Journalism-Guided Agentic In-context Learning for News Stance Detection

Anonymous ACL submission

Abstract

Understanding the overall stance of news articles is challenging due to their length and structural complexity. Yet, it is essential for supporting pluralistic and credible media environments. This paper introduces a novel stance detection dataset for Korean news, featuring annotations at both the article level and the segment level, informed by the narrative structure of news articles. Building on this resource, we propose an agentic in-context learning method that prompts a large language model (LLM) with segment-level stance predictions generated by a language model agent. Experiments across multiple LLMs demonstrate the effectiveness of the proposed framework for article-level stance detection and highlight its broader utility in enhancing diverse news recommendations and analyzing patterns of media bias.

1 Introduction

With the proliferation of digital platforms, online news consumption has become ubiquitous. In response, major news providers have shifted their publication channels from offline newspapers to online newspapers (Martens et al., 2018; Bhuller et al., 2024), and adopted personalized recommendation algorithms to enhance experience of news readers (Feng et al., 2020; Wu et al., 2023). However, such systems may inadvertently confine users within limited information environments, leading to filter bubbles and echo chambers that intensify political polarization (Flaxman et al., 2016; Duskin et al., 2024). To mitigate these effects, it is essential to automatically identify the perspectives embedded in news content and integrate them into recommendation algorithms, thereby promoting a more balanced media ecosystem.

Stance detection is a natural language processing task that aims to identify the perspective expressed in a text toward a specific target (Küçük and Can, 2020; Hardalov et al., 2022). Applying

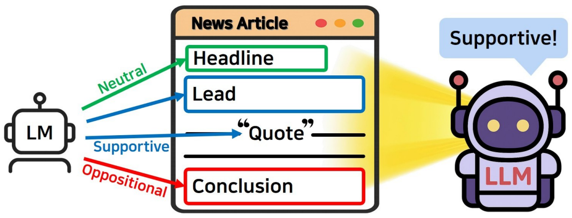


Figure 1: Key idea of SAAS, illustrating how article-level detection is performed by leveraging segment-level predictions generated by a language model agent.

stance detection to news articles can support balanced recommendations that reflect diverse viewpoints, thereby helping users make more informed decisions (Alam et al., 2022; Reuver et al., 2024). Additionally, it enables a data-driven understanding of media bias by allowing outlet-wise comparisons of stance distributions across a range of issues (Kuila et al., 2024).

Despite the growing need for stance detection methods in news articles, two significant gaps remain in prior research. First, most existing studies focus on short texts, such as individual sentences or tweets (Darwish et al., 2020; Glandt et al., 2021; Evrard et al., 2020). In contrast, news articles are often much longer, sometimes exceeding a thousand words. Within such lengthy texts, nuanced stances may vary across different segments. This makes it challenging for models to accurately infer the overall stance. Second, available datasets are mainly limited to high-resource languages (Li et al., 2021; Mascarell et al., 2021), such as English and German. The resource gap is even more significant for news article-level dataset. To enable more comprehensive and culturally grounded stance detection, it is crucial to develop datasets in non-major languages that reflect country-specific issues and linguistic nuances.

To address these gaps, this study introduces K-NEWS-STANCE, the first dataset for predicting the overall stance of full-length news articles in Korean. The dataset comprises 2,000 news articles, each manually annotated with its stance toward one of 39 nationwide issues. In addition to article-level annotations, we provide stance labels for smaller news components, including, the headline, concluding paragraph, and quotations within the body text. In total, K-NEWS-STANCE contains 19,650 segment-level stance annotations.

Building on this dataset, we propose a novel stance detection method, leveraging a large language model (LLM) through in-context learning, guided by a segment-level stance detection agent. We refer to this method as SAAS (**S**egment-level **A**gent for **A**rticle-level **S**tance Detection). As illustrated in Figure 1, the agent model is responsible for predicting stance labels for journalism-guided segments—such as the lead and quotations—which are incorporated into the prompt. This enables the LLM to better infer the overall stance of the article toward a given target issue. Experimental results show that SAAS outperforms existing methods, demonstrating the effectiveness of segment-level agency in article-level stance detection.

We make the following three key contributions.

- We introduce K-NEWS-STANCE, the first Korean dataset for article-level news stance detection, comprising 2,000 articles and 19,650 segments annotated with stance labels.
- We propose SAAS, an agentic in-context learning approach that predicts article-level stance by leveraging segment-level stance predictions generated by a language model agent.
- We present two case studies demonstrating the practical utility of SAAS in supporting pluralistic and trustworthy media environments.

