A Semantic-Invariant Human ChatGPT Comparison Corpus. *arXiv preprint arXiv:2309.02731*.
- Lionel Z Wang, Yiming Ma, Renfei Gao, Beichen Guo, Zhuoran Li, Han Zhu, Wenqi Fan, Zexin Lu, and Ka Chung Ng. 2024a. MegaFake: A Theory-Driven Dataset of Fake News Generated by Large Language Models. *arXiv preprint arXiv:2408.11871*.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024b. M4GT-bench: Evaluation benchmark for black-box machine-generated text detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024c. M4: Multi-generator, Multi-domain, and Multilingual Black-Box Machine-Generated Text Detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*.
- Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F Wong, and Lidia S Chao. 2023. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. *arXiv preprint arXiv:2310.14724*.
- Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2024. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. *arXiv preprint arXiv:2401.06431*.
- Soyoung Yoon, Sungjoon Park, Gyuwan Kim, Junhee Cho, Kihyo Park, Gyu Tae Kim, Minjoon Seo, and Alice Oh. 2023. Towards Standardizing Korean Grammatical Error Correction: Datasets and Annotation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems*.
- Biru Zhu, Lifan Yuan, Ganqu Cui, Yangyi Chen, Chong Fu, Bingxiang He, Yangdong Deng, Zhiyuan Liu, Maosong Sun, and Ming Gu. 2023. Beat LLMs at Their Own Game: Zero-Shot LLM-Generated Text Detection via Querying ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.
- George Kingsley Zipf. 2016. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Ravenio books.

A Genre Selection Criteria and Rationale

Practicality-Driven Domain Selection: Targeting Directly Problematic Use Cases Our selection of text types is driven by a focus on practical use cases where the use of LLMs for text generation is directly problematic. This contrasts with other genres where LLM use is either legitimate or not explicitly prohibited, but may still lead to indirect issues. We prioritize three genres that directly relate to violations of academic integrity, copyright, and research ethics—major issues that are of immediate concern in real-world applications of LLM-generated text.

Ensuring Broad Coverage in Writing Types Texts are classified into different types by their purpose, writing style, and usage. To ensure the broad applicability of our findings, we meticulously choose three genres—student-written argumentative essays, free-verse poetry, and research paper abstracts—to represent major facets of writing. Table 4 shows how each genre in KatFish contributes to this coverage.

Category	Type	Genres
Purpose	Expressive Writing	Poetry
	Informative Writing	Paper Abstract
	Persuasive Writing	Essay
Writing Style	Descriptive	Poetry
	Narrative	Poetry, Essay
	Expository	Paper Abstract
Usage	Persuasive	Essay
	Literary Writing	Poetry
	Academic Writing	Essay, Paper Abstract

Table 4: Summary of genre characteristics in KatFish, illustrating their complementary roles across purpose, writing style, and usage.

We choose free-verse poetry, a less conventional representative for literary text, to investigate how less conventional writing affects LLM-generated text detection. This allows us to gain insights that cannot be obtained from more structured forms.

B Essay Topics for Argumentative Writing Across Education Levels

Table 5 shows the essay topics used for writing argumentative essays at each education level.

Education Levels	Essay Topic
Elementary	다문화 가족을 대하는 본인의 자세 (Your Attitude Towards Multicultural Families)
	폭력 예방 방법 (Ways to Prevent Violence)
	과학의 발전에 대한 본인의 생각 (Your Thoughts on the Advancement of Science)
	미디어 발전과 사용방법 (Advancement of Media and its Usage Methods)
Middle	SNS상의 문제에 대한 본인의 생각 (Your Thoughts on the Issues Related to Social Media)
	e스포츠에 대한 본인의 생각 (Your Thoughts on Esports)
	전통과 악습에 대한 본인의 생각 (Your Thoughts on Tradition and Bad Practices)
	생물학적으로 다른 남/여에 대한 본인의 생각 (Your Thoughts on Biologically Different Males and Females)
High	인종차별에 대한 본인의 생각 (Your Thoughts on Racism)
	지적 재산권에 대한 본인의 생각 (Your Thoughts on Intellectual Property)
	평가에 대한 본인의 생각 (Your Thoughts on the Evaluation)

Table 5: Essay topics used for argumentative writing at different education levels.

C Prompt for Korean Text Generation via LLMs

Korean	Prompt Template
Essay	너는 과제로 주장글을 작성해야 하는 [EDUCATION LEVEL] 학생이야. 주어진 주제와 질문에 맞춰 에세이를 작성해줘. 한국어로만 작성해줘. 에세이만 출력해. 주제: [TOPIC] 질문: [ESSAY PROMPT]
Poetry	너는 [AGE GROUP] 시인이야. 주어진 시를 읽고 내용을 파악해줘. 그 후 너의 스타일로 너의 나이에 맞는 새로운 시를 작성해줘. 한국어로만 작성해줘. 새로운 시만 출력해. 주어진 시: [POEM]
Paper Abstract	초록을 제외한 논문을 줄테니, 이 논문의 초록을 작성해줘. 한국어로만 작성해줘. [PAPER]

English	Prompt Template
Essay	You are a [EDUCATION LEVEL] student who has to write an argumentative essay as an assignment. Write an essay according to the given topic and question. Write only in Korean. Output the essay only. Topic: [TOPIC] Question: [ESSAY PROMPT]
Poetry	You are a poet in your [AGE GROUP]. Read the given poem and understand its content. Then, write a new poem in your style that suits your age group. Write only in Korean. Output the new poem only. Given poem: [POEM]
Paper Abstract	I'll give you a paper without the abstract. Write an abstract for this paper. Write only in Korean. [PAPER]

Table 6: Prompt templates used for building the KatFish benchmark. The upper table shows the original Korean prompt templates used for data generation, and the lower table displays the translated English prompt templates.

D Data Statistics: Eojeol Counts

Table 7 presents the mean and standard deviation of the number of Eojeols in text written by each author for each text type. An Eojeol is the smallest unit in a Korean sentence, separated by spaces, and may consist of a single morpheme or a combination of multiple morphemes.

	Essay	Poetry	Paper Abstract
Human	152 \pm 54	52 \pm 52	91 \pm 27
GPT-4o	268 \pm 38	54 \pm 24	120 \pm 26
Solar	175 \pm 56	40 \pm 19	94 \pm 55
Qwen2	179 \pm 44	62 \pm 27	88 \pm 38
Llama3.1	211 \pm 28	77 \pm 30	133 \pm 33

Table 7: The average and standard deviation of Eojeol counts for each text type by author.

E Word Spacing Analysis

Alongside the three metrics related to word spacing in Section 3.1, we also analyze **Eojeol POS Diversity** and **Unspaced VX Diversity**. These metrics offer stylistic and grammatical distinctions in Korean text generation. Eojeol POS Diversity is computed by dividing the number of unique POS sequences by the total number of Eojeols in the text, where each sequence is defined at the Eojeol level. Eojeol POS diversity captures how diverse the syntactic structures are within individual Eojeols, reflecting differences in linguistic complexity and variability between human-written and LLM-generated text. Unspaced VX Diversity is computed by dividing the number of unique unspaced auxiliary verb stems by the total number of unspaced auxiliary verbs. We exclude the case where spacing is explicitly allowed. This analysis provides insights into the tendency of humans and LLMs to make consistent or varied spacing choices in VX-related word spacing.

Figure 4 shows that Eojeol VX Diversity is consistently higher for human-written text across all three genres, similar to the POS n-gram diversity results. Figure 5 illustrates that Unspaced VX Diversity results vary by genre, with humans scoring lower for essays and LLMs scoring lower for paper abstracts. This shows that genre-specific stylistic tendencies influence the spacing behavior of both humans and LLMs, reflecting variations in writing conventions and levels of formality.

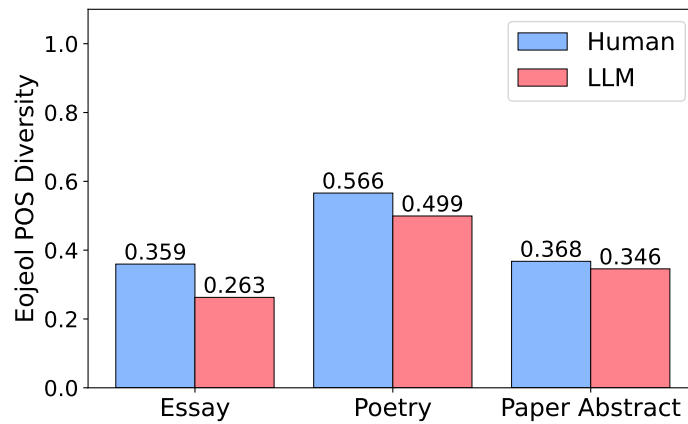


Figure 4: Comparison of Eojeol POS Diversity between human-written and LLM-generated Korean text.

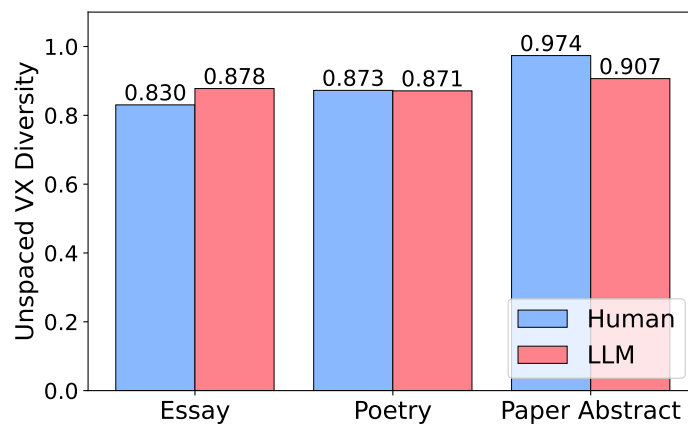


Figure 5: Comparison of Unspaced VX Diversity between human-written and LLM-generated Korean text.

F Part-of-Speech N-gram Diversity

Figure 6 compares the average POS n-gram diversity scores between human-written and LLM-generated Korean text. Excluding POS 4-gram and 5-gram in poetry, human-written text shows a higher diversity score than LLM-generated text. Poetry has shorter length and simpler structure compared to essays and paper abstracts, which can reduce the difference in POS n-gram diversity between human-written and LLM-generated text for 4-gram and 5-gram sequences.

