Automated Claim–Evidence Extraction for Political Discourse Analysis: A Large Language Model Approach to *Rodong Sinmun* Editorials

Gyuri Choi, Hansaem Kim*
Interdisciplinary Graduate Program of Linguistics and Informatics, Yonsei University
{gyuri1345, khss}@yonsei.ac.kr

Abstract

This study investigates the feasibility of automating political discourse analysis using large language models (LLMs), with a focus on 87 editorials from *Rodong Sinmun*, North Korea's official newspaper. We introduce a structured analytical framework that integrates Chain-of-Thought prompting for claim–evidence extraction and a GPT-4o–based automated evaluation system (G-Eval). Experimental results demonstrate that LLMs possess emerging discourse-level reasoning capabilities, showing notably improved alignment with expert analyses under one-shot prompting conditions. However, the models often reproduced ideological rhetoric uncritically or generated interpretive hallucinations, highlighting the risks of fully automated analysis. To address these issues, we propose a Hybrid Human-in-the-Loop evaluation framework that combines expert judgment with automated scoring. This study presents a novel approach to analyzing politically sensitive texts and offers empirical insights into the quantitative assessment of ideological discourse, underscoring the scalability and potential of automation-driven methodologies.

1 Introduction

Editorials in *Rodong Sinmun*, North Korea's official newspaper, function both as journalistic reports and instruments of political discourse that aid in the internalization and justification of the state's ideology and policies. As noted in Baek(2023), these editorials are structured more around persuasion than formal argumentation, with topics and supporting grounds intricately intertwined. Consequently, the discursive structure of these texts is difficult to discern without familiarity with the North Korean language and culture. Traditional rule-based information extraction and keyword-centric analytical approaches have shown limitations in capturing the indirect and ideologically laden nature of these texts, leading to a predominance of qualitative analyses. In South Korea, several studies (Lee, 1997; Kim, 2003; Jin, 2013; Kim and Cho, 2022) have examined *Rodong Sinmun* to investigate shifts in North Korean political policy, but these have also relied primarily on qualitative methods. This study presents the first framework designed to automatically extract and evaluate the core claims and supporting arguments of North Korean editorials by leveraging the step-by-step reasoning capabilities of large language models and automated evaluation metrics for text generation quality. By comparing the automated output with human analyses, we assess both the potential and interpretive risks of automation and propose a structure-based analytical method that extends beyond traditional qualitative approaches.

2 Related Work

LLMs have shown strong performance across various NLP tasks, including information extraction and automatic evaluation. However, their application to complex texts involving rhetoric, emotion, and ideology—such as political discourse—remains limited. In information extraction, Liu et al. (2024) used few-shot prompting to identify executable directives in emergency planning documents, and Xu et al. (2024) proposed ChatUIE, a unified framework for named entity recognition, relation extraction, and event extraction, addressing task imbalance and consistency. Yet, these methods are primarily designed for technical or practical texts, with limited applicability to political content. The "LLM-as-a-judge" paradigm has emerged as an alternative to human evaluation. For instance, Afzal et al. (2024) introduced G-Eval to assess responses from a RAG-based HR chatbot,

Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER), pages 1–17

showing high alignment with human judgments. Le Mens and Gallego (2025) found that LLMs can infer ideological positions in political texts with expert-level consistency. However, Stureborg et al. (2024) highlighted that evaluation results can vary significantly depending on prompt design, criteria, and temperature settings, indicating the need for more robust and systematic evaluation protocols.

Concerns over LLMs' political bias have also been raised. Yang et al. (2024) demonstrated that responses vary by model origin, size, and training time, while Kronlund-Drouault (2024) argued that model alignment may reflect dominant capitalist ideologies.

In sum, although progress has been made in information extraction, evaluation automation, and bias detection, integrated approaches for inferring and evaluating claim–evidence structures in political discourse remain underexplored. While Gao and Feng (2025) attempted stance analysis in journalistic texts, this study extends such methods to political texts, experimentally examining the combination of structural inference and automatic evaluation as a means of analyzing political rhetoric as a mode of governance.

3 System Architecture & Methodology

This section outlines the architecture and implementation of our LLM-based system for analyzing political discourse. We focus on 87 editorials from *Rodong Sinmun*, the official newspaper of North Korea, aiming to automatically extract political messages and quantitatively evaluate them based on criteria such as coherence, factuality, and relevance.

Based on pilot experiments regarding North Korea–related content and potential bias, GPT-4o demonstrated relatively stable performance in both understanding and restrained expression. As a result, all automated evaluations in this study were conducted using GPT-4o.

The evaluation consists of three main components:

(1) Logical coherence between the identified claim and its supporting evidence.

(2) Response quality, which include accuracy, relevance, and logical consistency.

(3) Hallucination rate, which detects external insertions or excessive rhetorical language beyond the source text.

Following the automated assessment, a subset of outputs was manually reviewed by human evaluators to verify the validity of the evaluations and identify common error types.

The overall system pipeline is summarized in Figure 1.

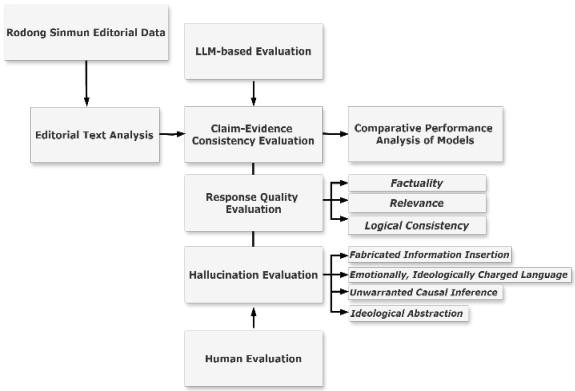

Figure 1: Structured Evaluation Pipeline for Analyzing Claim–Evidence Relations in Political Editorials. The pipeline systematically evaluates extracted claim–evidence pairs across coherence, response quality, and hallucination dimensions.

3.1 Prompt Design and Response Structure

To address the repetitive narrative structure and symbolic rhetoric typical of *Rodong Sinmun* editorials, we designed prompts that explicitly elicit chain-of-thought (CoT) reasoning. Inspired by Baek (2023), our CoT-based analysis prompt guides the model through three sequential steps:

(1) Identifying the main argument.

(2) Citing supporting evidence.

(3) Interpreting the underlying political objective.

This structure is intended to help the model uncover the implicit logic embedded in ideologically charged discourse.

