

# Navigating the Path of Writing: Outline-guided Text Generation with Large Language Models

**Yukyung Lee<sup>1,†</sup> Soonwon Ka<sup>2</sup> Bokyung Son<sup>2</sup> Pilsung Kang<sup>3</sup> Jaewook Kang<sup>2</sup>**

<sup>1</sup>Boston University <sup>2</sup>NAVER AI Platform <sup>3</sup>Seoul National University

ylee5@bu.edu pilsung\_kang@snu.ac.kr

{soonwon.ka, bo.son, jaewook.kang}@navercorp.com

## Abstract

Large Language Models (LLMs) have impacted the writing process, enhancing productivity by collaborating with humans in content creation platforms. However, generating high-quality, user-aligned text to satisfy real-world content creation needs remains challenging. We propose WritingPath, a framework that uses explicit outlines to guide LLMs in generating goal-oriented, high-quality text. Our approach draws inspiration from structured writing planning and reasoning paths, focusing on reflecting user intentions throughout the writing process. To validate our approach in real-world scenarios, we construct a diverse dataset from unstructured blog posts to benchmark writing performance and introduce a comprehensive evaluation framework assessing the quality of outlines and generated texts. Our evaluations with various LLMs demonstrate that the WritingPath approach significantly enhances text quality according to evaluations by both LLMs and professional writers.

## 1 Introduction

Writing is a fundamental means of structuring thoughts and conveying knowledge and personal opinions (Collins and Gentner, 1980). This process requires systematic planning and detailed review. Hayes (1980) describes writing as a complex problem-solving process and explores how planning and execution interact in writing. That is, writing involves more than merely generating text; it encompasses developing a proper understanding of the topic, gathering relevant subject matter, and implementing thorough structuring.

Recent advancements in Large Language Models (LLMs) have advanced the writing workflow, enhancing both its efficiency and productivity. One significant area of exploration is the collaborative

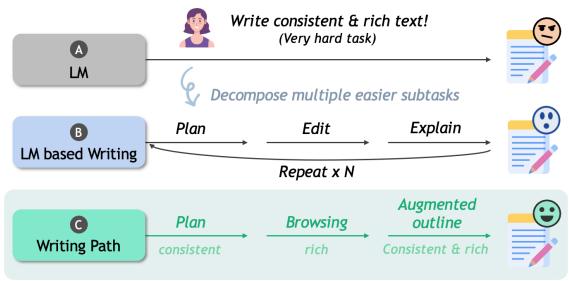


Figure 1: Comparative overview of writing approaches: (A) direct generation, (B) iterative writing involving planning, editing, and explaining, and (C) WritingPath method, which starts with a consistency-focused plan, incorporates information-rich browsing, and results in an augmented, consistent, and rich outline.

use of LLMs in writing processes (Lee et al., 2022; Mysore et al., 2023), as demonstrated by tools like Notion AI, Jasper, and Cohesive. The typical approach to incorporating LLMs involves the establishment of a writing plan and iterative improvement of interim outputs through revision Schick et al. (2023); Yang et al. (2022), as illustrated in Figure 1 (b) with a focus on utilizing the generative capabilities of LLMs to improve fluency, consistency, and grammatical accuracy. While these tools support users in creating content more efficiently, there remains room for improvement in maintaining consistent quality that accurately aligns with specific user intentions in production environments (Wang et al., 2024).

To address this, we propose **WritingPath**, a methodology designed to incorporate user intentions such as desired topic, textual flow, keyword inclusion, and search result integration into the writing process. WritingPath emphasizes the importance of systematic planning and a clear outline from the early stages of writing. Inspired by the structured writing plan of Hayes (1980) and the reasoning path of Wei et al. (2022), the WritingPath collects ideas and creates outlines that encapsu-

<sup>†</sup>Work done as a research intern at NAVER

late the user’s intentions before generating the final text. Furthermore, the initial outlines are further augmented with additional information through information browsing. Such a structured approach offers enhanced control over the text generation process and improves the quality of the content produced by LLMs.

We also utilize a multi-aspect writing evaluation framework to assess the intermediate and final productions from the WritingPath, offering a way to evaluate the quality of free-form text<sup>1</sup> without relying on reference texts. Taking into account that conventional Likert scales (1-5 ratings) (Clark; Hinkin, 1998) make it challenging to systematically compare and evaluate diverse writing outputs, particularly in creative tasks (Chakrabarty et al., 2023), our evaluation framework aims to provide more precise and reliable assessments for the outlines and final texts. For evaluation purposes, we construct a free-form blog text dataset incorporating a wide range of writing styles and topics from real users, including Beauty, Travel, Gardening, Cooking, and IT. Using this dataset, we evaluate how well the LLM outputs reflect the user’s intentions. Applying the WritingPath to various LLMs shows significant performance gains across all evaluated models. These results validate that our approach enables the models to maintain a stronger focus on the given topic and purpose, ultimately generating higher-quality text that more accurately reflects user intentions. Furthermore, to validate real-world applicability, we applied WritingPath to a commercial writing platform for beta testing from October 2023 to March 2024. The deployment demonstrated its effectiveness in supporting real users with structured content creation across diverse writing needs.

The main contributions of this study can be summarized as follows:

- We propose WritingPath, a novel framework that enhances the ability of LLMs to generate high-quality and goal-oriented pieces of writing by using explicit outlines.
- We customize a comprehensive evaluation framework that measures the quality of both the intermediate outlines and the final texts.

---

<sup>1</sup>Free-form text generation focuses on creating diverse texts tailored to specific information and user intentions, unlike story generation, which develops narratives with plots and characters

- We construct a diverse writing dataset from unstructured blog posts across multiple domains, providing useful information such as aligned human evaluation scores, such as metadata that can be used as input for LLM-based writing tasks, and aligned human evaluation scores for the generated texts.
- Our evaluation results indicate that the WritingPath markedly improves the quality of LLM-generated texts compared to methods that do not use intermediate outlines.

## 2 Design of WritingPath

We propose WritingPath, a systematic writing process to produce consistent, rich, and well-organized text with LLMs. Inspired by human writing processes, it consists of five key steps: metadata preparation, initial outline generation, information browsing, augmented outline creation, and final text writing (Figure 2). Each step is guided by a specific prompt configuration that aligns LLM output with specific step requirements.

The core components of WritingPath are those that generate outlines as they establish a structured writing plan. Research suggests that a well-structured outline significantly impacts the quality of the written text (Sun et al., 2022; Yang et al., 2022, 2023). The initial sketch is transformed into a detailed outline, including the flow, style, keywords, and relevant information from search results. This outline provides a clearer view of the final text to the LLMs. The specific steps are described as follows:

**Step 1: Prepare Meta Data** The first step establishes the writing direction and target reader using metadata  $m$ , which includes i) purpose, ii) type, iii) style, and iv) keywords. To simulate this process, we converted human-written texts into metadata (see Section 4.1 for details of the dataset).

**Step 2: Generate Title and Initial Outline** The second step generates the title  $t$  and initial outline  $O_{\text{init}}$  based on the metadata  $m$  from step 1, using the LLM function  $f_{\text{llm}}$  with a prompt configuration function  $\phi_s$ . Here,  $s$  indicates the step index, and for step 2, the prompt configuration is  $\phi_2$ :

$$t, O_{\text{init}} = f_{\text{llm}}(\phi_2(m)), \quad (1)$$

The initial outline  $O_{\text{init}}$  consists of main headers  $h_{i,0}$ , where  $i$  denotes the header sequence. This outline serves as the scaffolding of the text, organizing the main ideas and laying out the key points.