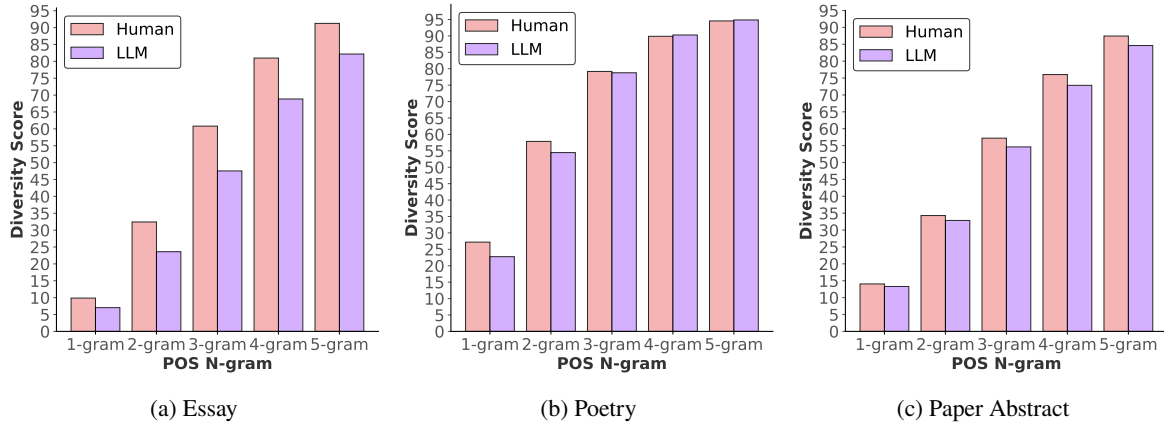


Figure 6: Comparison of POS n-gram diversity between human-written and LLM-generated Korean text.

G POS-Comma Analysis

Exploring when humans and LLMs use commas can give valuable information on stylistic differences between human-written and LLM-generated text. Section 3.3 extensively discusses how humans and LLMs differ in their quantitative use of commas. In this section, we investigate the differences at the lexical level.

G.1 Before Comma

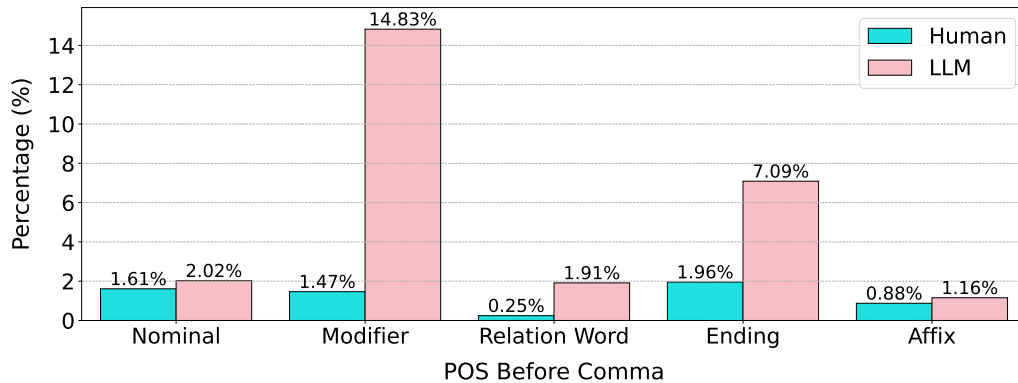
Figure 7 illustrates the ratio of each POS followed by a comma, calculated as the number of times a given POS appears before a comma divided by the total occurrences of that POS in the text. Endings and affixes are morpheme-level tags rather than POS tags. However, following the tagged output of the Bareun tagger, we include them in our analysis as individual categories. Predicates and interjections are excluded due to their sparsity.

Use of Commas After Endings While the frequency of a comma after a POS is consistently higher for LLM-generated text, the difference is particularly notable in endings. The ratios of a connective ending followed by a comma are 19.83%, 15.57%, and 28.01% for LLM-generated essays, poetry, and paper abstracts, respectively, while they are 4.10%, 4.68%, and 13.27% for human-written text. This discrepancy indicates that LLMs systematically overuse commas with connective endings, diverging from common patterns in human-written Korean text.

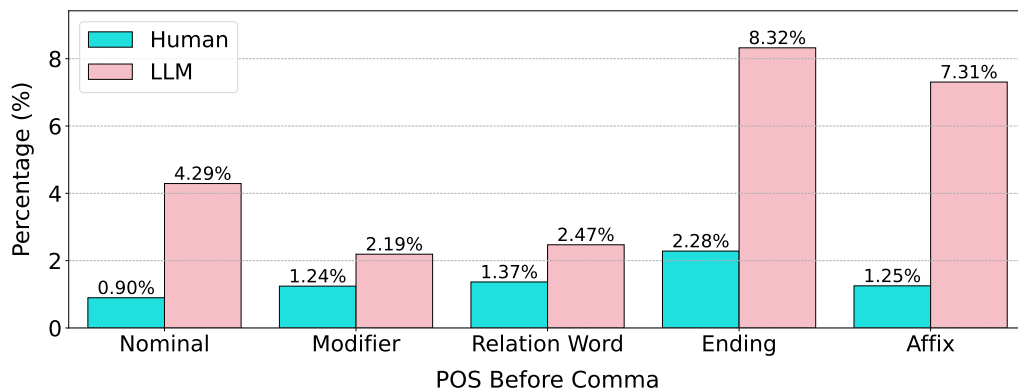
Use of Commas After Modifiers The Korean orthography guidelines state that it is natural not to use a comma after conjunctive adverbs such as *그리고* ('and' in function), *그러나*, *그런데* ('however'), and *그러므로* ('therefore'), as the functions of conjunctive adverbs and commas overlap. Unlike English, which often requires a comma in such cases, Korean does not, but its use remains a matter of stylistic preference. The tendency of LLMs to insert unnecessary commas after these modifiers suggests an influence of English punctuation norms on their output.

Influence of Multilingual Training The higher frequency of commas following connective endings and modifiers in LLM-generated Korean text may be influenced by the biases ingrained in multilingual language models. Prior research highlights how multilingual models exhibit preferences shaped by dominant training languages (Wendler et al., 2024). Given that LLMs are trained on multilingual data with English-dominated

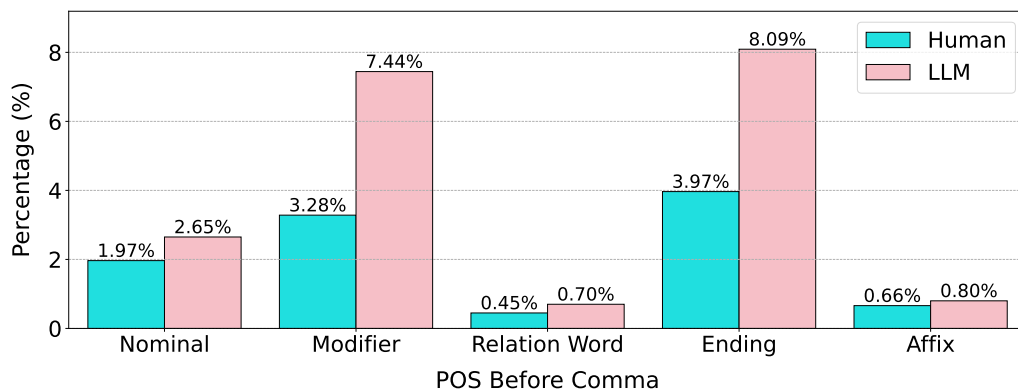
corpora, they may internalize English punctuation conventions, where commas frequently appear before conjunctions. This could lead LLMs to insert commas after connective endings more often than native Korean writers.



(a) Essay



(b) Poetry

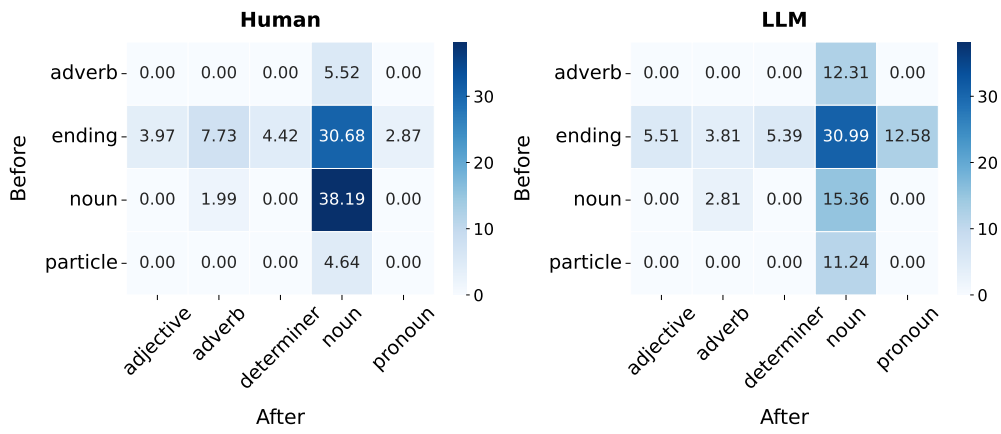


(c) Paper Abstract

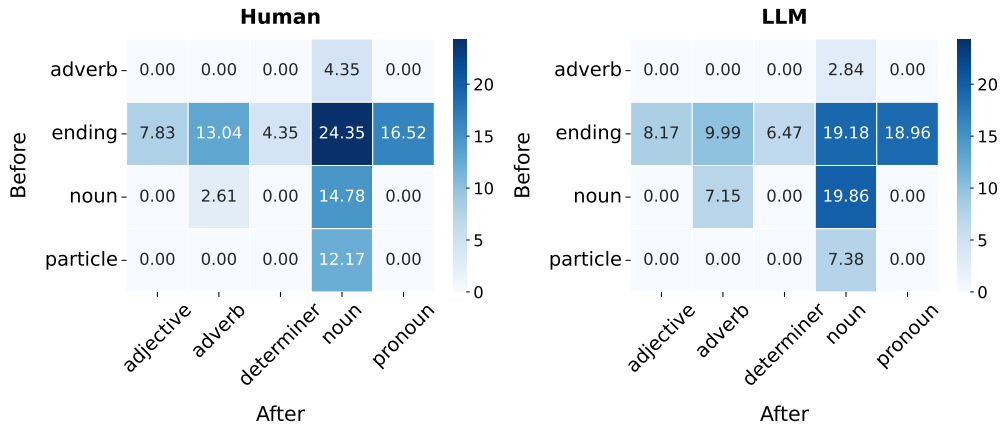
Figure 7: Histogram showing the percentage of each part of speech that appears before a comma, measured by the number of times a given POS is followed by a comma relative to its total occurrences in the text.