> You are an expert analyst specializing in political propaganda and editorial discourse. The following editorial text covers multiple domains—including politics, economics, and society—and embeds a specific political goal or ideological message.
>
> Please analyze the editorial by following the Chain-of-Thought (CoT) procedure below to uncover its political strategy and messaging intent:
>
> Step 1. Specify the editorial's metadata, including the title, publication date, and source.

Step 2. Read the entire editorial carefully and identify the central argument that is emphasized repeatedly or serves as the core theme. Step 3. Provide at least one direct quotation from the editorial that supports the identified argument. Step 4. Interpret the political objective or strategic message the editorial seeks to convey. Step 5. Synthesize your analysis in a clear and concise summary using the format below:

Table 1: Example of Chain-of-Thought Prompt Design for Political Editorial Analysis.

The model is instructed to produce responses in a fixed format, as shown below Table 2.

◆ Summary of Key Themes • Evidence [Provide justification based on a direct quote from the editorial or the editorial's title.] • Interpretation of Political Strategy or Message [Analyze and explain the intended political objective or strategic messaging embedded in the editorial.]

Table 2: Prompt Design for Output Format Standardization in Editorial Analysis.

To elicit structured reasoning from the model, the response format was designed to guide it beyond simple summarization by prompting it to autonomously construct a logical connection between the central claim and supporting evidence within the editorial. This allowed the model to demonstrate discourse-level analytical capabilities.

The experiment was designed to evaluate the language models' capacity for structured inference, using two prompting conditions: zero-shot and one-shot. The zero-shot setting assessed the model's ability to autonomously infer rhetorical structure without guidance, while the one-shot condition tested whether the model could reproduce a structured response based on a single expert-provided example. Only one demonstration was used in the one-shot setting; variations across examples or prompt sensitivity were not examined within the scope of this study. The number of shots was deliberately restricted to maintain experimental control, and multi-shot prompts were excluded due to potential risks of overfitting.

3.2 G-Eval–Based Automated Evaluation Framework

Model outputs were quantitatively assessed using the G-Eval framework. We adopted three evaluation dimensions tailored to claim–evidence extraction tasks: (1) Coherence, (2) Response Quality, and (3) Hallucination. All scores were generated automatically on a scale from 1-5 using GPT-4o.

The Coherence dimension evaluates the logical connection and inferential validity between the model-generated claim and the cited evidence from the original text. Rather than surface-level similarity, this metric focuses on semantic reasoning, aligning with prior research emphasizing structure-aware evaluation in discourse tasks (Yin and Roth, 2018).

The Response Quality dimension can be further subdivided into accuracy, relevance, and logical consistency. These subdimensions measure factual alignment with the source, reflection of the central theme, and the internal coherence of the generated response, respectively. This multidimensional approach mitigates the limitations of single-metric evaluation (Zhong et al., 2022).

The Hallucination dimension identifies instances where the model introduces unsupported content, rhetorical exaggerations, or ideologically skewed interpretations. This is particularly critical in the context of political discourse, as highlighted in recent work on hallucination in NLG systems (Ji et al., 2023). We assessed four types of hallucination: factual insertion, emotional overstatement, logical leap, and thematic generalization.

Detailed scoring guidelines for each category are provided in Appendix A.

4 Dataset & Models

This study analyzes the complete set of *Rodong Sinmun* editorials published in 2021, which consist of 87 articles in total. All texts were collected in their original Korean form. The editorials span a variety of topics, including politics, diplomacy, and economics. Each editorial was individually processed by the model for discourse-level analysis.

Four LLMs were selected for comparison: GPT-o3 Mini (OpenAI, 2024), Claude 3.7(Anthropic, 2024), Gemini 2.0 (Google, 2025), and EXAONE 7.8B (Research LG et al., 2025). The selection was based on general-purpose capability and adaptation to the Korean language. In particular, EXAONE 7.8B, although smaller in scale than the other

models, was included under the hypothesis that its linguistic alignment would offer advantages in interpreting the unique rhetorical patterns and ideological expressions found in *Rodong Sinmun*.

This setup aims to explore how differences in language adaptation affect the models' performance in political discourse analysis. Detailed model characteristics and configurations are summarized in Table 3.

Model	Key Characteristics
GPT-o3 Mini	Lightweight and fast; responsive to structured reasoning and chain-of-thought prompts.
Claude 3.7-Sonnet	Emphasizes consistency over complex reasoning; supports reflective and pragmatic generation.
Gemini 2.0-Flash	Optimized for long form processing and summarization; handles large context windows and multimodal input.
EXAONE 7.8B	Korean-specialized; includes CoT reasoning capabilities; excels in Korean style and vocabulary adaptation.

Table 3: Key Characteristics of Language Models Used in Political Editorial Analysis

The evaluation relied on GPT-4o (OpenAI), to serve as an automated judge to assess the outputs of each model. Scoring was performed along three dimensions: logical coherence between claims and evidence, overall response quality (accuracy, relevance, consistency), and hallucination detection.

4.1 Ground-Truth

To validate model outputs and ensure the external reliability of the evaluation results, we employed expert-written analytical reports [1] from the Korea Institute for National Security Strategy (KINSS). These reports were manually curated by specialists in North Korean politics, military affairs, and inter-Korean relations, providing in-depth qualitative interpretations of the political messages and rhetorical structures of each editorial. An example of individual editorial analysis is provided in Appendix B.

Using these expert analyses as reference, we conducted a secondary human review on a randomly sampled subset of model outputs following the automated evaluation. This step was intended to complement the AI-based scoring by enhancing the precision and trustworthiness of the results, through human verification.

5 Results & Analysis

This section presents the quantitative results of model responses to *Rodong Sinmun* editorials, as evaluated using GPT-4o. Detailed scoring criteria are provided in Appendix A.

Overall, the models failed to detect certain strategic rhetorical features such as the de-mystification of the Supreme Leader and exhibited limitations in integrating complex or contradictory issues into a coherent interpretation. However, in the 1-shot condition, model outputs showed greater structural and interpretive alignment with human analysis compared to the 0-shot condition, suggesting that prompt-based guidance positively influences discourse reasoning.

Based on these findings, we provide a comparative analysis in the following sections of how each model responds to political discourse, examining their capabilities in interpreting, structuring, and evaluating ideologically driven texts.