2 Related works

Stance detection on news articles Prior research on stance detection and related work involving news data can be broadly categorized as three. First, several studies have focused on stance detection in news headlines (Yoon et al., 2019; Bourgonje et al., 2017; Ghanem et al., 2018; Borges et al., 2019), aligning with the broader trend of applying stance detection to short texts, such as tweets (Ferreira and Vlachos, 2016). Pomerleau and Rao (2017) introduced a dataset for classifying the stance of a news headline toward an unverified claim. The second line of research addresses stance detection in full news articles, which are substantially longer than headlines or tweets (Mets et al., 2024; Lüüsi et al., 2024; Mascarell et al., 2021; Conforti et al., 2020) and therefore pose greater challenges. Reuver et al. (2024) showed that few-shot detection using LLMs struggles with this task. Third, several studies have examined framing and media bias, which are closely related to stance detection. Card et al. (2015) introduced an annotated corpus of news articles for 15 frames across different social issues, which has since been widely used in automated frame analysis (Kwak et al., 2020; Card et al., 2016; Roy and Goldwasser, 2020). Other studies have aimed to predict the political bias of news articles at different levels (Baly et al., 2018, 2020; Hong et al., 2023; Chen et al., 2020). Most recently, Lin et al. (2024b) employed LLMs to predict the political bias of news articles. This study addresses the problem of detecting the overall stance of a news article, a task that has been relatively understudied in prior research.

LLM-based stance detection Prior research on stance detection has primarily focused on modeling the relationship between a given text and a target to infer the expressed stance (Küçük and Can, 2020; ALDayel and Magdy, 2021). Early approaches relied on bag-of-words representations (Mohammad et al., 2016) and recurrent neural networks (Augenstein et al., 2016; Du et al., 2017). More recent studies have leveraged pretrained language models. For example, several works have explored methods based on masked language models (MLM) (He et al., 2022; Chai et al., 2022; Li and Caragea, 2021); among them, Li et al. (2021) proposed an uncertainty-aware self-training method for BERTweet, while Liu et al. (2022) introduced a stance detection model pretrained on 3.6 million news articles. Building on this, recent studies have investigated the use of instruction-tuned LLMs. A preliminary study by Zhu et al. (2023) employed in-context learning by ChatGPT for stance detection, finding that its performance lagged behind human annotators. Cruickshank and Ng (2023) compared in-context learning and fine-tuning using open LLMs. More advanced methods have since emerged: Lan et al. (2023) introduced a multi-agent framework in which LLM experts collaborate on stance prediction; Li et al. (2023) retrieved and filtered background knowledge from Wikipedia; and Zhang et al. (2024b) extracted diverse forms of stance-related knowledge from LLMs to train stance classifiers. Our work contributes to this line of research by proposing a novel LLM-based stance detection method that employs segment-level signals.

3 Problem and Dataset

3.1 Target Problem

We address the task of stance detection in news articles, which involves identifying the positional stance of a news article toward a given social issue. Formally, given a news article A covering a target issue T, the objective is to determine the overall stance of A toward T. The stance label L is categorized into one of three classes: *supportive*, *neutral*, or *oppositional*. A stance detection model $f(\cdot)$, which takes A and T as input, is tasked with predicting L. Model performance is evaluated using standard classification metrics.

The target problem represents a specialized case of stance detection, an NLP task aimed at determining the position or attitude expressed in a text regarding a particular target (Küçük and Can, 2020; Hardalov et al., 2022; Zhang et al., 2024a). While stance detection has been widely studied in the context of short-form content such as tweets, forum posts, or headlines, its application to long-form journalistic texts remains a formidable challenge due to the complex nature of news articles, which can be summarized in two key aspects.

First, professional journalism typically privileges verification over assertion (Kovach and Rosenstiel, 2021). Adhering to normative ideals of neutrality and balance, news articles often refrain from making overt evaluative claims. Instead, they rely on indirect cues, such as source selection (Zoch and Turk, 1998; Druckman and Parkin, 2005), narrative framing (Nelson et al., 1997; Gentzkow and Shapiro, 2010), and lexical subtleties (Simon and Jerit, 2007; Schuldt et al., 2011), to communicate a stance, if any, toward a given issue. Even when an article expresses a positional preference, it is frequently nuanced, hedged, or ambivalent, making it difficult for models to detect without a deep understanding of rhetorical and discursive context.