G.2 Before and After Comma

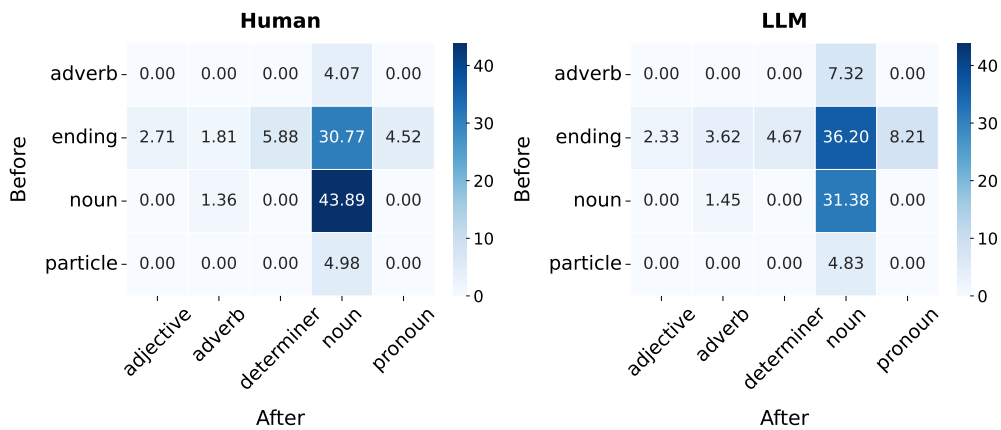
Figure 8 shows the normalized POS pair frequencies, calculated as the proportion of each POS pair relative to the total number of comma uses. Sparse cases have been omitted. Both humans and LLMs primarily use commas in noun-noun and ending-noun pairs. Especially in essays, however, human-written texts exhibit a more concentrated pattern, with a few POS pairs showing high frequencies. In contrast, LLM-generated text tends to display a more evenly distributed pattern across different POS pairs. This suggests that humans use commas more selectively, whereas LLMs apply them more broadly.



(a) Essay



(b) Poetry



(c) Paper Abstract

Figure 8: Normalized POS pair frequencies around a comma, expressed as the percentage of each POS pair relative to the total number of comma uses.

H Summary of Linguistic Features

This section presents a summary of the linguistic features employed in KatFishNet. The feature set falls into three primary categories: word spacing, POS combinations, and punctuation. Table 8 outlines each feature with a brief description.

Category	Feature	Description
Word Spacing	MMN-BN Space Ratio	The frequency of word spacing between a MMN and a BN.
	BN Space Ratio	The frequency of word spacing before a BN.
	VX Space Ratio	The frequency of word spacing before a VX, excluding the specific case of "-아"/"-어" (ENDING) + "지" (VX), where spacing is strictly prohibited.
POS Combinations	N-gram Diversity	The syntactic variability in a text by computing the ratio of unique POS n-grams to the total number of POS n-grams.
Punctuation	Comma Inclusion Rate	The proportion of sentences containing at least one comma out of all sentences in a text.
	Comma Usage Rate	The number of commas in a sentence divided by the total number of morphemes in that sentence.
	Relative Position	The position of each comma (counting the number of morphemes before it) divided by the total number of morphemes in the sentence.
	Segment Length POS Diversity Score	The average length of sentence segments split by commas. The diversity of part-of-speech pairs appearing before and after a comma.

Table 8: Overview of linguistic features used by KatFishNet across three categories.

I Implementation Details

When building the KatFish dataset, we access GPT-4o and Solar through their official APIs and use Qwen2 and Llama3.1 via Ollama⁵. We implement the proposed detection method, KatFishNet, using machine learning models provided by Scikit-learn. For the fine-tuning baseline, we use the chatgpt-detector-roberta⁶ model released by the HC3 dataset authors on HuggingFace as the base model and fine-tune it for five epochs using the KatFish dataset. When reproducing confidence-based and perturbation-based baselines, we use the implementations provided by MGTBench (He et al., 2023). For LLM paraphrasing and LLM prompting baselines, we access Exaone 3.5 through Ollama. We conduct experiments on a server equipped with an NVIDIA RTX A6000 with 48GB of memory.

⁵<https://ollama.com/search>

⁶<https://huggingface.co/Hello-SimpleAI/chatgpt-detector-roberta>

J Performance of KatFishNet with Different Machine Learning Models

Table 9 presents the performance of KatFishNet with different base machine learning models. Below is a brief overview of the models incorporated in KatFishNet:

- **Logistic Regression:** A statistical model used when the dependent variable is categorical. It applies a sigmoid function to the input values to compute probabilities, and the regression coefficients help interpret how each feature influences the outcome.
- **Random Forest:** An ensemble method that constructs multiple decision trees and combines their predictions to produce the final output. Each tree is trained on randomly selected subsets of the data and features, thus enhancing the diversity and generalization capability of the model.
- **Support Vector Machine:** A classification algorithm that seeks an optimal decision boundary to separate data points. It learns by maximizing the margin between the decision boundary and the closest data points. This approach is effective in high-dimensional spaces and can handle complex data structures well.

We find that KatFishNet, which leverages comma usage patterns, achieves the highest performance in detecting LLM-generated Korean text, regardless of the type of machine learning model used as the backbone. Looking ahead, we also plan to develop an expanded version of KatFishNet based on ensemble or voting methods that integrate predictions from a diverse set of machine learning models. By understanding and leveraging the unique strengths of each model, we anticipate that our approach will achieve even stronger detection performance and robustness.

Genre	KatFishNet		→ Solar	→ Qwen2	→ Llama3.1	Average
Essay	Word Spacing	Logistic Regression	86.00	80.63	71.91	79.51
		Random Forest	81.56	75.53	68.35	75.14
		Support Vector Machine	82.46	78.29	69.08	76.61
	POS Combinations	Logistic Regression	92.26	83.10	73.63	82.99
		Random Forest	91.85	80.08	75.08	82.33
		Support Vector Machine	89.03	73.60	76.02	79.55
	Punctuation	Logistic Regression	97.57	94.63	92.45	94.88
		Random Forest	96.07	95.12	90.29	93.82
		Support Vector Machine	96.26	95.49	91.33	94.36
Poetry	Word Spacing	Logistic Regression	71.85	65.56	43.81	60.40
		Random Forest	59.24	63.04	52.47	58.25
		Support Vector Machine	67.33	69.79	47.96	61.69
	POS Combinations	Logistic Regression	39.41	79.17	53.32	57.30
		Random Forest	45.99	69.27	55.52	56.92
		Support Vector Machine	43.66	75.88	56.53	58.69
	Punctuation	Logistic Regression	62.65	93.45	63.22	73.10
		Random Forest	63.21	90.63	60.91	71.58
		Support Vector Machine	60.76	87.92	60.65	69.77
Paper Abstract	Word Spacing	Logistic Regression	57.73	66.91	49.36	58.00
		Random Forest	51.60	49.67	43.70	48.32
		Support Vector Machine	55.86	63.38	46.99	55.41
	POS Combinations	Logistic Regression	47.47	70.05	42.47	53.33
		Random Forest	54.05	55.29	43.92	51.08
		Support Vector Machine	52.62	60.82	49.37	54.27
	Punctuation	Logistic Regression	78.99	77.47	70.41	75.62
		Random Forest	75.51	77.32	67.61	73.48
		Support Vector Machine	79.51	76.35	69.59	75.15

Table 9: Performance of detecting LLM-generated Korean text. We report the average performance (AUC-ROC) over five experiments.

K Ensemble of Linguistic Features

We conduct an additional experiment to explore the effectiveness of combining word spacing, POS combinations, and punctuation features for LLM-generated Korean text detection. Specifically, we apply an ensemble approach by averaging the predicted probabilities (ranging from 0 to 1) from separate logistic regression models. A value of 1 indicates that the text is predicted to be LLM-generated. Tables 10, 11, and 12 show the experimental results.

Feature Combination	→ Solar	→ Qwen2	→ Llama3.1	Average
Word Spacing	86.00	80.63	71.91	79.51
POS Combinations	92.26	83.10	73.63	82.99
Punctuation	97.57	94.63	92.45	94.88
All	97.67	93.93	88.05	93.21
POS + Punctuation	98.97	95.09	90.58	94.88
POS + Spacing	90.29	83.65	73.73	82.55
Punctuation + Spacing	97.89	94.74	90.35	94.32

Table 10: Performance on LLM-generated essay detection using individual and ensemble combinations of linguistic features. The ensemble method averages outputs from separate logistic regression classifiers.

Feature Combination	→ Solar	→ Qwen2	→ Llama3.1	Average
Word Spacing	71.85	65.56	43.81	60.40
POS Combinations	39.41	79.17	53.32	57.30
Punctuation	62.65	93.45	63.22	73.10
All	63.66	96.04	59.94	73.21
POS + Punctuation	55.30	94.83	63.40	71.17
POS + Spacing	52.09	80.27	48.69	60.35
Punctuation + Spacing	73.78	95.41	58.64	75.94

Table 11: Performance on LLM-generated poetry detection using individual and ensemble combinations of linguistic features. The ensemble method averages outputs from separate logistic regression classifiers.

Feature Combination	→ Solar	→ Qwen2	→ Llama3.1	Average
Word Spacing	57.73	66.91	49.36	58.00
POS Combinations	47.47	70.05	42.47	53.33
Punctuation	78.99	77.47	70.41	75.62
All	68.93	80.00	57.01	68.64
POS + Punctuation	71.70	82.82	60.93	71.81
POS + Spacing	53.22	70.76	44.98	56.32
Punctuation + Spacing	77.81	79.70	65.63	74.38

Table 12: Performance on LLM-generated paper abstract detection using individual and ensemble combinations of linguistic features. The ensemble method averages outputs from separate logistic regression classifiers.