5.1 Analysis of Claim–Evidence Coherence

To assess the models' actual reasoning capabilities, we focused on the logical coherence between the claims and their corresponding evidence within each editorial. This metric goes beyond surface-level response quality and aims to evaluate whether the model can accurately identify and connect semantic units through a valid inferential structure. It serves as a core indicator of the model's ability to engage in discourse-level reasoning. The coherence metric result from which model are summarized in the table below.

Model	0-shot	1-shot
GPT-o3 Mini	4.99	**5.00**
Claude 3.7	4.99	4.95
Gemini 2.0	4.95	4.99
EXAONE 7.8B	4.95	4.80

Table 4: Average Scores for Claim–Evidence Coherence Evaluation.

[1] https://www.inss.re.kr/publication/bbs/nk_list.do

GPT-o3 Mini demonstrated the most consistent performance across all prompt settings, achieving coherence scores close to the maximum (5.0). It reliably identified core claims, presented corresponding evidence in concrete terms, and explicitly established sentence-level logical connections.

Claude 3.7 also received favorable evaluations, particularly under the 1-shot condition, where its performance closely approached that of GPT-o3 Mini. However, approximately 4.6% of its responses were rated at level 3, as the supporting evidence tended to be explanatory rather than directly aligned with the central claim—indicating a relative weakness in evidential precision.

Gemini 2.0 consistently maintained a Claim–Evidence–Summary structure, demonstrating strength in formal coherence. Nevertheless, in approximately 1.1% of cases involving abstract topics, the evidence lacked sufficient specificity, leading to slightly weaker logical linkage. Although this had minimal impact on the overall average, it suggests a marginal decline in performance for more abstract editorial content.

EXAONE 7.8B exhibited relatively natural performance in identifying claims. However, in approximately 12.6% of its responses, relevant evidence was either missing or only weakly connected. This was particularly evident when the model overly fixated on the emotional and symbolic rhetoric of *Rodong Sinmun* editorials, repeatedly failing to shift toward logic-based analysis.

Overall, GPT-o3 Mini outperformed the other models in both structural explicitness and logical stability. Claude 3.7 showed strong consistency but lacked fine-grained alignment between claims and evidence. Gemini 2.0 offered solid structural scaffolding but weaker inferential integration, while EXAONE 7.8B showed a tendency to prioritize rhetorical affect over semantic reasoning. These findings suggest a persistent gap between surface-level fluency and genuine semantic inference in large language models. The Claim–Evidence coherence evaluation thus serves as a meaningful metric for quantifying this gap and may prove valuable for assessing LLM applicability in downstream NLP tasks involving reasoning over political discourse.

5.2 Response Quality Analysis

In this section, we quantitatively evaluate the response quality of each language model based on three sub-criteria: accuracy, relevance, and logical consistency. Each response was scored on a 5-point scale using the automated evaluation framework. The results are summarized as follows:

Model	Accuracy	Relevance	Logical Consistency
GPT-o3 Mini	4.94/ 4.85	4.98/ 4.96	5.00/ 5.00
Claude 3.7	4.90/ 4.75	4.90/ 4.95	5.00/ 4.96
Gemini 2.0	4.80/ 4.72	4.85/ 4.84	4.84/ 4.83
EXAONE 7.8B	4.31/ 3.86	4.84/ 4.20	4.82/ 4.33

Table 5: Average Scores for Response Quality Evaluation(0shot/1shot)

GPT-o3 Mini consistently achieved near-perfect scores across all categories, demonstrating the highest overall response quality. It reliably identified key arguments, maintained factually grounded and logically coherent reasoning, and exhibited stable performance regardless of prompt configuration.

Claude 3.7 also performed well, particularly in logical consistency and structured response construction. While 1-shot prompting improved its overall output quality, it occasionally introduced unsupported information or omitted critical details, resulting in slightly lower accuracy.

Gemini 2.0 excelled in structural composition, frequently utilizing repetitive rhetorical formats and summarization strategies. It was particularly effective in sequentially presenting leadership-oriented messages or policy narratives. However, its propensity to generalize topics or incorporate information not grounded in the source text resulted in a marked decline in accuracy.

EXAONE 7.8B showed consistently low response quality, with the weakest performance particularly in the 1-shot accuracy setting. Its outputs frequently included off-topic content and exhibited abrupt logical transitions. Performance gains remained minimal even with increased shot counts. For example, despite prompts designed to highlight Kim Jong Il's achievements in party-building and to signal the emergence of

Kimjongunism[2], the model often focused instead on Kim Jong Un's own contributions or produced analyses unrelated to the intended theme, indicating a lack of output reliability.

Category	Content
Editorial Title	Editorial on the 24th Anniversary of Kim Jong Il's Appointment as General Secretary (Oct 8, 1994)
Ground Truth	Emphasizes that the foundation of Party-building lies in Kim Jong Il's ideology and theory. Calls for the thorough establishment of the monolithic ideological system throughout the Party. Commemorates Kim Jong Il's leadership and achievements.
EXAONE 7.8B Output	Focuses on Kim Jong Un's Party-building achievements and reinforcement of socialism. Emphasizes organized Party operations in absolute obedience to Kim Jong Un. Highlights self-reliance and internal mobilization to solve funding and production issues.
Reasons for evaluation	*Accuracy score: 1* The AI misinterprets the editorial, which is intended to commemorate Kim Jong Il's legacy in Party-building and ideological leadership. While Kim Jong Un is briefly referenced in the original, EXAONE 7.8B places undue emphasis on his role, effectively shifting the main theme away from Kim Jong Il. This indicates a failure to distinguish symbolic continuity from thematic centrality in North Korean political discourse.

Table 6: A Case of Accuracy Error in EXAONE 7.8B's Interpretation of North Korean Political Discourse

In summary, GPT-o3 Mini provided consistently high-quality responses across all metrics. Claude 3.7 offered structurally sound outputs but lacked precision in content. Gemini 2.0 demonstrated strong formatting ability but limited inferential accuracy. EXAONE 7.8B struggled most with factual precision, though its Korean specialization suggests potential for future domain-specific tuning.

These findings highlight the partial success of CoT-based structuring while underscoring the persistent challenge of achieving fine-grained semantic inference. They suggest that although LLMs can simulate aspects of discourse analysis, their application to politically charged texts remains constrained.

In particular, despite its relatively low accuracy, EXAONE 7.8B demonstrated sensitivity to Korean rhetorical structures, indicating its potential as a domain-specific model. This points to the possibility of future performance improvements through model scaling and targeted fine-tuning.