Second and relatedly, the stance expressed in a news article is rarely concentrated in a single sentence or paragraph. Rather, it is often distributed across multiple textual layers, including headlines, leads, quotations, and framing devices. These layers may contain conflicting or ambiguous signals, especially in articles that attempt to present multiple sides of an issue. Accordingly, stance detection models must be capable of synthesizing fragmented and context-dependent cues across the entire document, a task made more challenging by the sheer length of the news texts.

Furthermore, despite recent advances in language understanding, LLMs often struggle to retain salient contextual information when processing long documents (Liu et al., 2024), leading to degraded performance (Reuver et al., 2024). This limitation is particularly pronounced in the news domain, where articles are significantly longer and more discursively layered than the short texts—such as tweets or single sentences—commonly used in prior stance detection research.

To address these challenges, we propose a hierarchical modeling approach that first infers the stance at the level of smaller discourse units (e.g., paragraphs or sections), and subsequently integrates these local predictions to determine the overall stance of the article. This architecture is designed to retain local context and capture dispersed stance cues in assessing how different parts of a news story contribute to its overall position on an issue.

3.2 Dataset: K-News-Stance

Despite prior research on news stance detection (Lüüsi et al., 2024; Mascarell et al., 2021; Liu et al., 2022), most existing studies focus on high-resource languages such as English and German. As a key contribution, this study introduces a new annotated corpus in Korean. The dataset includes manually labeled stance annotations for sub-components of news articles, enabling fine-grained analysis and facilitating the development of more advanced stance detection methods.

Raw data collection We collected Korean news articles published between June 2022 and June 2024 using BigKinds and Naver News. BigKinds, operated by the Korea Press Foundation (KPF), is a news platform that provides metadata (e.g., headlines, publishers) for weekly nationwide issues across diverse domains, including labor, gender, national, and international affairs. As a government-affiliated organization, the KPF curates a comprehensive news archive, ensuring that BigKinds captures social issues of national significance. From this archive, we randomly sampled 47 issues, maintaining temporal balance over the two-year period. Becuase BigKinds does not provide full text, we retrieved the corresponding content using the Naver News search API, a major news aggregator in Ko-

Target Issue: The National Assembly's Approval of the Ban on Dog Meat Consumption

Headline (*Supportive*): "Classic Kim Keon-hee"... Dog Meat Ban Wins Rare Praise from Gae-ddal

Body Text

-Lead (*Supportive*): A landmark bill banning the slaughter, breeding, and sale of dogs for consumption has passed the National Assembly, marking a pivotal moment in Korea's evolving stance on animal welfare. Commonly referred to as the 'Kim Keon-hee Law'

-Conclusion (*Supportive*): According to a 2023 national survey conducted by Korea Research International on behalf of the Animal Welfare Research Institute Aware, 94.5% of respondents reported not having consumed dog meat in the past year.

Overall Stance: *Supportive*

Headline (*Neutral*): Ban on Dog Meat Consumption: 'A Historic Victory' vs 'Awaiting Constitutional Appeal'

Body Text

-Lead (*Neutral*): The passage of Korea's so-called 'Dog Meat Ban Bill' on January 9 has triggered sharply divergent responses from advocacy groups and industry representatives. Animal rights organizations celebrated the vote as a watershed moment, declaring it "a historic victory for animal rights." Meanwhile, the Korea Dog Meat Association, which has fiercely opposed the legislation, condemned the decision as an infringement on basic rights, stating that "the freedom to choose one's occupation has been taken away." The group announced its intention to file a constitutional appeal.

-Conclusion (*Neutral*): The bill, formally titled, 'The Special Act on the Termination of Dog Breeding, Slaughter, Distribution, and Sale for Food Purposes,' bans all commercial activities involving dogs for human consumption, including breeding, slaughter, distribution, and sale.

Overall Stance: *Neutral*

Headline (*Oppositional*): With the ban on Dog Meat Passed, Longtime Boshintang Vender Says, 'I'm at a loss'

Body Text

-Lead (*Oppositional*): Confusion and concern are growing among dog meat industry workers following the National Assembly's approval of a bill that will outlaw the dog meat consumption. Although enforcement measures and penalties won't take effect until 2027, the law has already sparked fresh controversy. While officials argue that the legislation will end decades of bitter debate, questions are mounting about how to manage the sudden increase in abandoned dogs and unregulated breeding facilities.