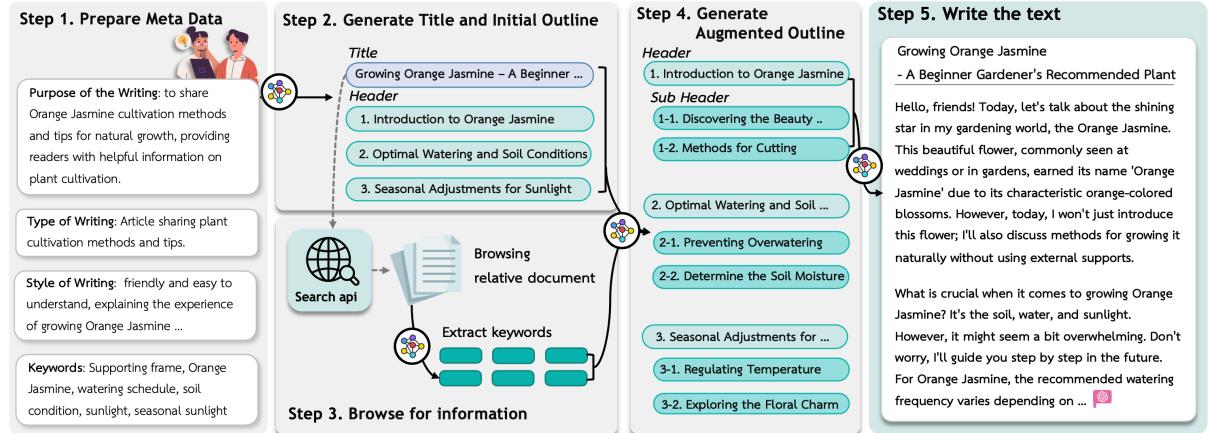


Figure 2: Main architecture of WritingPath, our proposed framework for guiding LLMs to generate high-quality text following a structured writing process. The WritingPath condenses text generation into five key steps. Inspired by human writing planning, it ensures alignment with specified writing goals.

**Step 3: Browse for Information** The third step enriches the text by collecting additional information and keywords to reinforce the initial outline. We use the search function  $f_{\text{search}}$  with the generated title  $t$  as the query to retrieve the top-1 blog document,  $D_{\text{sim}}$ :

$$D_{\text{sim}} = f_{\text{search}}(t) \quad (2)$$

In our implementation, we employ the NAVER search API<sup>2</sup> to retrieve the top-1 document among similar blog posts. From the blog document, we extract keywords  $K$  using the  $f_{\text{ilm}}$  with a prompt configuration  $\phi_3$ :

$$K = f_{\text{ilm}}(\phi_3(D_{\text{sim}})), \quad (3)$$

The extracted keywords constitute the additional information from the search results, leading to a more specific writing plan, improving the quality of the generated text.

**Step 4: Generate Augmented Outline** The fourth step refines the initial outline by adding subheadings and specific details to each section based on incorporating the keywords collected from the previous step. The augmented outline  $O_{\text{aug}}$  is generated using the LLM function  $f_{\text{ilm}}$  with a prompt configuration  $\phi_4$  that takes the title  $t$ , keywords  $K$ , and initial outline  $O_{\text{init}}$  as inputs:

$$\begin{aligned} O_{\text{aug}} &= f_{\text{ilm}}(\phi_4(t, K, O_{\text{init}})) \\ &= \{(h_{1,0}, \{h_{1,1}, h_{1,2}, \dots\}), \\ &\quad (h_{2,0}, \{h_{2,1}, h_{2,2}, \dots\}), \\ &\quad \dots\} \end{aligned} \quad (4)$$

The resulting augmented outline,  $O_{\text{aug}}$ , comprises headers ( $h_{i,0}$ ) and their corresponding subheaders

( $h_{i,j}$ ), where  $i$  denotes the header index, and  $j$  indexes the subheaders. This detailed structure serves as a comprehensive writing plan, breaking down the text into manageable parts and providing clear direction for the content.

**Step 5: Write the Text** Finally, the text for each section  $d^i$  is generated using the LLM function with a prompt configuration  $\phi_5$  that takes the title  $t$  and the corresponding section of the augmented outline  $O_{\text{aug}}^i$  as inputs:

$$d^i = f_{\text{ilm}}(\phi_5(t, O_{\text{aug}}^i)) \quad (5)$$

The final blog document  $D$  is then compiled by concatenating all sections:

$$D = \{d^1, d^2, \dots, d^n\} \quad (6)$$

WritingPath organically connects all steps of the writing process, employing an outline to aggregate and manage diverse information, and assists users in producing high-quality writing. The prompts utilized for the WritingPath are detailed in Figure 12 and 13.

### 3 Evaluation of WritingPath

Evaluating the effectiveness of WritingPath compared to existing writing support systems is challenging. Most previous studies do not directly utilize outlines in the writing process, resulting in a lack of systematic methods to assess outline quality. Even when outlines are used, evaluation relies only on human evaluation (Yang et al., 2023; Zhou et al., 2023). Moreover, current approaches heavily rely on human evaluation, which poses challenges for assessing full texts (Schick et al., 2023; Yang et al., 2022; Lee et al., 2023), as it requires evaluating multiple aspects of the written work. This

<sup>2</sup><https://developers.naver.com/docs/serviceapi>

challenge can arise in content creation workflows where scalable and consistent quality assessment helps maintain content standards.

To address these limitations, we propose an evaluation framework that combines human and automatic evaluation to assess the quality of generated outlines and final texts from multiple perspectives. This hybrid approach is designed to support real-world content creation workflows by combining systematic automated metrics with human assessment of nuanced writing aspects that require subjective judgment. The proposed method establishes clear evaluation criteria, enabling objective and reproducible validation of WritingPath’s effectiveness as a writing support system.

### 3.1 Outline Evaluation

#### 3.1.1 Automatic Evaluation

We adapt various metrics to evaluate the logical alignment, coherence, diversity, and repetition in outlines, following criteria established in linguistic literature (Van Dijk, 1977; Pitler and Nenkova, 2008; Tang et al., 2019; Elazar et al., 2021). Logical alignment, assessed through NLI-based methods, ensures that headers and subheaders are logically connected. Coherence evaluates thematic uniformity across sections, while diversity measures the breadth of topics covered. Repetition is analyzed to minimize redundancy and improve information efficiency. Note that coherence and diversity exhibit a trade-off relationship; maintaining coherence while covering a wide range of topics is essential to ensure the effectiveness of the outline in guiding the writing process. Detailed evaluation definitions are available in Appendix C.2.

#### 3.1.2 Human Evaluation

In addition to automatic evaluation metrics, we conduct a human evaluation to assess aspects of the generated outlines that are difficult to capture solely with automatic measures. These aspects include cohesion, natural flow, and redundancy. For augmented outlines, we also evaluate the usefulness of added information and overall improvement compared to the initial outline. Detailed evaluation definitions are available in Appendix C.3.

### 3.2 Writing Evaluation

Traditional evaluation metrics such as Likert scales are not well-suited for assessing creative tasks like long story generation (Chakrabarty et al., 2023). Acknowledging the need for more specific writing

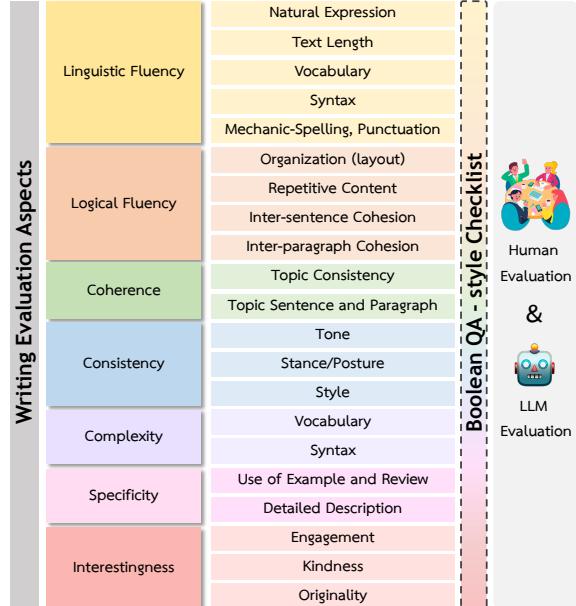


Figure 3: Breakdown of the seven key aspects used in writing evaluation, each with corresponding sub-aspects, employed in a Boolean QA-style checklist for human and LLM evaluation. This comprehensive framework ensures a multi-dimensional analysis of text quality.

evaluation methods, we employ CheckEval (Lee et al., 2024) to assess writing quality<sup>3</sup>. CheckEval decomposes the evaluation aspects into more granular sub-questions, forming a detailed checklist. These aspect-based checklists can make performance evaluations by either humans or LLMs more fine-grained. Moreover, by explicitly capturing the evaluator’s reasoning behind each rating, this approach enhances the explainability of the evaluation process. To adapt CheckEval, we identified 7 aspects and selected relevant sub-aspects for each. We formulated them as binary (Yes/No) questions. This resulted in a checklist-style evaluation sheet for each sub-aspect, enabling an intuitive and structured assessment of the generated texts. The prompts utilized for the writing evaluation are detailed in Figure 11. The evaluation criteria were selected based on prior linguistics research (Wolfe, 1997; Knoch, 2011; van der Lee et al., 2019; Celikyilmaz et al., 2020; Chhun et al., 2022; Sai et al., 2022; van der Lee et al., 2021) and finalized through a review and refinement process involving 6 writing experts. Details of the eval-

<sup>3</sup>Lee et al. (2024) reports a 0.65 spearman correlation between human and LLM evaluations for dialogue, which surpasses G-Eval (Liu et al., 2023). This demonstrates CheckEval’s potential as a reliable method for evaluating model-generated text quality.