In most cases, punctuation alone achieves detection performance that is either higher or comparable to that of ensemble models. This reinforces the strong effectiveness of punctuation-related features in detecting LLM-generated Korean text. Combining POS and spacing leads to the lowest performance among the ensembles, suggesting that these features alone do not capture the high-impact signals that punctuation features provide. Building on these findings, we plan to explore ensemble or voting approaches using multiple machine learning models, with a particular emphasis on punctuation features.

L Methods for Detection Machine-Generated Text

This section serves as supplementary material to Section 6, providing an overview of methodologies for detecting LLM-generated text. Methods for detecting machine-generated text can be broadly grouped into three main categories:

Watermarking Watermarking methods (Kirchenbauer et al., 2023; Guo et al., 2024) detect machine-generated text by embedding recognizable patterns during the text generation process. These methods intervene in the text generation process of LLMs by dividing the vocabulary into green and red lists and prioritizing the generation of tokens from the green list. If the proportion of green tokens in a generated text exceeds a certain threshold, the text is classified as machine-generated.

White-Box Detection White-box detection methods detect machine-generated text by analyzing the log probability (Mitchell et al., 2023) or log rank (Su et al., 2023a) of the text. These methods rely on full or partial access to the text generator, which makes them difficult to use with commercial LLMs that restrict such access. Recently, Mireshghallah et al. (2024) proposed the use of surrogate models (e.g., GPT-2) to approximate log probability or log rank.

Black-Box Detection Black-box detection methods require only the target text and do not rely on access to the text generator. For example, Zhu et al. (2023) introduced a paraphrasing-based detection approach using LLMs. Their method is based on the hypothesis that machine-generated text aligns more closely with the generation logic and statistical patterns learned by LLMs than human-written text. Accordingly, they proposed identifying a text as machine-generated if its paraphrased version closely resembles the original.

M Detection Results From the LLMs Used to Generate the KatFish Dataset

We conduct an additional experiment to investigate whether the LLM that generated a given text can reliably distinguish between its own outputs and human-written texts. Table 13 presents the performance of the LLM-generated Korean text detection task using the LLMs employed in the creation of the KatFish dataset. The results suggest that even LLMs struggle to identify their own generated text, highlighting the limitations of LLM prompting baseline.

Genre	Detection LLMs	→ Solar	→ Qwen2	→ Llama3.1	Average
Essay	GPT-4o	57.75	49.70	46.26	51.23
	Solar	46.50	48.11	47.38	47.33
	Qwen2	62.04	61.03	54.79	59.28
	Llama3.1	52.64	37.95	45.11	45.23
Poetry	GPT-4o	51.70	52.28	50.06	51.34
	Solar	49.99	51.15	44.34	48.49
	Qwen2	47.17	49.91	45.79	47.62
	Llama3.1	64.93	48.58	50.60	54.70
Paper Abstract	GPT-4o	48.20	47.70	47.77	47.89
	Solar	51.49	52.32	53.05	52.28
	Qwen2	50.00	50.50	49.52	50.00
	Llama3.1	49.80	49.32	49.68	49.60

Table 13: Performance Comparison of Different LLMs in Detecting LLM-Generated Korean Text. We report the average performance (AUC-ROC) over five experiments. We separately report the performance of the detection model for the task of distinguishing between human-written text and text generated by a specific LLM.

N Detection with Knowledge-Integrated Few-Shot Prompting

We present an additional experiment that incorporates linguistic insights into the prompt for LLM-generated text detection. The experiment focuses on comma usage patterns, identified as the most effective feature in our main analysis.

N.1 Experimental Setup

We follow the equivalent dataset and experimental settings described in Section 4.3. We use Exaone 3.5, a Korean-focused LLM for this additional experiment. We compare the new experimental results with the simpler prompting approach from Table 3.

N.2 Prompt Design

Zero-shot Learning

Distinguish between Korean text written by a human and Korean text generated by an LLM. Provide the basis for your decision under "Reasoning:". Indicate the probability that the given text was generated by an LLM on a scale from 0 to 1 under "Answer:".

Korean texts written by LLMs tend to use commas more frequently than those written by humans. Additionally, LLMs are more likely to place commas toward the latter part of a sentence. They also tend to use a wider variety of part-of-speech combinations around commas. Use these characteristics to help you distinguish the text.

Text to be classified: {t}
Reasoning:
Answer:

Few-shot Learning

Distinguish between Korean text written by a human and Korean text generated by an LLM. Provide the basis for your decision under "Reasoning:". Indicate the probability that the given text was generated by an LLM on a scale from 0 to 1 under "Answer:".

Korean texts written by LLMs tend to use commas more frequently than those written by humans. Additionally, LLMs are more likely to place commas toward the latter part of a sentence. They also tend to use a wider variety of part-of-speech combinations around commas. Use these characteristics to help you distinguish the text.

Here are examples of Korean texts written by a human and by an LLM:
Human-written Korean text 1: {h1}
Human-written Korean text 2: {h2}
LLM-generated Korean text 1: {g1}
LLM-generated Korean text 2: {g2}
Text to be classified: {t}
Reasoning:
Answer:

N.3 Three Prompting Methods

We compare three prompting methods:

- **No Knowledge:** The default LLM prompting approach as in Table 3 (label: LLM Prompting Exaone 3.5), where the prompt simply asks whether the text is human-written or LLM-generated, without providing specific linguistic insights.
- **Zero-shot Learning:** The new prompt that provides linguistic insights about commas, without example texts of human/LLM writing.
- **Few-shot Learning:** The new prompt that provides linguistic insights about commas and two examples each of GPT4o-generated and human-written Korean texts.

N.4 Findings

Table 14 shows the experimental results:

- For poetry and paper abstracts, prompts that provide linguistic guidance about comma usage improve detection performance over the baseline. While Zero-shot yields limited benefit in essays, the Few-shot prompt leads to a noticeable improvement.
- Providing both linguistic insights and representative examples further enhances detection accuracy in nearly all cases.

Genre	Prompting	Solar	Qwen2	Llama3.1	Average
Essay	No Knowledge	50.42	49.74	50.07	50.07
	Zero-shot	51.89	47.98	46.39	48.75
	Few-shot	56.10	51.70	47.92	51.90
Poetry	No Knowledge	50.53	50.16	49.42	50.03
	Zero-shot	62.51	66.03	50.06	59.53
	Few-shot	67.16	66.75	57.35	63.75
Paper Abstract	No Knowledge	48.60	46.41	47.18	47.39
	Zero-shot	52.80	64.26	62.25	59.77
	Few-shot	53.81	66.61	59.91	60.11

Table 14: Comparison of detection performance across three LLM prompting strategies.

These results highlight the usefulness of incorporating linguistic insights—specifically, the differences in comma usage—into the prompt. We believe the findings signify the importance of detailed linguistic analysis, which can be cataloged and utilized in LLM context-learning approaches. We plan to explore this approach in more detail in our future work. For example, we may integrate chain-of-thought or other advanced prompting techniques with our linguistic analysis to further improve LLM-generated text detection.

O Robustness of KatFishNet Against Human-Style Imitation via LLM Prompting

We investigate the robustness of KatFishNet in scenarios where LLMs are prompted to imitate human writing style. We experiment to see how effectively KatFishNet can detect text that has been deliberately revised to appear more **human-like**, focusing on comma-related linguistic features.

O.1 Experimental Setup

We follow the experimental settings described in Section 4.3. Specifically:

- We select 20% of the human-written texts and all texts generated by Solar for our data.
- We then feed each Solar-generated text back into Solar, instructing it to revise the text in such a way that it avoids LLM-like comma usage patterns (*e.g.*, frequent comma placement, commas toward the end of sentences, diverse POS combinations around commas). Essentially, we ask the model to fix the text to make it more closely resemble human writing.
- We combine these revised Solar texts with the 20% human-written texts to form a new test dataset.
- We evaluate KatFishNet (using comma-related features) on this revised dataset.

O.2 Prompt Design

We create prompts for essays, poetry, and paper abstracts, which all share the same structure except for specifying which text type is being revised. Below is the essay-specific prompt example:

Essay Prompt

I'm going to give you a Korean essay written by an LLM.
Please revise it to make it sound like it was written by a human.
Provide the revised essay under "Revised Essay:".

Essays written by LLMs tend to use commas more frequently than those written by humans.
They also tend to place commas toward the latter part of a sentence and use a wider variety of part-of-speech combinations around commas.
Keep this in mind as you revise the essay to make it more human-like.

Essay to revise: {t}
Revised Essay:

O.3 Experimental Results

Table 15 compares Original (the results from Table 3, under the “→ Solar” column using Punctuation-based features) with Revised (detection performance on the newly revised texts).

Genre	Original	Revised
Essay	97.57	90.83
Poetry	62.65	75.91
Paper Abstract	78.99	80.69

Table 15: KatFishNet performance on Solar-generated texts before and after human-style revision.

Observations:

- **Essays:** There is a performance drop from 97.57% to 90.83%, yet the score remains above 90. This suggests that while the LLM manages some level of style manipulation, KatFishNet still maintains strong detection performance.

- **Poetry & Paper Abstracts:** Interestingly, we observe a performance increase (62.65→75.91 for poetry; 78.99→80.69 for abstracts). This suggests that instructing the LLM to alter comma usage may yield text that is even more easily flagged by the punctuation-based features of KatFishNet. This may be because the LLM performs human-like revisions based on an internal conception of human-written text, which does not fully align with actual human writing patterns.

These findings indicate that comma-related linguistic features are not trivially shallow. While commas may seem trivial, their usage reflects deeper, inherent characteristics of LLMs. Without a nuanced understanding of Korean stylistic conventions, LLMs struggle to convincingly mimic human-like text. Their multilingual nature also contributes to this challenge. This aligns with our hypothesis that LLMs may still have difficulty fully capturing the stylistic intricacies of Korean writing.