-Conclusion (*Oppositional*): Mr. B, a long-time vendor in the industry, said "Once the suppliers disappeared, I had no choice but to shut down for a while because I simply couldn't get any meat." He added, "My rent is 1.6 million won per month. To stay afloat, I need to make at least 800,000 won per day, but now I'm barely making 200,000. At this rate, I won't be able to keep the doors open."

Overall Stance: *Oppositional*

Table 1: Data examples translated in English, of which the remaining body text is omitted for brevity. The colored highlight indicates the stance label for quotations (blue: supportive, red: oppositional).

rea. The data collection comprises 2,989 articles from 31 news outlets, which were then manually annotated for stance in the subsequent step.

Manual annotation In collaboration with a journalism scholar holding a Ph.D. in mass communication, we developed a manual annotation guideline for labeling the stance of a news article toward its target issue, as well as the stance expressed in its sub-components. The guideline is informed by features of the narrative structure known to signal stance, including information selection (Nelson et al., 1997; Zoch and Turk, 1998; Gentzkow and Shapiro, 2010; Druckman and Parkin, 2005), patterns of direct quotation (McGlone, 2005; Han and Federico, 2018; Song et al., 2023), lexical choices (Simon and Jerit, 2007; Schuldt et al., 2011), and cues that imply preferred interpretations or intended actions (Bolsen and Druckman, 2015; McIntyre, 2019).

The annotation process consists of two main tasks. The first involves classifying each article into one of four categories: straight news (factual reporting without interpretation), analysis (providing contextual background, implications, or future scenarios), opinion, and others (e.g., interviews). For stance annotation, we exclusively focus on analysis and opinion, as these genres are more likely to contain opinionated content (Alhindi et al., 2020). The second task asks annotators to assess the article's stance toward a given issue, classifying it as *supportive*, *neutral*, or *oppositional*. Beyond the article-level stance annotation, we also label four key structural components of the article: the headline, lead, conclusion, and direct quotations. In cases where stance is ambiguous, annotators are advised to consult additional articles addressing the same issue to enhance contextual understanding and maintain consistency in labeling.

Two annotators from our institution were recruited and trained to follow the annotation guidelines meticulously. The annotators labeled all 2,000 articles and 19,650 segments, achieving substantial inter-coder reliability, with Kripendorff's alpha ranging from 0.68 to 0.84 across different segments and article-level annotations. In cases of disagreement, annotation conflicts were resolved through discussion and consensus. The detailed guidelines and labeling interface are shown in Figure A4.

Table 1 presents three annotation examples that illustrate different stances on the same issue. As shown in these examples, segment-level stance labels offer important cues for interpreting the overall position of a news article toward the issue. The original example in Korean is in Table A7.

Dataset statistics The final dataset comprises 2,000 news articles covering 47 distinct issues. Following prior work on stance detection (Reuver et al., 2024), we divide the dataset into two splits such that each split contains a disjoint set of issues—24 for training and 23 for testing. Accordingly, the training and test sets consist of 999 and 1,001 articles, respectively. This issue-level split prevents models from relying on issue-specific cues when predicting stance labels. Table 2 presents descriptive statistics for the training and test sets, which are broadly comparable. On average, articles contain 1,483 characters, with lengths ranging from 376 to 8,185 characters. Each article includes an average of 7.8 direct quotations, with the number ranging from 0 to 45. Further analysis of label distributions and cross-segment associations is provided in Section A.1.

4 Proposed Method: SAAS

LLMs can be adapted to new tasks without parameter updates by providing task instructions in the input prompt, a technique known as *in-context learning*. However, applying this approach to the target task is suboptimal due to the length and structural complexity of news articles, which often leads to context loss (Liu et al., 2024) and degraded performance (Bertsch et al., 2025). To address this limitation, we propose SAAS (**S**egment-level **A**gent for **A**rticle-level **S**tance Detection), an agentic in-context learning framework that enhances LLM prompts by incorporating stance labels for shorter, journalism-guided structural segments of the article. These labels are predicted by a dedicated language model (LM) agent and used as auxiliary signals to support article-level stance inference.

Segment-level stance detection We employ an LM agent to infer stance labels for shorter segments of a given news article. We assume that the agent can accurately predict the stance of these segments, thereby assisting the LLM in inferring the overall stance of the full article based on segment-level signals.