Model		Automatic Evaluation				Human Evaluation			
Aspects Metrics		Logical Alignment NLI ( $\uparrow$ )	Coherence UCI ( $\uparrow$ ) / NPMI ( $\uparrow$ )	Diversity Topic Diversity ( $\uparrow$ )	Repetition Self-BLEU ( $\downarrow$ )	Cohesion ( $\uparrow$ )	Natural Flow ( $\uparrow$ )	Diversity ( $\uparrow$ )	Redundancy ( $\uparrow$ )
Eval Level		Header-Subheader		Outline	Outline	Outline	Outline	Outline	Outline
GPT-3.5	initial	-	0.60 / 0.31	0.60	32.03	<b>3.38</b>	2.70	2.77	2.73
	augmented	0.61	<b>1.33 / 0.51</b>	<b>0.61</b>	<b>17.33</b>	3.15	<b>2.78</b>	<b>3.54</b>	<b>3.13</b>
GPT-4	initial	-	0.80 / 0.49	0.67	24.81	<b>3.40</b>	2.86	3.06	2.86
	augmented	0.66	<b>1.61 / 0.52</b>	<b>0.68</b>	<b>13.12</b>	<b>3.40</b>	<b>2.98</b>	<b>3.74</b>	<b>3.43</b>
HyperCLOVA X	initial	-	0.75 / 0.41	0.74	18.04	<b>3.47</b>	2.96	2.82	3.22
	augmented	0.67	<b>1.82 / 0.54</b>	<b>0.75</b>	<b>11.50</b>	3.41	<b>3.48</b>	<b>3.93</b>	<b>3.79</b>

Table 1: Automatic and human evaluations on the quality of initial and augmented outlines from GPT-3.5, GPT-4, and HyperCLOVA X. Bold indicates the best result within a model.

ation criteria are in Figure 3, and the instructions and checklist used during the evaluation process are presented in Table 7.

## 4 Experimental Setting

### 4.1 Dataset

In this study, we constructed a Korean dataset based on real user-written blog posts to assess the effects of the WritingPath in real content creation scenarios. The dataset covers five domains frequently handled in content creation:: travel, beauty, gardening, IT, and cooking. We created a total of 1,500 posts for each model, resulting in 4,500 instances in total. For human evaluation, we randomly sampled 10% of the outlines and texts and assessed their scores. Final texts were evaluated by human experts, aligning model outputs with professional quality standards. Details of the dataset are in Appendix B.

### 4.2 Model

We conducted experiments using three models: GPT-3.5-turbo (Brown et al., 2020), GPT-4 (Achiam et al., 2023), and HyperCLOVA X (Yoo et al., 2024)<sup>4</sup>. For evaluation, we used GPT-4-turbo<sup>5</sup>. Additionally, we attempted to adapt the WritingPath approach to open-source models, including Llama2, Orion, and KoAlpaca. However, their outputs did not meet the quality standards necessary for fair comparison, and they were excluded from our analysis.

### 4.3 Human Evaluation

We conducted two separate human evaluation processes, involving a total of 12 carefully selected

<sup>4</sup>gpt-3.5-turbo, gpt-4-0125, HCX-003

<sup>5</sup>gpt-4-turbo; we chose GPT-4-turbo as the evaluation model because of its best performance at the time of this study.

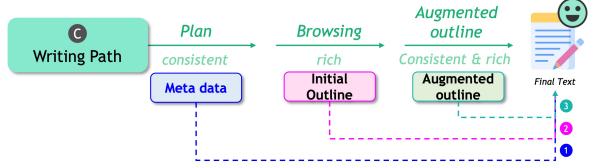


Figure 4: Overview of the main analysis steps in the WritingPath framework, covering meta-data only, initial outline, and augmented outline scenarios, respectively.

evaluators. For outline evaluations, which are relatively simple and short, we employed 6 native Korean speakers with experience in LLM. For the more detailed and rigorous writing evaluation, we recruited 6 professional writers and teachers as writing experts, each with over 10 years of expertise in Korean writing.

## 5 Experimental Results

### 5.1 Effectiveness of WritingPath

To verify that going through the WritingPath improves the final writing quality, we designed an analysis incorporating three cases (Figure 4): ① writing from metadata, ② writing from the initial outline, ③ writing from the augmented outline, where this final case corresponds to the complete WritingPath pipeline.

Figure 5 shows results from both (a) LLM and (b) human evaluation using CheckEval. Both consistently show progressive improvement as more components of the WritingPath are incorporated, while the model rankings are in different order between the two evaluation methods<sup>6</sup>. Specifically, The results show that using the augmented outline

<sup>6</sup>In the LLM evaluation, GPT-4 outperforms HyperCLOVA X, whereas the opposite trend is observed in human evaluations. These differences may be due to the use of GPT-4-turbo as the evaluation model and the self-enhancement bias discussed in Zheng et al. (2023).

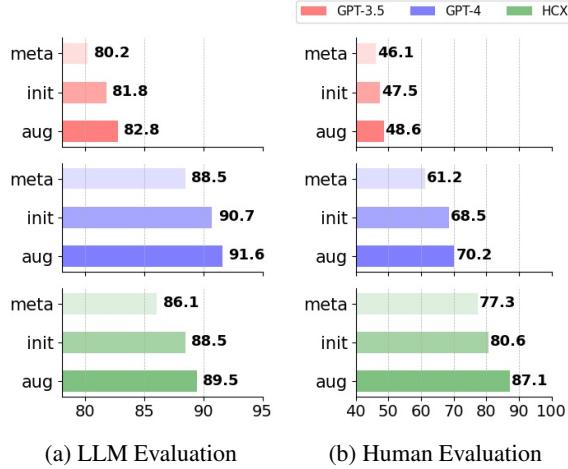


Figure 5: Main analysis steps on writing evaluation results by (a) LLM and (b) Human Evaluation.

(aug) leads to better writing quality compared to using only metadata (meta), indicating that the quality of writing improves significantly when the full WritingPath pipeline is employed. Furthermore, the augmented outline (aug) outperforms the initial outline (init), indicating that the content enrichment process further enhances writing quality. For a comprehensive analysis of writing quality, including human evaluation results for the final text across models, detailed improvement of text quality through the WritingPath, and Kendall tau correlations between various writing aspects and overall text quality, see Appendix D.

## 5.2 Outline Evaluation

Section 5.1 showed that using the augmented outline in the WritingPath pipeline led to better performance compared to using only the initial outline or metadata. To assess not only the impact of initial and augmented outlines on the quality of the final writing but also any differences in quality at the outline stage itself, we evaluated the initial and augmented outlines independently.