Furthermore, human annotators were able to engage in deeper contextual interpretation. For instance, they inferred that the editorial commemorating Kim Jong Il's appointment as General Secretary was strategically framed to reflect Kim Jong Un's recent title change at the 8th Party Congress. In contrast, LLMs restricted their responses to the immediate content of the editorial, failing to account for broader historical or institutional context.

5.3 Analysis of Hallucination Types

In our evaluation, hallucination was assessed using a single composite score (1–5) per response, rather than assigning separate scores for each hallucination type. The rubric defines four representative types—factual insertion, emotional or ideological embellishment, causal overreach, and thematic abstraction—not as independent evaluation axes, but as qualitative indicators that guided holistic judgment. This approach was adopted to reflect the entangled nature of hallucination in political discourse, where multiple error types often co-occur or reinforce one another in a single output.

[2] "Kim Jong-un-ism" was made known to the outside world through a report released by South Korea's National Intelligence Service on October 28, 2021. Although it has not been officially adopted as North Korea's state ideology, it has been used internally to establish Kim Jong-un's independent leadership system. This can be interpreted as an attempt to construct a new ideological framework following Kim Il-sung-ism and Kim Jong-il-ism.

Model	0-shot	1-shot
GPT-o3 Mini	4.52	4.21
Claude 3.7	4.75	4.52
Gemini 2.0	4.92	4.92
EXAONE 7.8B	4.47	4.57

Table 7: Average Scores for Hallucination Detection Evaluation.

GPT-o3 Mini recorded the lowest hallucination score '4.21' under the 1-shot condition, indicating the most factually grounded output among all models. The hallucination rate decreased under the few-shot setting, and the responses remained stable and centered on verifiable content.

Claude 3.7 produced structurally consistent outputs but frequently included emotionally charged phrases in leader-centric or mobilization-themed editorials. The uncritical reproduction of ideological rhetoric contributed to slightly elevated hallucination scores of '4.75' in the 0-shot condition, and '4.52' in 1-shot condition. Gemini 2.0 recorded the highest hallucination score of '4.92' across both 0-shot and 1-shot settings. Although its outputs exhibited structural consistency, they frequently employed symbolic or ideologically charged language that compromised factual grounding. Such hallucinations appeared to be systematically embedded in its output patterns.

Category	Content
Editorial Title	Let Us Thoroughly Implement the Tasks Set Forth in the First Year of the Five-Year Plan, Upholding the Spirit of the 2nd Plenary Meeting of the 8th Central Committee of the WPK *(Feb 14, 2021)*
Ground Truth	The editorial calls for the full implementation of the first-year tasks of the new Five-Year Plan, as set forth by the 2nd Plenary Meeting of the 8th Party Central Committee, which was held just a month after the 8th Party Congress.
Gemini2.0 Output	Describes the editorial as emphasizing unconditional obedience to the Party's decisions, framing it as a typical example of North Korean propaganda.
Reasons for evaluation	The editorial's call for practical implementation of Party decisions is reduced to a narrative of "unconditional obedience," reflecting interpretive bias. This reframes the text's original policy- and action-oriented message as ideological compliance, resulting in a distortion of its core intent.

Table 8: Case of Hallucination Induced by Ideological Framing in Gemini2.0's Interpretation of a North Korean Editorial

EXAONE 7.8B exhibited an increase in hallucination scores despite the addition of more shots, with the score rising from '4.47' in the 0-shot condition to '4.57' in the 1-shot setting. The model showed a marked tendency to overreact to North Korean rhetorical expressions—such as "great," "absolute," and "historic"—frequently producing formulaic constructions, overly generalized summaries, and nonfactual content. Although the outputs were linguistically fluent, they consistently demonstrated low semantic precision.

Category	Content
Model Output	The editorial was interpreted as emphasizing the strengthening of North Korea's political-ideological capabilities and the continued advancement of socialism, in response to Kim Jong Un's *historic* policy speech.
Reasons for evaluation	*Hallucination score: 5* The term *"historic"* was not used in the source text to glorify Kim Jong Un's speech. The model introduces a value-laden interpretation that is absent from the original, resulting in semantic distortion.

Table 9: Example of Hallucination in EXAONE 7.8B's Output: Overreaction to Rhetorical Expressions

Overall, GPT-o3 Mini exhibited the lowest hallucination frequency and highest factual alignment, with few-shot prompting further reducing rhetorical deviation. Claude maintained structural stability but was prone to rhetorical interference. Gemini demonstrated a distinctive tendency to produce outputs in which

hallucinations were structurally embedded, driven by interpretive framing that distorted the intended meaning. EXAONE 7.8B tended to prioritize rhetorical surface features—such as "great" and "absolute"—over semantic fidelity, frequently leading to repeated factual inaccuracies. These findings suggest that hallucination in political discourse is not merely a factual error but a higher-order generation failure where rhetorical form distorts intended meaning. The uncritical reproduction of ideological content by LLMs, especially in texts like North Korean editorials, demonstrates how hallucination can emerge as a structural phenomenon. This highlights the need for heightened model accountability when applying LLMs to politically sensitive discourse tasks.

5.4 Reliability of Automated Evaluation and Overall Model Interpretability

In this section, we compare the automated evaluation results generated by GPT-4o (G-Eval) with human assessments based on approximately 30% of the dataset. The goal is to analyze the alignment and reliability of the automated scoring system. The evaluation criteria were consistent across both approaches: claim–evidence coherence, response quality, and hallucination severity. The average Pearson correlation coefficients are reported below.

Dimension	0-shot Avg.	1-shot Avg.
Accuracy	0.64	0.78
Relevance	0.59	0.75
Logical Consistency	0.73	0.64
Coherence	0.55	0.60
Hallucination	0.65	0.77

Table 10: Average Spearman correlation coefficients between model predictions and human judgments across evaluation dimensions in 0-shot and 1-shot settings. Most correlations are statistically significant ($p < 0.05$), indicating meaningful alignment between model and human evaluations.

As shown in the table, GPT-4o generally produced judgments that were well aligned with those of human evaluators, exhibiting high correlation across all evaluation criteria. In particular, correlation coefficients improved across most dimensions under the 1-shot prompt setting, suggesting that example-based prompting helped the model better internalize evaluation criteria and align more closely with human judgments.