O.4 Further Justification

In our paper, we demonstrate that comma-related features are exceptionally useful for distinguishing between LLM-generated and human-written Korean texts. In particular:

- As discussed in Section 3.3, LLMs are trained on extensive multilingual datasets—often with a heavy emphasis on English. Since English comma usage differs considerably from Korean conventions, LLMs tend to adopt comma usage patterns that do not align with those commonly found in native Korean writing.
- As discussed in Section 5, comma usage reflects contextual and stylistic factors, making it highly variable depending on the intent of the writer. Consequently, it is challenging for LLMs to precisely mimic human writers.
- As detailed in Appendix G, we conduct an extensive analysis of the differences in part-of-speech patterns surrounding commas between human-written and LLM-generated texts.

Together, these results strongly support our hypothesis that comma-related features go beyond shallow surface-level analysis and capture critical stylistic and syntactic nuances inherent to Korean writing.

O.5 Future Directions

Adversarial attacks that aim to deceive LLM-generated text detectors represent an intriguing avenue for future exploration. We plan to explore prompt-engineered adversarial attacks, leveraging the linguistic distinctions found between human- and LLM-generated Korean texts, as part of our subsequent research.

P Morphological Analysis

We conduct a morphological analysis of Korean text written by humans and LLMs to analyze the frequency of different parts of speech. We perform the analysis using the HanNanum POS tagger, provided by the KoNLPy. Table 16 shows the English names of each POS along with their corresponding Korean names. In Korean, words are classified into five word types (nominal, predicate, modifier, interjection, relation word) based on their function and into nine parts of speech based on their meaning (noun, pronoun, numeral, verb, adjective, determiner, adverb, interjection, particle).

English	Korean	Description
Nominal	체언	A nominal is a part of speech that includes nouns, pronouns, and numeral classifiers, functioning as the subject or object in a sentence.
Predicate	용언	A predicate is a part of speech that includes verbs and adjectives, expressing actions, states, or descriptions while serving as the main component of the predicate in a sentence.
Modifier	수식언	A modifier is a part of speech that includes adverbs and determiners, providing additional information about other words, such as nouns or verbs, by describing or qualifying them.
Interjection	독립언	An interjection is a part of speech that expresses sudden emotions or reactions and stands independently within a sentence, often without a grammatical connection to other words.
Relation Word	관계언	A relation word, or particle, is a part of speech that links nouns, pronouns, or phrases to other words in a sentence, indicating grammatical relations such as subject, object, or possession.
Ending	어미	An ending is a part of speech attached to the stem of a verb or adjective, modifying its meaning, tense, mood, or form, and determining the sentence's grammatical structure.
Affix	접사	An affix is a part of speech that attaches to a root word, altering its meaning or grammatical function, and includes prefixes, suffixes, infixes, and circumfixes.
Symbol	기호	A symbol is a part of speech that includes non-alphabetic characters such as punctuation marks, numbers, and special signs, used to convey meaning or structure in writing.
Foreign Language	외국어	Foreign language refers to words or phrases borrowed from other languages, used within a text to convey specific meanings, often retaining their original form and pronunciation.

Table 16: English names of each part of speech along with their corresponding Korean names.

Figure 9 visualizes the distribution of POS usage in human and LLM-written Korean text.

P.1 Higher Usage of Endings and Predicates in Human-written text

Endings and predicates significantly contribute to the diversity of sentence structure and expression. Humans naturally use a variety of sentence endings and predicates to connect sentences fluidly, express emotions, or reinforce arguments. These elements are essential for shaping the flow and rhythm of writing.

In contrast, LLM-generated text show relatively lower usage of endings and predicates. This indicates that LLMs often create simpler and more formulaic sentence structures. In creative or emotionally rich writing, they tend to struggle with replicating the nuanced flow of sentences that human writers achieve. Guided by statistical patterns rather than the natural rhythm of language, LLMs frequently produce sentences that lack the dynamic quality characteristic of human writing.

P.2 Higher Usage of Nominals in LLM-generated text

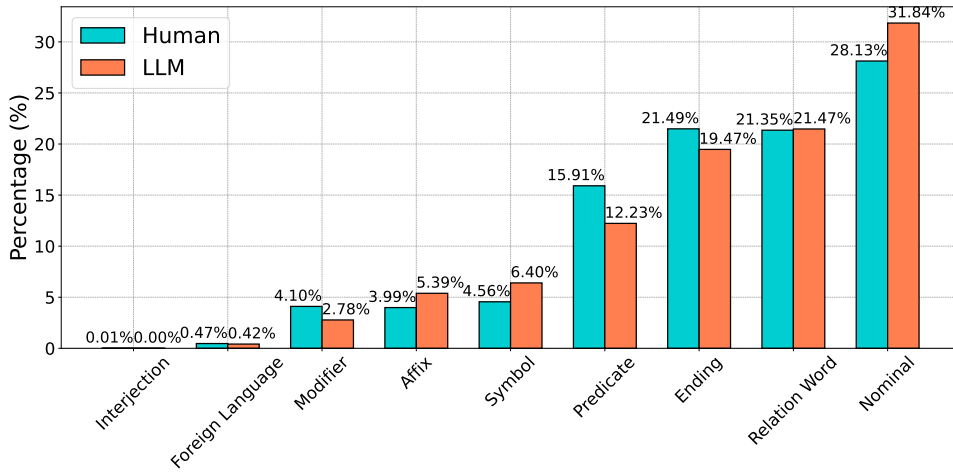
LLMs often rely heavily on nouns, focusing on generating sentences that are technical or primarily aimed at delivering information. Nouns play a key role in clearly conveying topics or concepts, and LLMs frequently build sentences around them, resulting in expressions that are direct and concise.

In contrast, human writers use nouns less frequently than LLMs, opting for a more balanced mix of verbs, adjectives, and adverbs to add depth and nuance to their writing. Human writing typically exhibits greater flexibility and complexity in sentence structures, reflecting context and emotion more effectively.

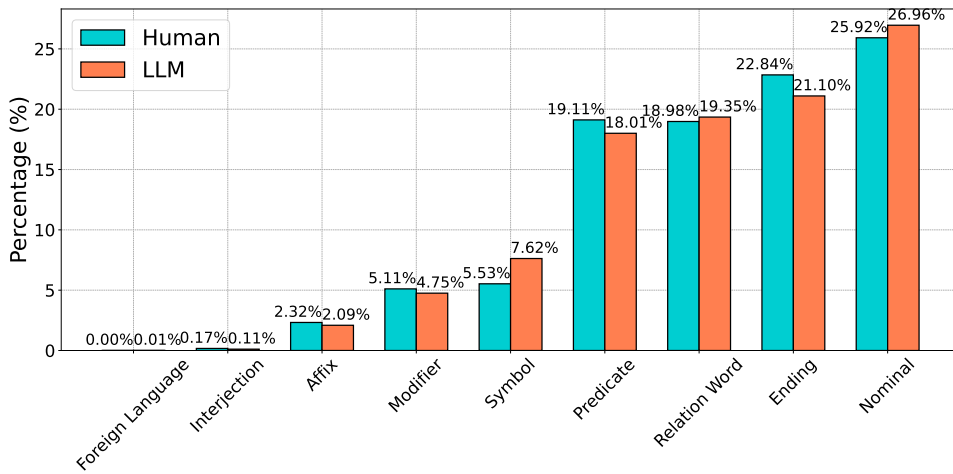
P.3 Overall Analysis

Human Writing is More Dynamic in Expression and Structure The abundant use of endings and predicates makes human writing appear more natural, capturing the flow of emotions and logic. This is because humans tend to connect sentences organically and emphasize a variety of expressions.

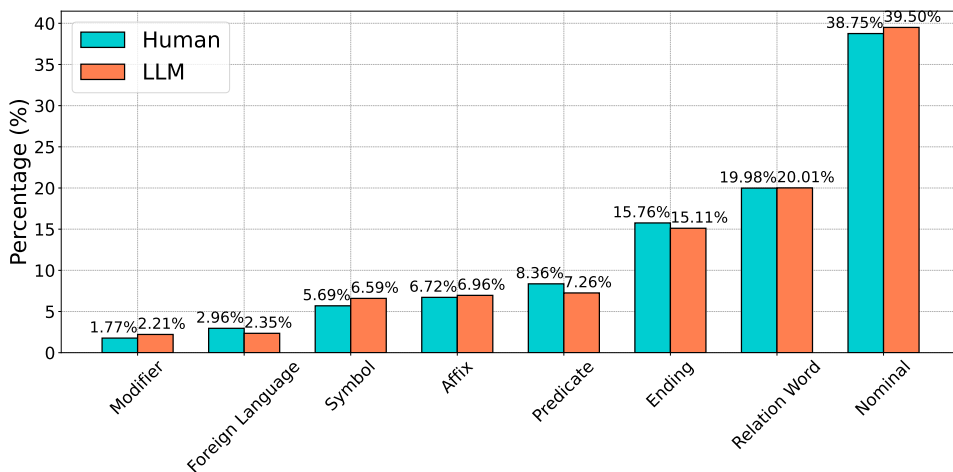
LLM-generated text are Optimized for Information Delivery The higher frequency of nouns indicates that LLMs focus on clarity and brevity, often structuring sentences in a more repetitive manner.



(a) Essay



(b) Poetry



(c) Paper Abstract

Figure 9: Histogram visualizing the distribution of part-of-speech usage in human-written and LLM-generated text.

Q Lexical Diversity Analysis

We use Zipf's Law (Zipf, 2016) and Heap's Law (Heaps, 1978) to compare and analyze the lexical diversity between human-written text and text generated by LLMs. Zipf's Law is an empirical rule that describes the frequency distribution of words in natural language texts. Zipf's Law illustrates the relationship between word frequency and rank, showing that frequently used words appear much more often than less common words. Heap's Law describes the relationship between the size of a text and the number of unique words (vocabulary size) it contains. Heap's Law examines how the number of unique words increases with the total word count, indicating how often new words are introduced as the text lengthens. Figure 10 shows Zipf's Law and Heap's Law for Korean text written by humans and LLMs.