**Automatic Evaluation** To see the effects of the outline augmentation module, we conducted automatic evaluations on the initial and augmented outlines using criteria described in Section 3.1.1. The results in Table 1 show significant improvements in Coherence and Repetition aspects for the augmented outlines compared to the initial ones, indicating that the outline augmentation process enhances content consistency and reduces unnecessary repetition. Notably, although Diversity and Coherence are often considered trade-offs, the augmented outlines in our study maintained Diversity

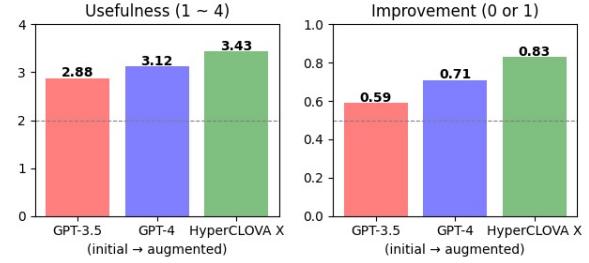


Figure 6: Evaluation of augmented outlines showing all models surpass the effectiveness threshold with scores in Usefulness above 2 and Improvement over 0.5, indicating universal enhancements from the initial outlines.

while improving Coherence. This suggests that the outline expansion module can increase consistency without compromising content diversity. Detailed performance across various domains is in Table 3.

**Human Evaluation** As described in Section 3.1.2, we conducted human evaluations to assess the cohesion, natural flow, diversity, and redundancy of initial and augmented outlines. The augmented outlines demonstrated significant improvements in all aspects except cohesion, which slightly declined or remained stable. Nevertheless, the overall performance of the augmented outlines surpassed that of the initial outlines. Further evaluations of the augmented outlines were conducted on usefulness and improvement, which indicated the extent of useful information added and overall quality enhancement compared to the initial outlines. As shown in Figure 6, all models demonstrated improvements in both metrics, validating the power of the browsing step. Detailed performance across various text domains is in Table 4.

## 6 Real-World Deployment

WritingPath was integrated into a commercial blogging platform as a writing assistance feature and tested for six months. In the service environment, additional considerations such as safety filtering and content quality control measures were necessary for reliable content generation. The system architecture of CLOVA for Writing by NAVER is depicted in Figure 7.

The serving pipeline integrates multiple components for reliable service operation. It integrates user request handling, content filtering, Kafka pipeline, and retrieval. Requests pass through a Gateway with rate limiting and are filtered for harmlessness. Specifically, the system includes emergency filtering and safety classification before pass-

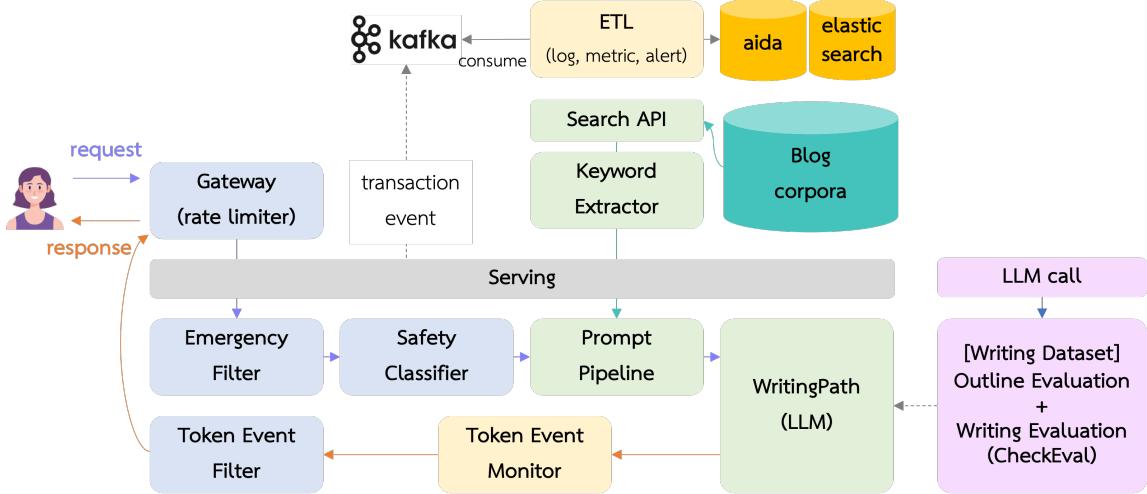


Figure 7: Real-world deployment pipeline of WritingPath.

ing requests to WritingPath. Additionally, a token event monitoring system tracks model usage, followed by token event filtering over output anomalies.

## 7 Conclusion

We introduced WritingPath, a framework that enhances the ability of LLMs to generate high-quality and goal-oriented writing by employing explicit outlines. Designed for real-world content creation, our approach uses structured guidelines from the early stages to ensure consistent quality control.

We verified the impact of WritingPath by conducting a comprehensive evaluation that incorporates automatic and human evaluations covering a wide range of aspects. Our experimental results demonstrate that texts generated following the full WritingPath approach, which includes the use of augmented outlines, exhibit superior performance compared to texts produced using only initial outlines or without any intermediate outlines. We also proposed a framework for assessing the WritingPath’s intermediate outlines, which found that augmented outlines have better inherent quality than initial outlines, demonstrating the importance of outline augmentation steps. We hope that this work will contribute to the research and development of more reliable AI-assisted writing solutions.

## 8 Acknowledgments

This research project was conducted as part of the NAVER HyperCLOVA-X and CLOVA for Writing projects. We express our gratitude to Nako Sung for his thoughtful advice on writer LLMs and to the

NAVER AX-SmartEditor team. Additionally, we would like to thank the NAVER Cloud Conversational Experience team for their practical assistance and valuable advice in creating the blog dataset. We appreciate Jaehee Kim, Hyowon Cho, Keonwoo Kim, Joonwon Jang, Hyojin Lee, Joonghoon Kim, Sangmin Lee, and Jaewon Cheon for their invaluable feedback and evaluation. We also thank DSBA NLP Group members for their comments on the paper.

## References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *ArXiv*, abs/2006.14799.
- Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. 2023. Art or artifice? large language models and the false promise of creativity. *arXiv preprint arXiv:2309.14556*.

- Yanran Chen and Steffen Eger. 2023. **MENLI**: Robust evaluation metrics from natural language inference. *Transactions of the Association for Computational Linguistics*, 11:804–825.
- Cyril Chhun, Pierre Colombo, Chloé Clavel, and Fabian M. Suchanek. 2022. Of human criteria and automatic metrics: A benchmark of the evaluation of story generation. In *International Conference on Computational Linguistics*.
- LA Clark. 8c watson, d.(1995). constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3):309–319.
- Allan Collins and Dedre Gentner. 1980. A framework for a cognitive theory of writing. In *Cognitive Processes in Writing*, pages 51–72. Erlbaum.
- Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.
- Filip Graliński, Anna Wróblewska, Tomasz Stanisławek, Kamil Grabowski, and Tomasz Górecki. 2019. **GEval**: Tool for debugging NLP datasets and models. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 254–262, Florence, Italy. Association for Computational Linguistics.
- JR Hayes. 1980. identifying the organization of writing processes. *Cognitive processes in writing: An interdisciplinary approach*, pages 3–10.
- Timothy R Hinkin. 1998. A brief tutorial on the development of measures for use in survey questionnaires. *Organizational research methods*, 1(1):104–121.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*, 9(8):1735–1780.
- Pei Ke, Hao Zhou, Yankai Lin, Peng Li, Jie Zhou, Xiaoyan Zhu, and Minlie Huang. 2022. CTRLEval: An unsupervised reference-free metric for evaluating controlled text generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2306–2319, Dublin, Ireland. Association for Computational Linguistics.
- Ute Knoch. 2011. Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, 16(2):81–96. Studies in Writing Assessment in New Zealand and Australia.
- Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.
- Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. *CHI '22*, New York, NY, USA. Association for Computing Machinery.
- Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladzhak, Frieda Rong, Rose E Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael S. Bernstein, and Percy Liang. 2023. Evaluating human-language model interaction. *Transactions on Machine Learning Research*.
- Yukyung Lee, Joonghoon Kim, Jaehoo Kim, Hyowon Cho, and Kang Pilsung. 2024. Checkeval: Robust evaluation framework using large language model via checklist. In *First Workshop on Human-Centered Evaluation and Auditing of Language Models*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.
- Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2023. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. *arXiv preprint arXiv:2311.09180*.
- Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages

- 186–195, Honolulu, Hawaii. Association for Computational Linguistics.
- Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. [A survey of evaluation metrics used for nlg systems](#). 55(2).
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36.
- Timo Schick, Jane A. Yu, Zhengbao Jiang, Fabio Petroni, Patrick Lewis, Gautier Izacard, Qingfei You, Christoforos Nalmpantis, Edouard Grave, and Sebastian Riedel. 2023. [PEER: A collaborative language model](#). In *The Eleventh International Conference on Learning Representations*.
- Guojin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2024. Kmmlu: Measuring massive multitask language understanding in korean. *arXiv preprint arXiv:2402.11548*.
- Keith Stevens, Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. [Exploring topic coherence over many models and many topics](#). In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island, Korea. Association for Computational Linguistics.
- Xiaofei Sun, Zijun Sun, Yuxian Meng, Jiwei Li, and Chun Fan. 2022. [Summarize, outline, and elaborate: Long-text generation via hierarchical supervision from extractive summaries](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6392–6402, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Hongyin Tang, Miao Li, and Beihong Jin. 2019. [A topic augmented text generation model: Joint learning of semantics and structural features](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5090–5099, Hong Kong, China. Association for Computational Linguistics.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. [Human evaluation of automatically generated text: Current trends and best practice guidelines](#). *Computer Speech & Language*, 67:101151.
- Chris van der Lee, Albert Gatt, Emiel van Miltenburg, Sander Wubben, and Emiel J. Krahmer. 2019. [Best practices for the human evaluation of automatically generated text](#). In *International Conference on Natural Language Generation*.
- Teun Adrianus Van Dijk. 1977. [Text and context: Explorations in the semantics and pragmatics of discourse](#).
- Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, et al. 2024. Weaver: Foundation models for creative writing. *arXiv preprint arXiv:2401.17268*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Sara Cushing Weigle. 2002. [Gptscore: Evaluate as you desire](#). *Cambridge University Press*.
- Edward W. Wolfe. 1997. [The relationship between essay reading style and scoring proficiency in a psychometric scoring system](#). *Assessing Writing*, 4(1):83–106.
- Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2023. [DOC: Improving long story coherence with detailed outline control](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3378–3465, Toronto, Canada. Association for Computational Linguistics.
- Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. 2022. [Re3: Generating longer stories with recursive reprompting and revision](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4393–4479, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, et al. 2024. Hyperclova x technical report. *arXiv preprint arXiv:2404.01954*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). *ArXiv*, abs/2306.05685.
- Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. 2023. [Recurrent-gpt: Interactive generation of \(arbitrarily\) long text](#). *Preprint*, arXiv:2305.13304.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texxygen: A benchmarking platform for text generation models. *SIGIR*.

## A Related Work

### A.1 Collaborative Writing with Language Models

Recent works that explore collaboration with LLMs during the writing process can be categorized into two aspects: 1) Outline Planning and Draft Generation, and 2) Recursive Re-prompting and Revision.

**Outline Planning and Draft Generation** involves incorporating the writer’s intents and contextual information into LLM prompts to create intermediate drafts. Dramatron (Mirowski et al., 2023) is a system for collaborative scriptwriting that automatically generates outlines with themes, characters, settings, flows, and dialogues. DOC (Yang et al., 2023) improves the coherence of generating long stories by offering detailed control of their outlines, including analyses of generated outlines and suggestions for revisions to maintain consistent plot and style.

Building on these works, our WritingPath mimics the human writing process by structuring it into controllable outlines. While our approach shares similarities with DOC in terms of utilizing outlines, we diverge from focusing solely on story generation and propose a novel outline generation process that incorporates external knowledge through browsing. Our aim is to sophisticatedly control machine-generated text across a wide range of writing tasks.

**Recursive Reprompting and Revision** technique extends the potential of LMs to assist not only with draft generation but also with editing and revision processes. This approach employs LLM prompt chains such as planning - drafting - reviewing - suggesting revisions in an iterative fashion to enhance the quality of written content. Re3 (Yang et al., 2022) introduces a framework for maintaining the long-range coherence of draft generation. It operates separate rewriter and edit modules in its prompt chain to check and refine plot relevance and long-term factual consistency. PEER (Schick et al., 2023) proposes a recursive revision framework based on the concept of self-training, where the model autonomously selects the editing operations for revision and provides explanations for the modifications it makes. RECURRENTGPT (Zhou et al., 2023) utilizes a recursive, language-based mechanism to simulate LSTM (Hochreiter and Schmidhuber, 1997), enabling the generation of coherent and extended texts. While these works are relevant to collaborative writing

with LMs, direct comparisons with our approach are unfeasible. These studies focus on specific tasks like story generation, requiring task-specific training and datasets, which are unavailable in Korean for our writing tasks.

Our WritingPath differs from previous works in its goals for utilizing LLMs in the writing process. Instead of relying on an ad-hoc recursive writing structure that may be inefficient, we establish a systematic writing plan that guides the generation process from the very beginning. Furthermore, we focus on free-form text generation rather than story generation and do not require separate training for writing, planning, or editing.

### A.2 Integrating External Information

Existing approaches have explored various methods to inject external knowledge into LLMs to improve their performance on text-generation tasks. For instance, Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), and Toolformer (Schick et al., 2024) have developed techniques to connect LLMs with external search tools, enabling them to gather relevant information and generate more informative and accurate responses. However, despite these contributions to improving LLMs’ access to information (Asai et al., 2024), they inherently fall short of fully reflecting the diversity and complexity of the writing process (Chakrabarty et al., 2023).

Our work distinguishes itself from previous approaches by focusing on emulating the modern writing planning process. With this structured approach, an LLM can efficiently produce high-quality text, significantly contributing to improving the control and quality of the generated text.

### A.3 Writing Evaluation

It is well-known that supervised metrics such as ROUGE and BLEU are ill-suited for evaluating natural language generation output, especially for open-ended writing tasks. Traditionally, such evaluation has depended on rubric-based human evaluation, which is a costly and time-consuming task (Weigle, 2002). Recent advancements in LLMs have led to the exploration of new paradigms that utilize LMs for evaluating LM-generated text (Gralínski et al., 2019; Fu et al., 2023). However, to effectively assess free-form text, a more customized and interactive evaluation framework is needed.

We utilize CheckEval (Lee et al., 2024), a fine-grained and explainable evaluation framework, to assess free-form text writing. By customizing a

checklist with specific sub-questions for each writing aspect, we provide a more reliable and accurate means of evaluating writing quality.

## B Details of Dataset

We selected 20 blog posts for each domain<sup>7</sup>, resulting in 100 seed data points. For seed data construction, we generated metadata, including purpose, topic, keywords, and expected reader, based on the title and content of the blog posts. This metadata is the input to the WritingPath, helping the model understand the context of the post and generate relevant outlines and text. We created a test dataset of 1,100 instances per model under evaluation using the seed data. Each data point includes the outputs of each WritingPath step: an outline, additional information, an augmented outline, and the final text. With analysis experiments as well, we generated a total of 1,500 posts for each model, resulting in 4,500 instances in total. For human evaluation, we randomly sampled 10% of the outlines and texts and assessed their scores. The final texts were evaluated by human experts, and the dataset aligns the generated outputs from three models with the human scores.

## C Details of Evaluation

### C.1 Compensation Details

Outline evaluators were compensated with a 6,000 KRW ( $\approx$  4.2 USD) gift card for their 30-minute participation. And writing experts were compensated at a rate of 9,000 KRW ( $\approx$  6.6 USD) per one-writing sample.

### C.2 Automatic Evaluation - Outline

- Logical alignment: Based on [Chen and Eger \(2023\)](#), we utilize Natural Language Inference (NLI) which examines whether the headers and subheaders within an outline logically connect, ensuring the structural integrity necessary for coherent argumentation<sup>8</sup>.
- Coherence: Through Topic Coherency metrics such as NPMI ([Stevens et al., 2012](#)) and UCI ([Lau et al., 2014](#)), this aspect assesses the thematic uniformity across the sections of outline, verifying a consistent narrative.
- Diversity: We measure the breadth of topics addressed by applying Topic Diversity metrics

([Dieng et al., 2020](#)), aiming to ensure that the content of outline is comprehensive and varied.