However, for logical consistency, the 1-shot setting resulted in a lower correlation compared to 0-shot. This may indicate that the model became overly dependent on the prompt example, simplifying its reasoning process.

In addition, while G-Eval showed generally strong correlations with human raters, its score distribution tended to be skewed toward the higher end. For instance, in the coherence dimension, several responses received scores of 4.0 or above from the automatic evaluator, despite containing clear logical inconsistencies or stylistic issues. This pattern suggests that the model applies the scoring rubric conservatively and is relatively reluctant to assign lower scores. It indicates that GPT-4o tends to make more lenient judgments than human evaluators. While G-Eval achieves a reasonable degree of quantitative reliability, these findings underscore the need for calibrated interpretation when relying on its absolute scores.

When comparing the performance across models, clear differences emerged. While all models demonstrated a baseline capacity for structured generation and output stability, their ability to interpret political discourse varied significantly. GPT-o3 Mini and Claude 3.7 responded stably to the rhetorical structure and stylistic patterns of *Rodong Sinmun* editorials, maintaining thematic flow and coherent response composition. In contrast, EXAONE 7.8B and Gemini 2.0 frequently overreacted to political symbolism and emotional rhetoric, resulting in repeated hallucinations and semantic distortion.

These findings suggest that beyond surface-level fluency, LLMs still face structural limitations in interpreting the ideological abstraction and rhetorical complexity inherent in political discourse. This study empirically delineates the boundary between what LLMs can and cannot do in the context of political discourse analysis and underscores the need for a Hybrid Human-in-the-Loop framework where human interpretation complements, rather than simply post-processes, automated outputs.

6 Conclusion

This study empirically examined the potential for automating political discourse analysis by applying LLMs and an automated evaluation framework to North Korea's *Rodong Sinmun* editorials. Unlike

traditional approaches to political text analysis that rely primarily on rule-based methods or qualitative interpretation, this research proposed a structure-based framework that automatically extracts Claim–Evidence structures through CoT-based prompting and quantitatively evaluates the generated responses using the G-Eval framework. This approach offers a scalable and systematic methodological alternative that maintains the interpretive depth of qualitative analysis while ensuring the consistency and reproducibility of automation.

Experimental results show that LLMs can produce responses with a reasonable level of coherence and logical linkage. In particular, under the 1-shot condition, model outputs demonstrated high alignment with expert evaluations. The study also revealed differences in rhetorical structure interpretation across models and quantified tendencies in hallucination generation. These findings suggest that LLMs can serve not only as generative tools but also as potential instruments for analyzing and structuring ideological discourse.

At the same time, the models tended to reproduce or overinterpret political symbolism and emotional rhetoric without critical reasoning. This often manifested as exaggerated ideological framing or uncritical mimicry of rhetorical forms, revealing interpretive risks inherent in politically sensitive text generation.

G-Eval, the GPT-4o-based automatic evaluation system, achieved a fair degree of correlation with human assessments but failed to fully capture subtle contextual errors or rhetorical distortions. This highlights that automatic evaluation may also reflect model-internal biases, and excessive reliance on a single model could undermine both reliability and interpretive precision. Accordingly, a Hybrid Human-in-the-Loop framework that complements automatic scoring with expert judgment is proposed as a necessary strategy for high-fidelity political discourse analysis.

This study's methodology is distinguished from prior CoT prompting approaches by its design innovation. The prompt structure not only extracted claims and evidence but also guided higher-order inference by incorporating a political objective interpretation step. Evaluation criteria were also tailored to political discourse, including coherence, ideological consistency, and rhetorical hallucination. The experimental design aligned prompt construction, response generation, automatic scoring, and human reference evaluation in a tightly structured sequence, making it a well-organized empirical attempt to assess both the capabilities and limitations of LLM-based analysis.

As a pioneering study, this work demonstrates the viability of quantifying rhetorical structures in political texts and empirically evaluates the promise and constraints of automatic scoring systems. Future research may extend this framework to various languages and genres of political discourse, thereby further validating the generalizability and real-world applicability of LLM-based analytical methods.

7 Limitations

This study represents an empirical attempt to automate political discourse analysis, but it also has several limitations.

First, the evaluation of generated responses was conducted by a single expert, which may introduce subjective bias. The classification of hallucination errors also relied solely on GPT-4o's automated judgment, lacking inter-rater agreement measures or explicit annotation guidelines.

Second, the prompt was designed as a fixed five-step structure, but no ablation study was conducted to assess the contribution of each step. In addition, the one-shot prompt relied on a single demonstration, and its sensitivity to prompt selection was not systematically evaluated.

Third, while hallucination scores were calculated as composite values across four error types, the individual frequency and influence of each type were not analyzed. This limits the ability to make fine-grained comparisons of model-specific error patterns.

Lastly, the analysis focused on *Rodong Sinmun* editorials from a specific period, meaning the temporal scalability and cross-genre generalizability of the proposed methodology remain untested.

These limitations highlight the need for further research to improve the precision of automated systems and to capture the multilayered nature of political discourse. Future work could include ablation analyses of prompt components, few-shot designs with diverse examples, the development of annotation protocols with multiple evaluators, and the application of this framework across genres and languages to enhance both the robustness and scalability of automated political text analysis.

References

Baek, Seungjoo (2023). Analyzing the textual organization of North Korean Rodong Sinmun editorials. The Journal of Language & Literature, 94, 85-118. 10.15565/jll.2023.3.93.85

Lee Hang-Dong (1997). A Study on North Korea's Policy Change - With an Content Analysis of the Editorials in the Ro-Dong Sinmun. Korean Political Science Review, 31(4), 131-160.

Kim Yong Hyeon. (2003). A study on the political change in North Korea through the analysis of Rodong Shinmun: 1945-1950. NORTH KOREAN STUDIES REVIEW, 7(1), 107-127.

Heegwan chin. (2013). The Relations of Japan-North Korea and Chongryon Policy of North Korea: Through the analysis of Rodong Shinmun (1946-2010). The Korean Journal of Unification Affairs, 25(1), 361-396.

Kim. Jeong-ho, & Cho. Yunyoung (2022). North Korea"s Political-Economic Changes Under the Five-Year Economic Development Plan: Kim Jong-Un Regime"s rising ruling tasks and challenges on the 「Rodong Newspaper」. Journal of Peace Studies, 23(1), 155-178.