First, as seen in Zipf's Law, text written by humans show a steep decline in the frequency of commonly used words, indicating that a wide variety of words are used. In contrast, text written by LLMs show a higher concentration of frequently used words, with noticeable repetition of certain vocabulary. This reflects the tendency of LLMs to repeatedly use words that frequently appear within learned patterns. Next, as observed in Heap's Law, text written by humans demonstrate a consistent increase in the number of unique words as the total word count rises, showcasing lexical diversity. On the other hand, text written by LLMs show a slower increase in unique words compared to humans, indicating relatively lower lexical diversity. This suggests that LLMs tend to use vocabulary repetitively. In conclusion, LLMs exhibit lower lexical diversity when writing texts compared to humans.

Finding. LLMs tend to use a narrower range of vocabulary and frequently repeat specific word patterns when writing, unlike humans who typically show greater lexical diversity.

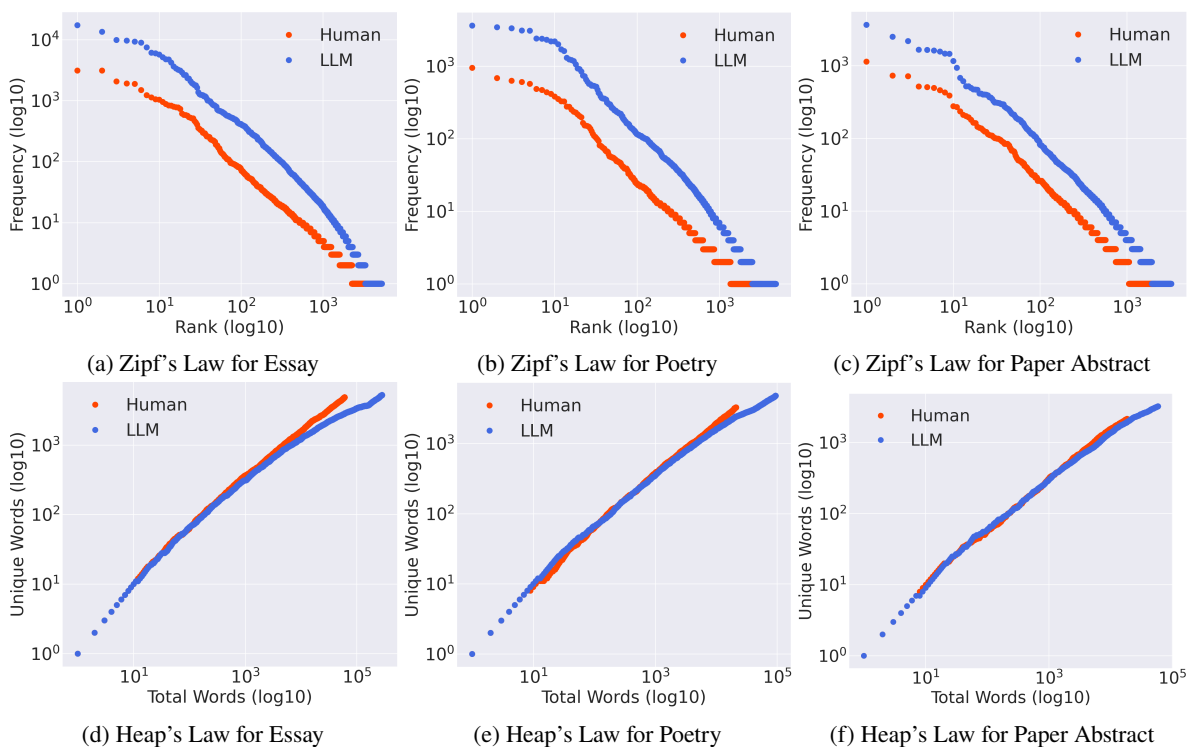


Figure 10: Comparison of Zipf's Law and Heap's Law between Korean text written by humans and those generated by LLMs.

R Expert Evaluation of Human-Written vs. LLM-Generated Korean Text

R.1 Expert Evaluation Settings

Data Preparation for Expert Evaluation We randomly select a total of 75 texts written by humans and four different LLMs. First, we randomly select text written by humans. For essays, we select two writings for each education level. For poetry, we select one poem each written by individuals under the age of 10 and over 40, and two poems each by individuals in their 10s and 20s & 30s. For paper abstracts, we select three different papers. We select essays and research papers written by LLMs that are on the same topics as those written by humans, and for research abstracts, we select abstracts from the same papers. Overall, we select a total of 30 essays, 30 poems, and 15 paper abstracts from humans and four LLMs.

Evaluator We request qualitative analysis for expert evaluation from three native Korean speakers specializing in Korean literature or Korean language education. Among them, one is a university student majoring in Korean literature, and the other two are current high school Korean language teachers.

Evaluation Rubric We design an evaluation rubric to assess text written by humans and generated by LLMs from multiple perspectives. For essays and paper abstracts, we categorize the evaluation criteria into Language, Organization, and Content, creating specific subcategories for each. In poetry, we evaluate based on Poetic Diction, Organization, Content, and Creativity, also detailing subcategories for these main categories. Each detailed evaluation item is rated on a 3-point scale to standardize the assessment process across different text types. This structured approach allows us to systematically compare the qualitative aspects of human and LLM-written Korean text. We have our rubric reviewed by an independent expert, a high school Korean language teacher specializing in Korean language education, alongside the evaluators involved in the expert evaluation. The evaluation rubrics we designed can be found in Figure 11.

Essay

- **Language** focuses on grammatical accuracy and semantic clarity, emphasizing the importance of clear communication in essays, which is crucial for conveying arguments effectively.
- **Organization** evaluates the logical structure, which is essential for maintaining a coherent flow of ideas and ensuring the reader can follow the argument easily.
- **Content** addresses the purpose of essays, which is to present arguments. Criteria like argument clarity, use of evidence, comprehension, and extension of ideas beyond the given passage are central because essays are often judged on their analytical depth and ability to engage with the topic.

Poetry

- **Poetic Diction** examines imagery and poetic devices, which are fundamental in poetry for evoking emotions and creating depth.
- **Organization** looks at the completeness of the poetic structure, acknowledging that form and structure are as vital as content in poetry.
- **Content** emphasizes emotion, sensitivity, and thematic clarity, highlighting the role of poetry in conveying complex emotions and abstract themes.
- **Creativity** evaluates the originality of content, diction, and organization, recognizing the reliance of poetry on creative expression.

Paper Abstract

- **Language** focuses on grammatical accuracy and sentence cohesion.
- **Organization** evaluates the abstract structure, ensuring that it follows the conventional format that guides readers through the research purpose, methods, and findings.
- **Content** targets the clarity of the research topic, purpose, and results, ensuring that abstracts succinctly summarize the essential components of the study.

	Criteria	3 (Excellent)	2 (Average)	1 (Poor)
표현 Language	문법의 적절성 Grammatical Accuracy	문법 오류가 거의 없다. There are almost no grammatical errors.	약간의 문법 오류가 있으나 전체적으로 이해하는 데 문제는 없다. There are some grammatical errors, but they do not interfere with overall understanding.	문법 오류가 자주 나타나며 이해하기 어려운 부분이 있다. Grammatical errors appear frequently, making it difficult to understand certain parts.
	문장 표현의 명확성 Semantic Clarity	문장이 명확하고, 의도가 효과적으로 전달된다. Sentences are clear, and the intent is effectively conveyed.	문장이 대체로 명확하나 일부 불분명한 부분이 있다. Sentences are generally clear, but some parts lack clarity.	문장이 모호하거나 전달하려는 의도가 불명확하다. 불필요하거나 장황한 표현이 사용되었다. Sentences are ambiguous, or the intent is unclear. Unnecessary or verbose expressions are used.
구성 Organization	글 구조의 적절성 Logical Structure	논리적으로 구성되어 있으며 전체적으로 일관성을 가진다. (예: 서론 - 본론 - 결론) The essay is structured logically and is coherent throughout the whole essay. (e.g., introduction - body - conclusion.)	글의 구조가 대체로 논리적이거나 일부 부분에서는 일관성이 부족하다 The essay is generally structured logically, but some parts lack coherency.	글의 구조가 논리적이지 않아 주요 내용을 따라가기 어렵다. The essay lacks a logical structure, making it difficult to follow and understand the main points.
내용 Content	주장의 명료성 및 일관성 Clear and Consistent Argument	주장이 명확하게 전달되고 글 전체에 일관되게 유지된다. The argument is clearly conveyed and consistent throughout the essay.	주장이 대체로 명확하나 일부 부분에서 불분명하다. The argument is generally clear but is unclear in some parts.	주장이 명확하지 않거나 일관성이 부족하다. The argument is unclear or lacks consistency.
	근거의 적절성 Usage of Relevant Supporting Evidence	근거가 주장을 효과적으로 뒷받침하며 논리적이다. The evidence effectively supports the argument and is logical.	근거가 대체로 주장을 뒷받침하나 일부 약한 근거가 있다. The evidence generally supports the argument, but some evidence is weak.	근거가 부족하거나 주장을 뒷받침하지 못한다. The evidence is insufficient or fails to support the argument.
	지문의 사고의 확장성 Novelty Beyond the Given Passage	지문에서 제시한 정보 이외의 본인의 사고 및 아이디어가 담겨 있다. The essay contains novel thinking and ideas beyond the information provided in the passage.	지문에서 제시하는 정보에 다소 국한되어 작성하였다. The essay is somewhat limited to the information presented in the provided passage.	지문에서 제시하는 정보에 의존하여 작성하였다. The essay relies solely on the information presented in the passage.
	지문 독해력 Comprehension of the Given Passage	지문을 정확히 이해하고 적절한 답변을 제시하였다. The given passage is thoroughly understood, and an appropriate response is provided.	지문을 대체로 이해하였으나 일부 부분에서 오해가 있거나 답변을 누락하는 경우가 있다. The passage is generally understood, but there may be some misunderstandings or omissions in the response.	지문을 제대로 이해하지 못하였고 답변이 부적절하거나 답변을 하지 않았다. The passage is insufficiently understood, and the response is inappropriate or missing.