- Repetition: Self-BLEU ([Zhu et al., 2018](#)) is used to gauge the degree of redundancy within the outline, prioritizing efficiency in information presentation by minimizing repetition.

### C.3 Human Evaluation - Outline

The human evaluation criteria are based on aspects considered in previous studies on text coherence, relevance, and quality assessment ([Yang et al., 2022, 2023](#); [Zhou et al., 2023](#); [Ke et al., 2022](#)). For both initial and augmented outlines, the human evaluation is performed on the following five aspects, using a 1-4 point scale:

- Cohesion: Evaluates whether the title and outline are semantically consistent.
- Natural Flow: Assesses whether the outline flows in a natural order.
- Diversity: Evaluates whether the outline consists of diverse topics.
- Redundancy: Assesses whether the outline avoids semantically redundant content.

Furthermore, we use two additional aspects for evaluating the augmented outline:

- Usefulness of Information: Assesses whether the augmented outline provides useful information beyond the initial outline.
- Improvement: Evaluates whether significant improvements have been made in the augmented outline compared to the initial outline, using a binary scale.

<sup>7</sup><https://blog.naver.com/>

<sup>8</sup>we utilize gpt-4-turbo for NLI evaluation

Model	Linguistic Fl. binary	Logical Fl. binary	Coh. binary	Cons. binary	Comple. binary	Spec. binary	Int. binary	Overall binary
GPT-3.5	51.66	31.14	46.29	88.11	66.43	21.14	35.14	48.56
GPT-4	68.00	60.57	72.86	89.26	80.29	54.14	66.29	70.20
HyperCLOVA X	<b>89.71</b>	<b>84.46</b>	<b>91.14</b>	<b>98.06</b>	<b>92.57</b>	<b>74.00</b>	<b>80.00</b>	<b>87.13</b>

Table 2: Human evaluation results for writing quality of final text (aug) across models.

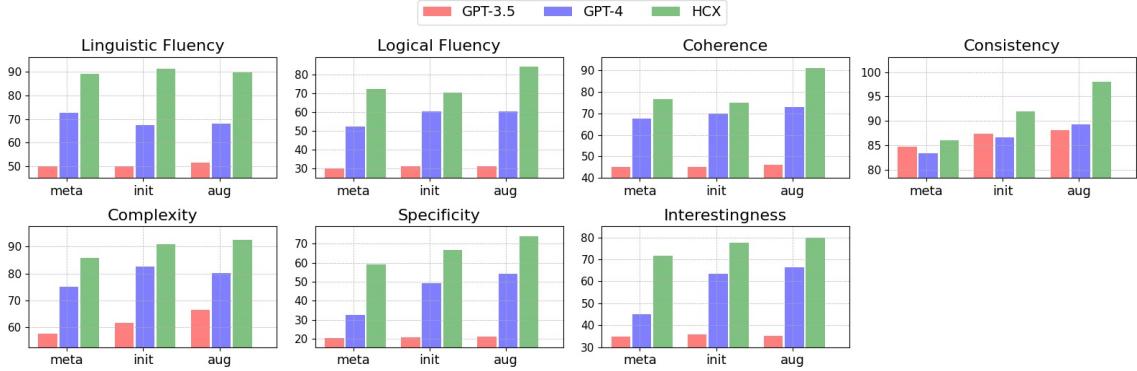


Figure 8: Human evaluation results for writing quality (meta, init, aug) over various CheckEval aspects.

## D Further Analysis of Writing Quality

To further analyze the quality of the text generated through the complete WritingPath pipeline, we conducted a human evaluation based on the CheckEval framework. The results are presented in Table 2. The analysis by six writing experts showed that GPT-4 and HyperCLOVA X generally performed better than GPT-3.5 in terms of writing quality. HyperCLOVA X exhibited higher scores in specificity compared to other models, which is consistent with the findings reported in KMMLU (Son et al., 2024) regarding the advantages of language-specific models. Detailed performance metrics across various domains and further LLM evaluations can be found in Table 5, 6. Furthermore, We consider seven key aspects (Section 3) for evaluating the quality of writing. CheckEval’s binary responses for each aspect allow for identifying the specific factors contributing to the assessments. We found that logical fluency, coherence, consistency, and specificity significantly contribute to the improvement of text quality through the WritingPath (Figure 8).

During the evaluation of the writing quality, writing experts assigned binary overall quality ratings (1 for high quality, 0 for low quality) to the texts. We employed the Kendall tau correlation to examine the relationship between the overall binary ratings and the scores for each evaluation aspect. The analysis (Figure 9) revealed a significant corre-

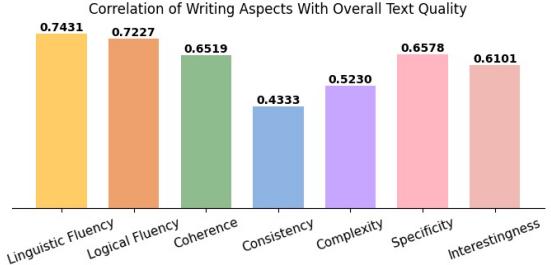


Figure 9: Kendall tau correlations between various writing aspects and overall text quality.

lation for all the aspects we designed. Interestingly, logical fluency, specificity, and coherence, which were found to be particularly important in determining the perceived quality of written content, are among the aspects that showed the most significant improvement through the WritingPath (Figure 8).

The progressive improvement in these aspects can be attributed to the effectiveness of using outlines. The initial outline (init) helps organize information more logically and coherently compared to using only metadata (meta), while the augmented outline (aug) further enhances the consistency and richness of the content. These findings highlight the importance of using outlines in the writing process and demonstrate how their gradual enhancement leads to better-structured, more coherent, and content-rich texts, ultimately improving the overall quality of the written output.

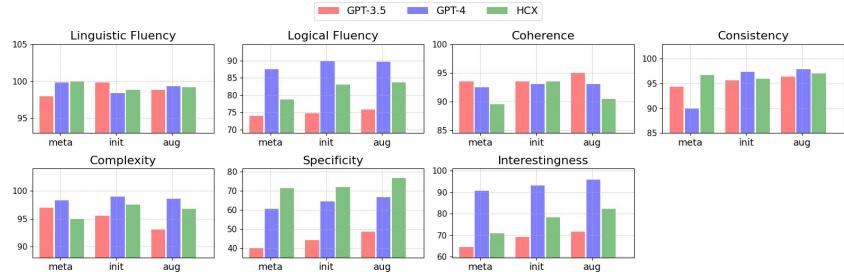


Figure 10: LLM Evaluation

Model Metrics	Category	Outline Type	Logical Alignment	Coherence	Diversity	Repetition	
			NLI ( $\uparrow$ )	UCI ( $\uparrow$ ) / NPMI ( $\uparrow$ )	Topic Diversity ( $\uparrow$ )	Self-BLEU ( $\downarrow$ )	
Eval Level			Header-Subheader	Outline	Outline	Outline	
GPT 3.5	Beauty	Initial	-	0.638 / 0.298	0.488	48.01	
		Augmented	0.483	<b>1.506 / 0.553</b>	<b>0.513</b>	<b>23.79</b>	
	Travel	Initial	-	0.835 / 0.454	<b>0.708</b>	21.56	
		Augmented	0.575	<b>1.646 / 0.540</b>	0.670	<b>13.21</b>	
	Gardening	Initial	-	0.496 / 0.206	0.591	26.52	
		Augmented	0.658	<b>1.291 / 0.575</b>	<b>0.592</b>	<b>19.24</b>	
GPT 4	Cooking	Initial	-	0.543 / 0.352	0.641	15.94	
		Augmented	0.686	<b>1.003 / 0.411</b>	<b>0.712</b>	<b>13.71</b>	
	IT	Initial	-	0.491 / 0.235	0.523	25.67	
		Augmented	0.667	<b>1.180 / 0.463</b>	<b>0.560</b>	<b>16.69</b>	
	HyperCLOVA X	Beauty	Initial	-	0.908 / <b>0.574</b>	0.657	36.50
		Augmented	0.577	<b>1.854 / 0.573</b>	<b>0.658</b>	<b>18.80</b>	
		Travel	Initial	0.717 / <b>0.534</b>	<b>0.691</b>	17.61	
		Augmented	0.615	<b>1.690 / 0.530</b>	0.688	<b>10.63</b>	
		Gardening	Initial	0.833 / 0.398	0.676	20.33	
		Augmented	0.724	<b>1.559 / 0.555</b>	<b>0.681</b>	<b>13.43</b>	
HyperCLOVA X	Cooking	Initial	-	0.693 / <b>0.468</b>	0.720	13.85	
		Augmented	0.701	<b>1.512 / 0.464</b>	<b>0.745</b>	<b>10.98</b>	
	IT	Initial	-	0.854 / 0.454	0.625	16.80	
		Augmented	0.702	<b>1.448 / 0.471</b>	<b>0.633</b>	<b>11.77</b>	

Table 3: Detailed outline automatic evaluation results.