Liu, Z., Liu, Y., Zhang, Z., Di, L., Wei, F., & Wang, Y. (2024, February). Method for extracting power emergency plan information based on LLM Prompt Learning. In Proc. of SPIE Vol (Vol. 13080, pp. 130800G-1).

Xu, J., Sun, M., Zhang, Z., & Zhou, J. (2024). ChatUIE: Exploring Chat-based Unified Information Extraction using Large Language Models. arXiv preprint arXiv:2403.05132. https://arxiv.org/abs/2403.05132

Afzal, A., Kowsik, A., Fani, R., & Matthes, F. (2024). Towards Optimizing and Evaluating a Retrieval Augmented QA Chatbot using LLMs with Human in the Loop. arXiv preprint arXiv:2407.05925.

Le Mens, G., & Gallego, A. (2025). Positioning Political Texts with Large Language Models by Asking and Averaging. Political Analysis, 1–9. https://doi.org/10.1017/pan.2024.29

Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Large language models are inconsistent and biased evaluators. arXiv preprint arXiv:2405.01724.

Yang, K., Li, H., Chu, Y., Lin, Y., Peng, T. Q., & Liu, H. (2024). Unpacking Political Bias in Large Language Models: Insights Across Topic Polarization. arXiv preprint arXiv:2412.16746.

Kronlund-Drouault, P. (2024). Propaganda is all you need. arXiv preprint arXiv:2410.01810.

Gao, Q., & Feng, D. W. (2025). Deploying large language models for discourse studies: An exploration of automated analysis of media attitudes. PLOS ONE, 20(1), e0313932. https://doi.org/10.1371/journal.pone.0313932

Wenpeng Yin and Dan Roth. 2018. TwoWingOS: A Two-Wing Optimization Strategy for Evidential Claim Verification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 105–114, Brussels, Belgium. Association for Computational Linguistics.

Zhong, M., Liu, Y., Yin, D., Mao, Y., Jiao, Y., Liu, P., ... & Han, J. (2022). Towards a unified multi-dimensional evaluator for text generation. arXiv preprint arXiv:2210.07197.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.

OpenAI. (2024, April 17). Introducing the OpenAI o3 mini models. OpenAI. https://openai.com/index/openai-o3-mini/

Anthropic. (2024, June 20). Claude 3.5 Sonnet: faster, smarter, more accessible. Anthropic. https://www.anthropic.com/news/claude-3-7-sonnet

Google. (2025, February 15). Gemini model updates: February 2025. Google Blog. https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/

Research, L. G., Bae, K., Choi, E., Choi, K., Choi, S. J., Choi, Y., ... & Yun, H. (2025). EXAONE Deep: Reasoning Enhanced Language Models. arXiv preprint arXiv:2503.12524. Appendices

Ministry of Unification. 2021. North Korea Knowledge Dictionary. Seoul: Ministry of Unification.

A Appendix: Evaluation Criteria

1. Claim–Evidence Coherence Evaluation Criteria

Score	Description	Example	Evaluation Rationale
5	The supporting evidence clearly aligns with the claim, and the overall logical structure and meaning flow are consistent and coherent.	- The editorial calls for honoring Kim Jong Il's immortal revolutionary achievements on the occasion of Gwangmyeongseong Day, and for implementing the decisions of the 8th Party Congress to advance socialism. - Emphasis is placed on advancing socialism through the ideological and leadership continuity between Kim Jong Il and Kim Jong Un. - Citations from classical works like *Socialism is a Science* reinforce ideological legitimacy. - Statements such as "Respected Comrade Kim Jong Un is fulfilling the patriotic and strong nation-building aspirations of the great General" reinforce leadership succession. - The Party Congress is framed as a concrete political and economic action plan.	The claim and supporting evidence are organically connected, coherently linking theoretical legitimacy, leadership succession, and actionable goals. Logical structure and persuasive force are both strong.
4	The supporting evidence includes partial contradictions or tension with the claim, slightly weakening the legitimacy of the response.	- The editorial reaffirms Kim Jong Il's achievements while emphasizing the importance of implementing the 8th Party Congress decisions. - Citations of Kim Jong Il's ideological works are presented, but also suggest a gap with current realities. - Leadership succession is emphasized yet focus fluctuates between Kim Jong Il and Kim Jong Un. - References to the Party Congress include future goals, but also allude to unresolved challenges.	The response emphasizes socialist continuity, but the evidence introduces interpretive tension: theoretical-practical gaps, dual leadership focus, and incomplete outcomes. This weakens coherence and justification.
3	The relationship between the claim and evidence is unclear, and the justification lacks sufficient detail, making judgment difficult.	- The editorial commemorates Kim Jong Il's achievements and vaguely connects them to future direction. - Kim Jong Un's leadership is highlighted, with Kim Jong Il's ideology only symbolically mentioned.	While evidence exists, it lacks sufficient clarity or strength to justify the claim. Semantic links between ideological succession and practical content are weak or vague.

		- Citations like "Socialism is a Science" are disconnected from concrete policy. - The Party Congress is mentioned, but without detail or evaluative substance.	
2	The evidence includes both support and implicit contradiction, leading to interpretive ambiguity.	- The editorial commemorates Kim Jong Il while centering current leadership on Kim Jong Un. - Kim Jong Un's leadership is emphasized as the practical driver of socialism, while Kim Jong Il's ideology remains symbolic. - Party Congress content is referenced, but with limited mention of implementation or results.	The response expresses dual emphasis but lacks cohesion. It does not clearly justify the original ideological claim, and contradictory focus on past vs. present undermines consistency.
1	The response relies on irrelevant, fragmented, or selectively cited content that artificially supports the claim.	- The editorial introduces Kim Jong Un's international standing and recent achievements, rather than substantiating the ideological continuity from Kim Jong Il. - Citations from Kim Jong Il's works are missing or peripheral. - Party Congress goals are sidelined in favor of recent military or technological advancements.	The claim of ideological succession is not substantively supported. The response introduces unrelated topics (e.g., foreign evaluation, defense) and lacks logical justification.

Table A 1. Scoring Criteria for Claim–Evidence Coherence

2. Response Quality Evaluation Criteria

To evaluate the quality of model-generated responses, we adopt a three-dimensional framework comprising:

- **Accuracy**: The degree to which the response faithfully reflects factual content presented in the original editorial.
- **Relevance**: The extent to which the response directly captures the core themes, ideological messages, and strategic intent of the source text.
- **Logical Consistency**: The internal coherence of the response, including the logical progression of ideas, structural clarity, and stability of the narrative.