(a) Essay

	Criteria	3 (Excellent)	2 (Average)	1 (Poor)
시어 Poetic Diction	심상의 명확성 Vivid Imagery	심상이 매우 명확하게 드러나며, 독자에게 강한 인상을 남긴다. The imagery is very vivid and leaves a strong impression on the reader.	심상이 대체로 드러나지만, 일부 부분에서 약하다. The imagery is generally present, but some parts are weak.	심상이 제대로 드러나지 않아, 독자에게 전달되지 않는다. The imagery is too weak and does not reach the reader.
	시적 장치의 활용 Use of Poetic Devices	시적 장치가 매우 효과적으로 사용되었다. Poetic devices are used very effectively.	시적 장치가 사용되었으나, 효과가 미흡하다. Poetic devices are used, but their effect is limited.	시적 장치가 거의 사용되지 않았다. Poetic devices are barely used.
구성 Organization	시 구조의 완성도 Poetic Structure	시의 구조가 잘 구성되어 있으며 주제를 효과적으로 전달하기에 알맞다. The poem is well structured and leveraged effectively to convey the theme.	시의 구조가 대체로 잘 구성되어 있으나 개선될 여지가 있다. The poem's structure is generally well-organized, but there is room for improvement.	시의 구조가 미흡하여 주제를 전달하는데 방해가 된다. The poem is poorly structured, hindering the delivery of the theme.
내용 Content	정서 및 감수성 Emotion and Sensitivity	정서와 감수성이 깊이 드러난다. The poem deeply expresses emotion and sensitivity.	정서와 감수성이 대체로 드러나지만 깊이가 부족하다. Emotion and sensitivity are generally expressed but lack depth.	정서와 감수성이 제대로 드러나지 않거나 전혀 느껴지지 않는다. Emotion and sensitivity are not well expressed or are not felt at all.
	주제의 명확성 Clarity of the Theme	주제가 명확하게 드러나며, 시 전체에 일관되게 유지된다. The theme is clearly expressed and consistent throughout the poem.	주제가 대체로 명확하나, 일부 부분에서 불분명하다. The theme is generally clear, but it is unclear in some parts.	주제가 명확하지 않거나, 일관성이 부족하다. The theme is unclear or lacks consistency.
창의성 Creativity	창의성 (시어, 구성, 내용) Creativity (Diction, Organization, Content)	시어, 구성, 또는 내용에서 독창적이고 신선한 아이디어가 돋보이며, 일반적으로 지 않는 방식으로 시의 주제를 효과적으로 전달하고 있다. The poem stands out with original and fresh ideas in theme, organization, or content, effectively conveying the theme in an unconventional way.	시어, 구성, 또는 내용에서 일부 창의적인 요소를 포함하고 있지만, 전반적으로 평범하거나 예측 가능한 방식으로 주제를 전달하고 있다. The poem includes some creative elements in diction, organization, or content but conveys the theme in a generally predictable or conventional way.	시어, 구성, 내용에서 창의성이 부족하고, 흔하거나 진부한 방식으로 주제를 전달하고 있다. The poem lacks creativity in diction, organization, and content and conveys the theme in a common or cliché manner.

(b) Poetry

	Criteria	3 (Excellent)	2 (Average)	1 (Poor)
표현 Language	문법의 적절성 Grammatical Accuracy	문법 오류가 거의 없다. There are almost no grammatical errors.	약간의 문법 오류가 있으나 전체적으로 이해하는 데 문제는 없다. There are some grammatical errors, but they do not interfere with overall understanding.	문법 오류가 자주 나타나 이해하기 어려운 부분이 있다. Grammatical errors appear frequently, making it difficult to understand certain parts.
	문장의 가독성 Sentence Cohesion	문장이 명확하고 이해하기 쉽다. The sentences are clear and easy to understand.	일부 문장이 모호하거나 이해하기 어렵다. Some sentences are ambiguous or hard to understand.	전반적으로 문장이 복잡하고 이해하기 어렵다. Overall, the sentences are complex and difficult to understand.
구성 Organization	초록 구조의 적절성 Abstract Structure	초록이 명확한 구조를 가지고 있다. (예: 연구 배경 - 기존 연구 한계점 - 제안하는 연구 방법론 - 연구 기여도 및 발견 사항) The abstract has a clear structure. (e.g., research background - limitations of previous research - proposed methodology - contributions and findings)	초록의 구조가 다소 적절치 않으나 연구 내용을 이해하는 데 문제는 없다. The structure of the abstract is somewhat inappropriate but does not hinder understanding of the research content.	초록의 구조가 적절치 않다. The structure of the abstract is inappropriate.
내용 Content	연구 주제 및 목적의 명확성 Clarity of the Research Topic and Purpose	연구 주제와 목적이 명확히 정의되어 있으며 이해하기 쉽다. The research topic and purpose are clearly defined and easy to understand.	연구 주제와 목적이 다소 모호하거나 이해하기 어렵다. The research topic and purpose are somewhat ambiguous or hard to understand.	연구 주제와 목적이 불명확하며 이해하기 어렵다. The research topic and purpose are unclear and difficult to understand.
	연구 결과 요약 Summary of the Research Results	연구 결과가 명확하게 요약되어 있다. The research results are clearly summarized.	연구 결과 요약이 다소 불충분하거나 불명확하다. The summary of the research results is somewhat insufficient or unclear.	연구 결과 요약이 거의 없거나 매우 불명확하다. There is little to no summary of the research results, or it is very unclear.

(c) Paper Abstract

Figure 11: Evaluation rubrics for expert evaluation.

Guidelines for Evaluation We provide our evaluation rubric along with the texts to evaluators with the following guidelines: 1) Essay: We provide the topic of the essay, the education level of the author, and the essay prompt along with the text to be evaluated. 2) Poetry: For poetry, we include the age group of the author along with the text being evaluated. 3) Paper Abstract: We provide the title of the research paper and the full content of the paper along with the abstract to be evaluated. These guidelines ensure that evaluators have all necessary context to accurately assess the texts across different categories and criteria. Figures 12 and 13 respectively show an example essay for expert evaluation, and its English translation. We ensure that evaluators do not make biased assessments by not providing information on whether each evaluated text is written by a human or generated by LLMs.

3. 에세이 (주장하는 글)

[주제1]

주제	#1	다문화 가족을 대하는 본인의 자세	학교	초등학교
지문/질문	<p>다문화 가족은 아버지 혹은 어머니가 외국 국적의 사람일 경우를 말합니다. 국가 정책을 통해 다문화 가정에 대한 이해 교육을 펼쳐 국민들의 생각을 변화시키려 하고 있지만, 여전히 사회나 학교 안에서 차별과 편견이 남아있습니다. 생김새와 문화가 한국인과는 다르다는 이유로 다문화 가정인 아이들이 사람들이 보는 시선에 의해 상처받는 경우가 많습니다. 다문화 가족의 아이들이 차별 혹은 편견 받지 않고 우리와 동일한 한국인으로 인정받고 존중하는 사회를 만들기 위해서는 어떠한 노력이 필요할까요? 여러분이 생각하기에 가장 먼저 실천되어야 하는 것이 무엇인지 작성해주시고 그 이유에 관해 설명해주세요.</p>			
고유번호	본문			
에세이_1_1	<p>우리의 사회 안에서, 차별과 편견이 여전히 존재하고 있습니다. 특히, 생김새나 문화가 조금 다른 다문화 가정의 아이들이 이로 인해 상처받는 경우가 많습니다. 모든 사람이 동등한 인간이라는 것을 인정하고 존중하는 세상을 만들어야 합니다. 그러나, 그런 사회를 만들기 위해서 가장 먼저 실천되어야 할 일은 바로 교육입니다. 첫째로, 학교교육에 다문화 이해의 중요성을 강조해야 합니다. 우리 아이들에게는 생김새나 국적, 인종 등의 차이가 존재한다는 사실 자체를 알려주는 것뿐 아니라, 그 차이에 대해 공포감이나 편견을 가지지 말아야 합니다. 다양한 배경과 문화 속에서 성장하는 사람들에 대한 이해와 존중의 중요성을 가르쳐야 합니다. 둘째로, 사회 전반적인 다문화 인식 개선 교육도 필요합니다. 특히, 매체를 통해 다문화 가정이 보여지는 방식을 바꾸어야 합니다. 외모나 문화적 차이에 주목하지 말고 그들의 인간성을 강조하는 내용으로 변화시켜야 합니다. 이를 통해 국민들에게 다문화 가족 역시 우리 사회의 중요한 구성원임을 알리는 교육이 필요합니다. 셋째로, 법률적인 보호와 지원도 중요합니다. 차별이나 편견에 대해 강력하게 처벌하고, 다문화 가정의 아이들이 편견 없이 공평한 기회를 누릴 수 있도록 제도적 보완이 필요합니다. 마지막으로, 우리 모두 스스로의 행동을 바꾸어야 합니다. 우리는 이미 우리의 사회에서 다양한 차이가 존재한다는 사실에 익숙해져야 합니다. 다른 사람들과 다르다는 이유로 그들을 판단하거나 편견을 가질 대신, 이해하고 존중하는 자세를 보여줘야 합니다. 우리 모두 함께 노력하여 우리 주변의 다문화 가족들이 차별이나 편견 받지 않고 동일한 인간으로 인정받고 존중받는 세상을 만들어야 합니다. 이 모든 것이 이루어진다면, 우리는 진정으로 공평하고 다채로운 사회를 경험하게 될 것입니다.</p>			

Figure 12: Example essay for expert evaluation.

3. Essay (Argumentative Essay)

[Topic 1]

Topic	#1	Your Attitude Towards Multicultural Families	Education Level	Elementary School
Passage/ Question	<p>A multicultural family is a family where the father or mother holds a foreign nationality. Although efforts are being made to change public perception through national policies that promote education on multicultural families, discrimination and prejudice still persist in society and schools. Many children from multicultural families are often hurt by the way others look at them simply because their appearance or culture differs from that of Koreans. What can we do to create a society where children from multicultural families are treated equally as Koreans, free from discrimination or prejudice? Please write down what you think should be implemented first, and explain the reason for your choice.</p>			
ID	Text			
Essay_1_1	<p>Discrimination and prejudice still exist in our society. Children from multicultural families are often hurt simply because they look or act differently. We must create a world that recognizes and respects every person as an equal human being. However, the first step toward building such a society is education. First, schools need to stress the importance of understanding multiculturalism. Children should be taught not only that differences in appearance, nationality, and race exist but also that these differences should not be feared or met with prejudice. It's crucial to teach the value of understanding and respecting people from diverse backgrounds and cultures. Second, there must be a broader effort to improve multicultural awareness across society. This includes changing how the media portrays multicultural families. Instead of focusing on physical or cultural differences, the emphasis should be on their humanity. This shift in perspective will help convey to the public that multicultural families are valuable members of our society. Second, it is necessary to improve multicultural awareness throughout society. In particular, the way multicultural families are portrayed in the media must change. Rather than focusing on differences in appearance or culture, the emphasis should be on their humanity. Through this approach, it is crucial to educate the public that multicultural families are also vital members of our</p>			

Figure 13: Example essay for expert evaluation (English translation).