Model	Category	Outline Type	Cohesion	Natural Flow	Diversity	Redundancy	Usefulness	Improvement
GPT 3.5	Beauty	Initial	<b>3.542</b>	<b>2.958</b>	2.833	2.917	-	-
		Augmented	3.208	2.875	<b>3.417</b>	<b>3.375</b>	3.000	0.542
		Initial	3.625	2.833	3.167	2.917	-	-
	Travel	Augmented	<b>3.708</b>	<b>3.125</b>	<b>3.750</b>	<b>3.542</b>	3.458	0.708
		Initial	2.958	2.375	<b>2.542</b>	2.417	-	-
		Augmented	<b>3.042</b>	<b>2.833</b>	3.375	<b>2.667</b>	2.542	0.708
GPT 4	Cooking	Initial	<b>3.292</b>	<b>2.417</b>	2.417	2.583	-	-
		Augmented	2.708	2.333	<b>3.458</b>	<b>2.833</b>	2.542	0.458
		Initial	<b>3.458</b>	<b>2.917</b>	2.875	2.792	-	-
	IT	Initial	<b>3.000</b>	<b>2.917</b>	3.250	3.042	-	-
		Augmented	3.083	2.750	<b>3.708</b>	<b>3.250</b>	2.875	0.542
		Augmented	<b>3.000</b>	2.792	<b>3.833</b>	<b>3.583</b>	3.042	0.750
HyperCLOVA X	Beauty	Initial	3.375	2.750	3.083	3.167	-	-
		Augmented	<b>3.542</b>	<b>3.208</b>	<b>3.708</b>	<b>3.583</b>	3.208	0.833
		Initial	3.542	2.792	3.042	2.833	-	-
	Travel	Augmented	<b>3.625</b>	<b>3.083</b>	<b>3.792</b>	<b>3.542</b>	3.333	0.750
		Initial	<b>3.792</b>	3.125	3.083	2.917	-	-
		Augmented	3.625	<b>3.208</b>	<b>3.833</b>	<b>3.500</b>	3.208	0.667
	Gardening	Initial	<b>3.292</b>	<b>2.708</b>	2.833	2.333	-	-
		Augmented	3.208	2.625	<b>3.542</b>	<b>2.917</b>	2.833	0.542
		Initial	<b>3.000</b>	<b>2.917</b>	3.250	3.042	-	-
	Cooking	Augmented	3.025	3.375	<b>3.792</b>	<b>3.792</b>	3.250	0.750
		Initial	<b>3.500</b>	2.958	2.750	3.250	-	-
		Augmented	<b>3.000</b>	2.625	<b>3.208</b>	<b>3.917</b>	3.083	0.750
	IT	Initial	<b>3.292</b>	2.625	2.833	3.250	-	-
		Augmented	3.042	<b>3.208</b>	<b>3.917</b>	<b>3.708</b>	3.083	0.750

Table 4: Detailed outline human evaluation results.

Model	Category	Linguistic Fl.	Logical Fl.	Coh.	Cons.	Comple.	Spec.	Int.	Overall
<i>GPT 3.5</i>	Beauty	96.00	75.42	90.00	94.17	89.58	52.50	63.06	80.10
	Travel	100.00	78.54	97.08	96.94	97.08	65.83	81.67	88.16
	Gardening	98.67	75.63	95.00	96.11	89.17	49.17	78.89	83.23
	Cooking	99.50	72.92	95.42	97.78	95.00	35.83	84.17	82.94
	IT	100.00	76.67	97.50	96.94	94.17	39.17	50.83	79.33
	Total	98.83	75.83	95.00	96.39	93.00	48.50	71.72	82.75
<i>GPT 4</i>	Beauty	99.17	89.79	97.50	99.44	99.17	84.58	98.61	95.47
	Travel	99.00	90.21	91.67	97.22	96.67	70.00	96.94	91.67
	Gardening	99.67	90.00	94.17	98.33	100.00	74.17	97.78	93.44
	Cooking	99.67	89.58	93.75	98.61	97.92	63.33	96.67	91.36
	IT	99.17	88.96	88.33	96.11	99.17	41.67	76.11	84.22
	Total	<b>99.33</b>	<b>89.71</b>	<b>93.08</b>	<b>97.94</b>	<b>98.58</b>	66.75	<b>93.22</b>	<b>91.23</b>
<i>HyperCLOVA X</i>	Beauty	100.00	88.33	98.33	100.00	90.00	90.42	91.38	94.07
	Travel	99.50	83.75	90.42	97.50	91.25	79.58	92.77	90.68
	Gardening	99.33	88.13	93.33	98.61	95.00	70.42	84.16	89.85
	Cooking	98.67	82.29	88.75	96.67	90.42	87.92	91.11	90.83
	IT	98.50	76.04	81.25	92.50	87.08	55.00	52.50	77.55
	Total	99.20	83.71	90.42	97.06	90.75	<b>76.67</b>	82.38	88.60

Table 5: Detailed writing LLM evaluation results.

Model	Category	Linguistic Fl.	Logical Fl.	Coh.	Cons.	Comple.	Spec.	Int.	Overall
<i>GPT 3.5</i>	Beauty	51.43	30.29	51.43	92.57	67.14	14.29	29.29	48.06
	Travel	68.00	56.00	68.57	87.43	82.14	47.14	55.71	66.43
	Gardening	45.14	30.86	44.29	87.43	52.86	18.57	30.00	44.16
	Cooking	46.29	8.86	21.43	80.00	72.86	7.14	32.86	38.49
	IT	47.43	29.71	45.71	93.14	57.14	18.57	27.86	45.65
	Total	51.66	31.14	46.29	88.11	66.43	21.14	35.14	48.56
<i>GPT 4</i>	Beauty	72.00	67.43	82.86	93.71	86.43	56.43	72.14	75.86
	Travel	76.00	71.43	82.86	89.71	86.43	67.14	79.29	78.98
	Gardening	66.29	61.14	75.71	85.71	70.71	54.29	61.43	67.90
	Cooking	63.43	50.86	61.43	90.86	82.14	47.14	62.14	65.43
	IT	62.29	52.00	61.43	86.29	75.71	45.71	56.43	62.84
	Total	68.00	60.57	72.86	89.26	80.29	54.14	66.29	70.20
<i>HyperCLOVA X</i>	Beauty	92.00	87.43	91.43	99.43	92.14	69.29	81.43	87.59
	Travel	95.43	91.43	95.71	99.43	100.00	84.29	85.00	93.04
	Gardening	88.57	85.71	97.14	99.43	95.71	75.71	82.86	89.31
	Cooking	90.86	85.71	90.00	98.29	96.43	81.43	82.14	89.27
	IT	81.71	72.00	81.43	93.71	78.57	59.29	68.57	76.47
	Total	<b>89.71</b>	<b>84.46</b>	<b>91.14</b>	<b>98.06</b>	<b>92.57</b>	<b>74.00</b>	<b>80.00</b>	<b>87.13</b>

Table 6: Detailed writing human evaluation results.