Each dimension is scored on scale from 1-5, with detailed scoring rubrics are provided in Tables **A**2-1~3.

Score	Description	Example	Rationale
5	The response fully and precisely matches the facts in the source text.	- The editorial commemorates Gwangmyeongseong Day by glorifying the revolutionary legacy of Kim Jong Il and calls for implementing the 8th Party Congress decisions to achieve new victories in socialism.	The response accurately reflects all key elements of the editorial—textual citations, leadership references, and policy phrases—without distortion or exaggeration. It

		- The model highlights Kim Jong Il's ideological contributions, the leadership of Kim Jong Un, and their joint role in advancing North Korean-style socialism. - Citations from classic texts such as *"Socialism is Science"* and *"Our Socialism Centered on the Masses Is Invincible"* are used to emphasize ideological legitimacy. - The response reflects Kim Jong Un's role in realizing Kim Jong Il's patriotic vision and the Party's five-year plan goals.	faithfully reproduces the core message.
4	Most information is accurate, but some details are slightly exaggerated or inaccurate.	- The editorial emphasizes Kim Jong Il's legacy while highlighting the importance of completing a ten-year national economic development strategy from the 8th Party Congress. - The model blends Kim Jong Il's classical thought with Kim Jong Un's pragmatic economic policies, focusing on self-reliance and science-based development. - However, the editorial actually refers to a five-year plan, not a ten-year strategy.	Although the overall context is preserved, certain terms (e.g., mislabeling "five-year plan" as "ten-year strategy") and emphasis (e.g., focus shifting from Kim Jong Il to Kim Jong Un) reduce factual precision.
3	Some factual inaccuracies exist, but the main idea is generally retained.	- The editorial commemorates both Kim Jong Il's revolutionary legacy and Kim Il Sung's anti-Japanese struggle. - The model emphasizes Juche ideology and identifies military buildup and foreign policy independence as Kim Jong Un's main agenda. - Kim Il Sung is referenced more than Kim Jong Il, whose writings are underrepresented. - The economic and ideological themes of the actual editorial are not sufficiently addressed.	Although the response touches on socialist legitimacy, the main message—emphasizing Kim Jong Il's leadership and the 8th Party Congress—is diluted by unrelated content. Core intent is partially retained but weakened.
2	Frequent factual errors lead to confusion in conveying the core message.	- The editorial is framed around Kim Jong Un's 10th year in power, emphasizing defense strategy and self-reliant military buildup. - Kim Jong Il's legacy and Gwangmyeongseong Day's ideological significance are omitted.	The key theme—celebrating Kim Jong Il's accomplishments—is missing. The focus is shifted to unrelated topics (e.g., military policy), causing thematic distortion and confusion.
1	The response is largely fabricated or factually inconsistent.	- Gwangmyeongseong Day is described as Kim Jong Un's birthday. - The editorial is said to focus on capitalist reforms and reconciliation with the U.S., citing a fictional quote from Kim Jong Un: "True socialism is	The description entirely contradicts the actual editorial. Key terms, individuals, and messages are either fabricated or misrepresented. Factual and

		impossible without the inflow of capital and technology."	contextual integrity is severely compromised.

Table A 2-1. Scoring Rubric for Accuracy

Score	Description	Example	Rationale
5	The response accurately captures the editorial's core themes.	- The editorial commemorates Gwangmyeongseong Day by honoring Kim Jong Il's revolutionary legacy and calls for implementing the 8th Party Congress decisions to achieve new socialist victories. - The model emphasizes the ideological centrality of Kim Jong Il's contributions and Kim Jong Un's leadership, citing key texts and Party goals.	The response reflects all key elements: the holiday's significance, leadership succession, and ideological continuity. Topic alignment and internal structure are coherent and consistent.
4	Minor divergence occurs as the response includes peripheral information.	- The editorial commemorates Kim Jong Il but focuses on Kim Jong Un's recent economic policies and leadership philosophy. - While ideological continuity is mentioned, emphasis shifts to economic achievements. - References to the 8th Party Congress are included but secondary.	While the response remains within the broader thematic scope, the main focus leans toward peripheral content. Core themes are partially reflected but slightly diluted.
3	Only part of the main theme is captured; secondary content dominates.	- The editorial highlights Kim Jong Un's governance and international strategy under sanctions. - Kim Jong Il is referenced symbolically, and policy content focuses on foreign relations. - The 8th Party Congress is framed as a diplomatic rather than ideological initiative.	The response emphasizes a different priority (e.g., foreign strategy), only lightly touching on the intended ideological focus of the editorial. Central themes are marginally addressed.
2	The response misinterprets the topic or lacks clear connection to the source.	- The editorial discusses Kim Jong Un's military leadership and weapons development as key to the nation's future. - Gwangmyeongseong Day is mentioned symbolically but disconnected from its ideological context.	The response shifts away from the intended subject—Kim Jong Il's legacy—and misrepresents the thematic core. Ideological continuity is largely absent.
1	The response is entirely unrelated to the editorial's main themes.	- The editorial focuses on North Korea's youth policy and university education reform. - No mention is made of Kim Jong Il, Kim Jong Un, or the 8th Party Congress.	The response bears no thematic connection to the original editorial. It discusses unrelated content, making the response irrelevant.