R.2 Expert Evaluation Results

Figure 14 presents the expert evaluation results for essays written by humans and those generated by LLMs. We compare the results for text written by humans, commercial LLMs (GPT-4o and Solar), and open-source LLMs (Qwen2 and Llama3.1). We report the average evaluation results of the three human evaluators. Our analysis reveals that essays generated by LLMs receive better evaluations than those written by humans across seven metrics. Notably, essays by LLMs are distinctly rated higher in GRAMMATICAL ACCURACY compared to those written by humans. This is likely because LLMs are trained on extensive data, enabling them to learn grammatical rules and thereby generate texts with fewer grammatical errors. Overall, human-written essays score well in NOVELTY BEYOND THE GIVEN PASSAGE and COMPREHENSION OF THE GIVEN PASSAGE. One reason is that humans tend to weave in creative elements that go beyond the given context. However, human-written essays receive lower ratings in these two metrics compared to LLM-generated essays, which is an interesting result showing that LLMs not only properly understand and write about the given passage but also have the capability to utilize knowledge beyond the information provided in the passage.

The evaluators provide descriptive comments for each sample. The three main characteristics identified in LLM-generated essays are the excessive use of commas, repetition, and linguistic maturity. LLMs overuse commas, especially after adverbial case markers and connective endings. LLMs frequently repeat expressions or present paraphrased versions of the same sentence. Furthermore, LLM-generated text often exceed the expected writing level for the given educational stage, exhibiting complex vocabulary, advanced logic, and intricate sentence structures.

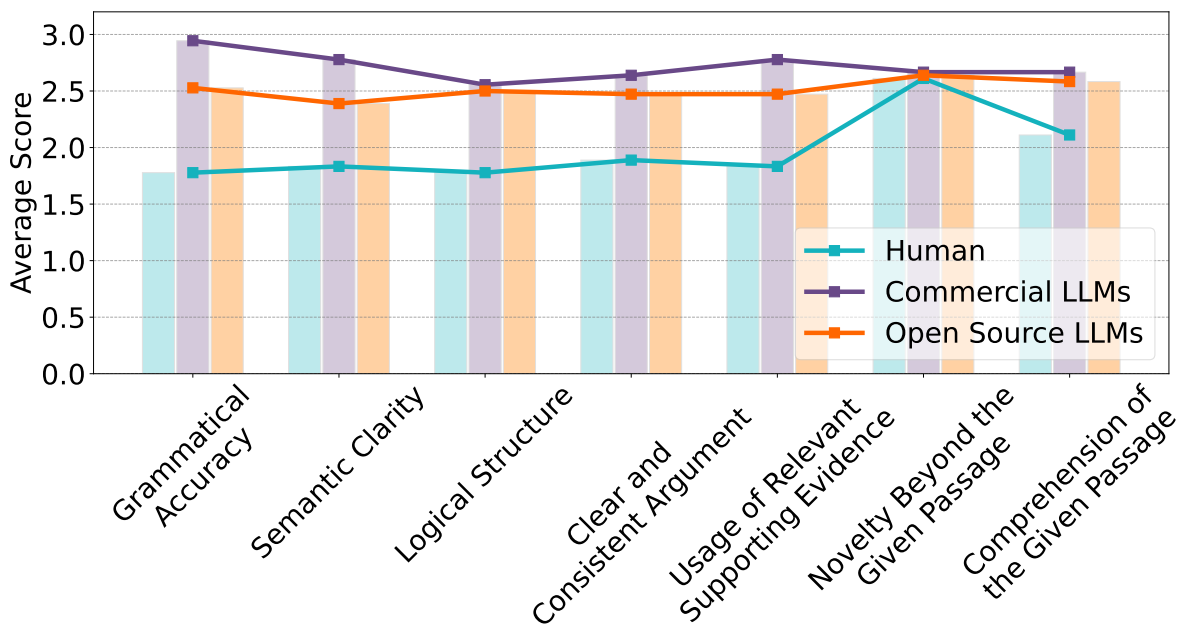


Figure 14: Results of expert evaluations for essays.

Figures 15 and 16 display the results of expert evaluations for poetry and paper abstracts, respectively. Our analysis indicates that for poetry, commercial LLMs generally receive higher ratings than open-source LLMs, and the evaluations for poetry written by humans and commercial LLMs are quite similar. For paper abstracts, those generated by commercial LLMs receive the highest ratings, while abstracts written by humans and open-source LLMs are rated similarly. Overall, while LLM-generated essays consistently outperform human-written ones across all metrics, the differences in evaluations for poetry and paper abstracts are less pronounced.

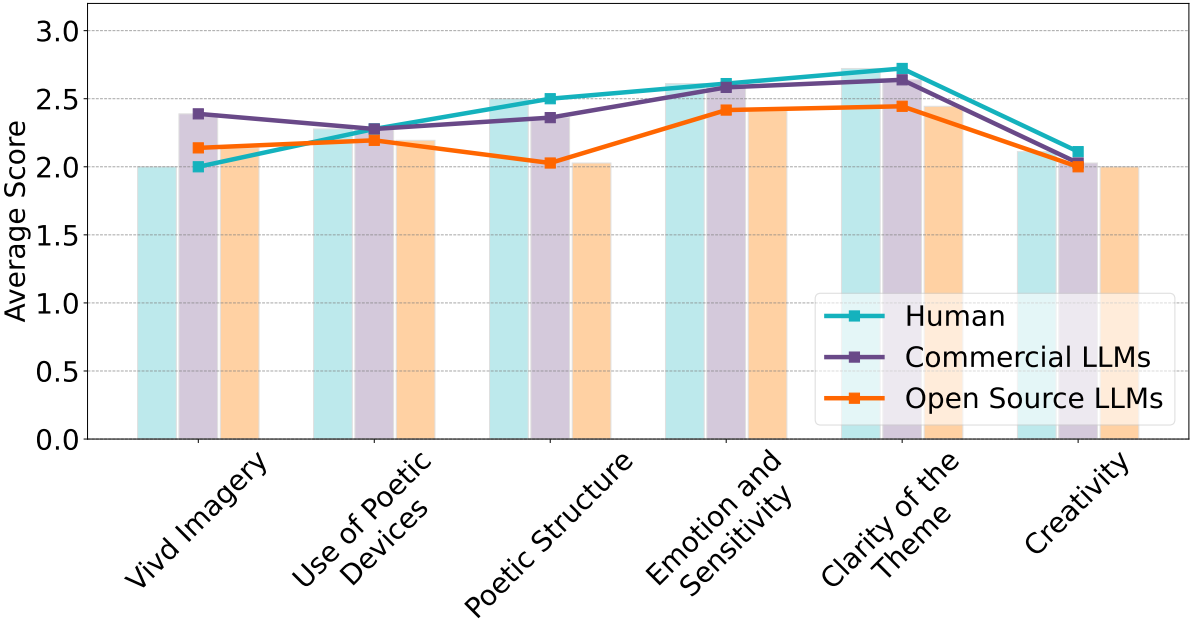


Figure 15: Results of expert evaluations for poetry.

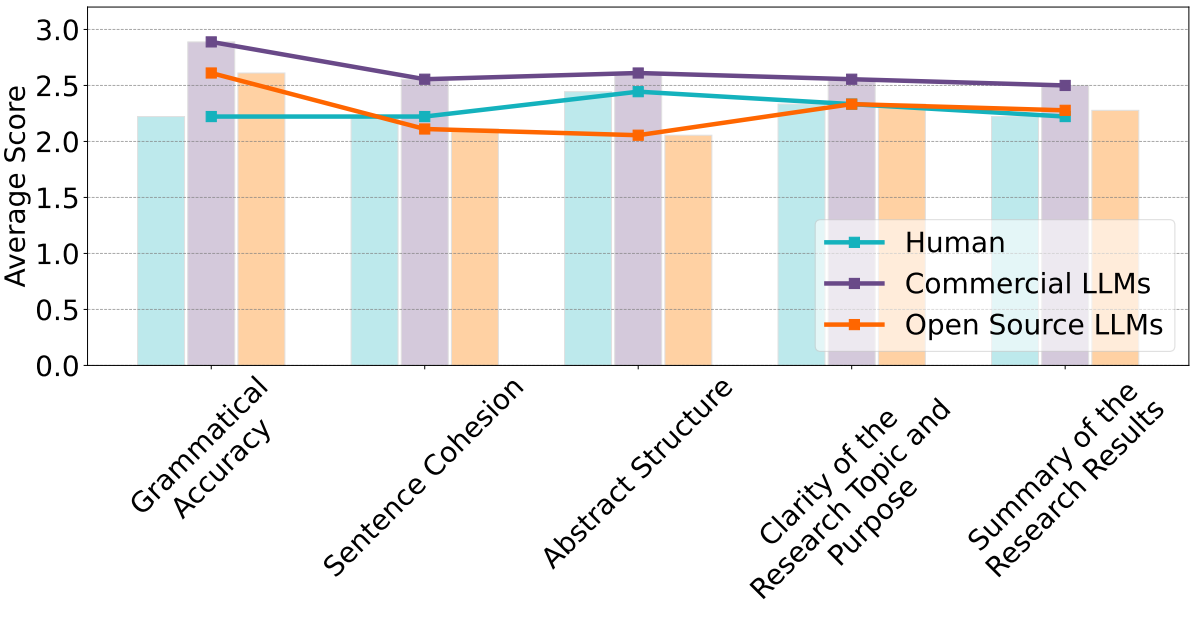


Figure 16: Results of expert evaluations for paper abstracts.