Aspect	Subaspect	Descriptions
Linguistic Fluency	Natural Expression	Does the given text read naturally without any unnatural rhythm or excessively emphasized parts? 주어진 글이 부자연스러운 리듬이나 과도하게 강조된 부분 없이 자연스럽게 읽히나요?
	Text Length	Is the length of the text suitable for the purpose and is it not excessively verbose or overly concise? 텍스트의 길이가 목적으로 적합하며 과도하게 장황하거나 지나치게 간결하지는 않은 글인가요?
	Vocabulary	Is the vocabulary appropriate for the context, not overly complex, and suitable for the topic and reader? 어휘가 맥락에 맞지 않거나 지나치게 복잡하지 않고, 주제와 독자에 적합한가요?
	Syntax	Is the composition and sentence structure of the given text correct? 주어진 글의 구성과 문장의 구조가 올바른가요?
	Mechanic-Spelling, Punctuation	Is the spelling and punctuation of the given text correctly applied? 주어진 글의 철자와 문장부호가 올바르게 적용되었나요?
Logical Fluency	Organization (layout)	Does the given text have a clear and effective structure (layout)? 주어진 글은 명확하고 효과적인 구조 (레이아웃)를 가지고 있나요?
	Repetitive Content	Is the text free of repetitive or unnecessary content? 텍스트 내에서 반복되는 내용이나 불필요한 내용이 없는 글인가요?
	Inter-sentence Cohesion	In the text, are the sentences well connected and progressing naturally and logically? 글 내에서 문장들이 잘 연결되어 있어 자연스럽고 논리적으로 진행이 되나요? Did you use conjunctions appropriately to improve readability? 가독성을 높이기 위한 접속사를 적절하게 사용했나요?
	Inter-paragraph Cohesion	Are the paragraphs in the text logically connected and progressing with each other? 텍스트 내의 단락들이 논리적으로 연결되어 서로 진행되나요?
Coherence	Topic Consistency	Is the entire article consistently progressing with the central theme as the focus? 전체 글이 중심 주제를 중심으로 일관되게 진행되나요?
	Topic Sentence and Paragraph	Does each paragraph of the article have a clear subtopic centered around the main idea? 글의 각 문단이 주요 아이디어를 중심으로 명확한 소주제를 가지고 있나요?
Consistency	Tone	Is a consistent narrative tone and style maintained throughout the entire text? 텍스트 전체에서 일관된 서술 어조와 어투가 유지되나요? Is there no sudden change in tone in the context of the writing? 글의 맥락에서 급격한 어조 변화가 없는 글인가요?
	Stance/Posture	Does the author present a consistent opinion on the topic in the writing? (Should not present conflicting opinions on the same subject) 저자는 글에서 주제에 대한 일관된 의견을 제시하나요? (동일한 대상에 대한 상반된 의견을 제시하지 않아야 함)
	Style	Does the given text maintain a consistent style type (spoken language, written language, informal, formal, etc.)? 주어진 글이 일관된 스타일의 유형 (구어체, 문어체, 반말, 존댓말 등의 유형)을 유지하나요? Do you consistently use abbreviations and acronyms when necessary? 필요 시 약어와 머리글자가 일관되게 사용되나요?
Complexity	Vocabulary	Is it a clear text that does not excessively use uncommon or complex words? 일반적이지 않거나 복잡한 단어들이 과도하게 등장하지 않는 명료한 글인가요? Is the definition of unfamiliar and difficult words provided and are they used appropriately in context? 낯설고 어려운 단어에 대한 정의가 되어 있고 문맥에 맞게 잘 사용되었다나요?
	Syntax	Is the given text clearly structured without excessively complex sentence structures? 주어진 글이 과도하게 복잡한 문장 구조를 가진 문장을 없이 명확하게 구성되어 있나요? Do the first sentences of each paragraph start differently? (Asking if the text has paragraphs that do not all start the same way) 각 문단의 첫 문장이 다양하게 시작되나요? (각 문단의 시작이 모두 동일하지 않은 글인지 질문)
Specificity	Use of Examples and Review	Is the example appropriately connected to the topic of the article? 예시가 글의 주제와 적절하게 연결되어 있나요? Was the author's personal experience mentioned specifically? 작성자의 개인적인 경험에 구체적으로 언급되었다나요?
	Detailed Descriptions	In the writing, were specific numerical values such as ratios and quantities mentioned? 글에서 구체적으로 비율, 수량과 같은 수치들이 언급되었나요? When introducing details in a writing, do you appropriately utilize context or background information? 글에서 세부 사항을 소개할 때 맥락이나 배경 정보를 적절하게 활용하나요?
Interestingness	Engagement	Was the blog post written based on an appealing storytelling approach? (It's okay if an exaggerated tone is included) 블로그 글이 매력적인 스토리텔링 접근 방식을 기반으로 작성되었나요? (과장된 어조가 포함되어도 괜찮음)
	Kindness	Was the written blog post written in a friendly tone for the readers? 작성된 블로그 글은 독자들에게 친근한 어조로 작성되었나요?
	Originality	Does the written blog post include the author's unique ideas or perspectives? 작성된 블로그 글에는 작성자의 독특한 아이디어나 관점이 포함되어 있나요? Does the writer's personal experience add freshness to the writing? 작성자의 개인적인 경험에 글에 신선함을 더하나요?

Table 7: Evaluation principles.

## Writing Evaluation Prompt

You will be given one text written for a blog post.  
Your task is to rate the written text on one metric.  
Please read and understand these instructions carefully.  
Keep this document open while reviewing and refer to it as needed. You are a writing expert! it is crucial to apply a robust evaluation.

## Evaluation Criteria:

{aspect} - {definition}

### Guidelines###

1. Read these guidelines completely.
2. Read the Written Text attentively.
3. Comprehend the questions and the meaning of the {aspect}.
4. Answer each question with 'yes' or 'no', without any explanations.
5. Use the prescribed answer format.

### Output Format###

Q: [Question] A: [Answer]

Q: [Question] A: [Answer]

...

### Questions###

Q. {question}

Blog text: {writing}

Your Answers:

Figure 11: Writing Evaluation Prompt for Checklist-based Assessment.

## WritingPath Prompt

### Prompt for Metadata construction (step #1):

We aim to systematically organize blog posts by dividing them into four categories:

1. the purpose of the post
2. the type of post
3. the style of the post
4. keywords.

An example of the expected format is provided below.

{examples}

Similar to the example provided, please categorize the blog post below in detail according to

1. purpose, 2. type, 3. style, and 4. keywords, where keywords are composed of words.

==Blog post==

{original blog text}

### Prompt for Generation of Title and Initial Outline (step #2):

Based on the metadata, I plan to create the title and a simple table of contents for the article.

Below is an example of the desired format.

{example}

Following the example above, based on the post information provided below, only create "==Title==" and a brief "==Initial Outline==".

Do not generate an excessively long table of contents.

The table of contents should not be a simple list;

do not write it in paragraph form. Do not create subheadings.

Only the title and table of contents should be generated.

The table of contents must be numbered in sequence.

You must strictly follow the format for the title and table of contents below.

==Meta data==

{meta data}

Figure 12: WritingPath Prompt for Each Stage (Step 1 and 2).

## WritingPath Prompt

### Prompt for Generation of Augmented Outline (step #4):

Map the necessary additional information below to create an augmented outline.  
Here is an example.

{example}

Following the method above, create an ==Augmented Outline==.  
Specifically, incorporate new information as subheadings under the existing headings,  
ensuring that each heading and its subheadings are themed consistently.

==Additional Information==  
{additional information from browsing}  
==Initial Outline==  
{initial outline}

### Prompt for Generation of Text (step #5):

Based on the title and current table of contents below,  
I plan to write the i + 1th paragraph suitable for a blog post.  
Writing should naturally follow the flow of the post information and the augmented outline.  
Write in a friendly and attractive tone like bloggers, making it interesting for the reader.  
The written content should be engaging and captivating for the reader.

==Augmented Outline==  
{augmented outline}  
==Meta Data==  
{meta data}

Below are the title and current table of contents for writing the blog post.  
==Title==  
{title}  
==Current Outline==  
{current section}

Figure 13: WritingPath Prompt for Each Stage (Step 4 and 5).