Table A2-2 Scoring Rubric for Relevance

Score	Description	Example	Rationale
5	The response is logically coherent and maintains a consistent tone and structure throughout.	- The editorial commemorates Gwangmyeongseong Day, celebrates Kim Jong Il's revolutionary legacy, and emphasizes implementing the 8th Party Congress decisions. - The narrative connects Kim Jong Il's ideological works to Kim Jong Un's leadership and the Party's five-year plan, presenting them as a unified framework. - The overall discourse links past ideology with present execution, achieving both ideological and practical persuasiveness.	The flow from Kim Jong Il's thought → Kim Jong Un's leadership → Party policy is clear and cohesive. Paragraph transitions are smooth, and the logical structure is firm and well-developed.
4	The overall tone is consistent, but minor gaps or ambiguities in logical flow appear.	- The editorial underscores Kim Jong Il's thought and ties it to Kim Jong Un's leadership and the importance of implementing Party decisions. - However, the final connection to economic self-reliance is slightly abrupt.	The response generally maintains consistency, but certain thematic transitions (e.g., ideological → economic emphasis) are underdeveloped or weakly linked.
3	There are some logical issues, but the overall flow is intact.	- The editorial highlights Kim Jong Il's legacy, justifies Kim Jong Un's socialist path, and introduces the 8th Party Congress. - However, unrelated themes such as youth enthusiasm and national defense appear toward the end.	While the initial structure is coherent, the later introduction of unrelated themes weakens cohesion. The response contains logical fragmentation but retains general structure.
2	The logical flow is frequently broken, with unclear transitions or inconsistent development.	- The editorial references Kim Jong Il's achievements, then abruptly shifts to Kim Jong Un's diplomacy and traditional cultural restoration. - Connections between ideological and policy narratives are unclear.	The narrative lacks smooth transitions, and the logical linkage between ideas is weak or missing. Shifts in topic reduce persuasiveness and structural integrity.
1	The response is logically unstable and lacks consistent structure.	- The editorial commemorates North Korea–China friendship, Kim Il Sung's anti-Japanese resistance, and Kim Jong Un's inter-Korean diplomacy. - The discourse shifts to sports diplomacy and peaceful socialism, suggesting reinterpretation of Kim Jong Il's life. - These themes are unrelated to the editorial's original purpose.	The response mixes unrelated topics without a coherent central thread. The lack of logical progression or consistent focus severely undermines credibility.

Table A2-3. Scoring Rubric for Logical Consistency

3. Hallucination

Hallucination is assessed based on the presence of factual inaccuracies, emotionally or ideologically exaggerated rhetoric, speculative causal inferences, or abstraction that distorts the original intent of the editorial. Four primary types are considered:

- **Factual Insertion:** Mentions of people, policies, or events not found in the source.
- **Ideological/Emotional Embellishment:** Rhetorical flourishes or glorifying language absent from the original text.
- **Causal Overreach:** Speculative or ideologically motivated inference not supported by the editorial.
- **Thematic Abstraction:** Specific claims reduced to vague ideological values.

Score	Description	Example	Rationale
5	The response is heavily distorted with pervasive embellishment, ideological exaggeration, or speculative interpretation that severely undermines the factual structure.	- Gwangmyeongseong Day is described as honoring "the greatest thinker in human civilization," and Kim Jong Il's philosophy is said to guide global politics for the next 1,000 years. - Kim Jong Un is claimed to have "defeated all imperialist powers," and the 8th Party Congress is described as "humanity's final revolutionary blueprint."	A combination of extreme exaggeration, ideological inflation, and factual fabrication severely undermines the editorial's meaning. This is a prototypical example of Level 5 hallucination.
4	Multiple sentences include ideological overreach or speculative narratives that diverge from the editorial's meaning.	- The editorial declares Kim Jong Il's thought a "universal ideological guide for humanity," and predicts global socialist unification under Kim Jong Un's leadership. - It frames the 8th Party Congress as a final confrontation to end capitalism.	Phrases like "global unification" and "end of capitalism" are ideologically motivated hallucinations not supported by the editorial. Interpretation diverges significantly from the source.
3	Promotional rhetoric or emotional expressions limit the clarity of factual content.	- The editorial says the people are "engraving in their hearts the legendary love and genius leadership of the General." - Kim Jong Un is portrayed as upholding "the red banner of Juche socialism with blood and sweat."	Phrases such as "genius leadership" and "red banner" introduce sentimentality that dilutes factual clarity. While the intent remains, persuasive value declines due to rhetorical overload.
2	Minor flowery or rhetorical expressions are added, but the meaning and facts remain intact.	- Gwangmyeongseong Day is called a "sacred day" to commemorate Kim Jong Il's immortal thought. - The editorial praises Kim Jong Un's leadership in achieving a socialist powerhouse.	Although phrases like "sacred day" and "immortal thought" are embellishments, they don't distort the factual core or argument. Rhetoric is present but minimally invasive.
1	No hallucinations observed; the response is	- The editorial commemorates Kim Jong Il's revolutionary	The response is factual, concise, and objective.

	factually grounded and faithful to the source.	achievements and affirms new socialist victories under Kim Jong Un's leadership. - Citations from Kim Jong Il's writings support the legitimacy of socialism, and the Party's five-year plan is emphasized.	There is no exaggeration, distortion, or ideological inflation—making it a reliable and hallucination-free summary.

Table A3. Scoring Rubric for Hallucination

B Appendix: Analysis Report on *Rodong Sinmun* Editorial

"Let Us Internalize the Juche Ideology as Our Worldview and Philosophy of Life" (May 7 Editorial)

❍ **These editorial urges readers to thoroughly embody the Juche ideology as both a worldview and a philosophy of life, applying it comprehensively to work and daily life.**

❍ **Juche is framed as "the eternal guiding ideology of our revolution," with emphasis placed on the roles of Kim Il-sung, Kim Jong-il, and Kim Jong-un:**
– Kim Il-sung is credited with founding the ideology, Kim Jong-il with its systematization and theoretical development, and Kim Jong-un with its succession and further advancement.
– Notably, the editorial highlights Juche ideology instead of the officially codified Kimilsung–Kimjongilism stated in the Party Charter, suggesting a renewed focus on internal strength and subjective power.

❍ **The editorial elaborates on what it means to internalize Juche as a worldview and philosophy of life:**
– Taking it as the starting point for all thought and action, and treating it as an absolute standard in life and struggle
– A necessary condition for preserving ideological purity and unity within the revolutionary ranks
– A key requirement for significantly strengthening internal power and achieving new victories in the revolution
– The fundamental guarantee for fully carrying through and completing the revolution

❍ **It also outlines the tasks required to internalize Juche ideology:**
– **Deep understanding of its significance**:
∆ Recognizing that Juche is the sole guiding ideology for our era and future
∆ Studying and fully internalizing its principles, legitimacy, scientific basis, and vitality
∆ Arming oneself with the proud history and revolutionary traditions
∆ Engaging in systematic and comprehensive study of the classical works of Kim Il-sung, Kim Jong-il, and Kim Jong-un
– **Integrating ideological study with revolutionary practice**:
∆ Grasping the truth and traction of Juche through the tangible superiority of socialism in our style
∆ Engraving the Party's line and policies, which embody Juche, onto one's consciousness
∆ Creating more socialist wealth through struggles aimed at comprehensive development of socialism in our style
– **Enhancing the roles of Party and working people's organizations**

Table B: Structured Expert Interpretation of One of the *Rodong Sinmun* Editorials Published in May 2021