	Train	Test	Total
# Articles	999	1001	2000
(S/N/O)	(314/346/339)	(323/330/348)	(637/676/687)
# Issues	24	23	47
# Characters (max)	5451	8185	8185
# Characters (mean)	1478	1489.14	1483.58
# Characters (median)	1348	1318	1335
# Characters (min)	376	413	376
# Quotations (max)	45	20	45
# Quotations (mean)	7.93	7.62	7.78
# Quotations (median)	7	7	8
# Quotations (min)	0	0	0

Table 2: Descriptive statistics (S:Supportive, N:Neutral, O:Oppositional)

Specifically, we analyze the following subcomponents of news articles, grounded in journalism research (Mencher and Shilton, 1997):

- Headline: Conveys the core message of the article and is designed to be clear and easily understood at a glance.
- Lead: Typically the first paragraph. Following the inverted pyramid structure, it summarizes the most important information and generally addresses at least three of the six classic questions (5Ws and 1H): Who, What, Where, When, Why, and How.
- Conclusion: The final paragraph, which often reinforces the main points or offers final context or interpretation.
- Quotations: Direct speech from sources, included to provide evidence, perspectives, or rhetorical impact.

We consider two types of LMs for segment-level stance prediction: (1) an LLM that performs in-context learning without parameter updates, and (2) a fine-tuned MLM trained on the segment-level annotations. According to the comparison experiments, we adopt a fine-tuned MLM for the best-performing variant.

Article stance prediction To predict the overall stance of an article toward a target issue, we prompt an LLM with task instructions while augmenting segment-level stance labels predicted by an LM agent. These labels are embedded into the article using an XML-like format, enabling the LLM to incorporate them as contextual cues during inference. The proposed method is model-agnostic and can be applied to any instruction-following LLM. The prompt format and an example are shown in Figure A2.

5 Evaluation

We conduct evaluation experiments to assess the effectiveness of the proposed method for article-level news stance detection in comparison to existing approaches. Detailed experimental settings are provided in Section A.2.

5.1 Baseline Methods

We employ four fine-tuned methods as baselines, each demonstrating state-of-the-art performance on stance detection benchmarks. **(1)** RoBERTa (Liu et al., 2019) is fine-tuned for article-level stance detection. **(2)** CoT Embeddings (Gatto et al., 2023) is a fine-tuned RoBERTa model on the explanation trace of an LLM for determining the stance of a given news article. **(3)** LKI-BART (Zhang et al., 2024c) is an encoder-decoder model that incorporates contextual knowledge from an LLM into stance detection by prompting the LLM with both the input and target. **(4)** PT-HCL (Liang et al., 2022) is a hierarchical contrastive learning method designed to distinguish between target-invariant and target-specific features. Model checkpoints are provided in Section A.2.

5.2 Results

Table 3 presents the evaluation results of the baseline and proposed methods on the test set of K-NEWS-STANCE. Table 3a reports the performance of existing state-of-the-art methods, all trained on the training set. Table 3b shows the performance of LLM-based in-context learning methods, including an additional baseline that uses instruction-only prompting and two variants of the proposed SAAS, each listed in a separate row. The second and third columns indicate whether each model incorporates advanced prompting techniques, such as chain-of-thought reasoning (Wei et al., 2022) or few-shot sample augmentations (Brown et al., 2020). The remaining columns present performance across different LLM backbones used for article-level stance prediction. Specifically, we evaluate three proprietary LLMs—GPT-4o-mini, Gemini-2.0-flash, and Claude-3-haiku—used without parameter updates, alongside one open-weight model, Exaone-2.4b, which has been instruction fine-tuned. These models are selected for their strong performance in Korean language understanding (LG AI Research, 2024).

We implemented two SAAS variants based on the source of segment-level stance prediction. The first assumes an idealized *oracle* scenario in which segment-level stance labels are perfectly accurate, serving as an upper bound on the model's potential performance. The second variant replaces the oracle with a fine-tuned RoBERTa model trained on segment-level annotations from the training set, simulating realistic prediction conditions. We select the RoBERTa model as the representative segment-level agent due to its competitive performance in article-level stance detection, as shown in Table A4.

Three primary observations emerge from Table 3. First, among the baseline methods, PT-HCL achieves the highest performance, with an accuracy of 0.617 and an F1 score of 0.618, followed by the fine-tuned RoBERTa model. These results highlight the effectiveness of contrastive learning and standard MLM fine-tuning for stance detection.

Second, among the LLM-based in-context learning baselines, Gemini-2.0-flash stands out as the best-performing LLM when combined with CoT prompting, reaching an accuracy of 0.661 and F1 of 0.657. This outperforms all fine-tuned baselines. The effectiveness of CoT and few-shot prompting varies across different LLMs, suggesting model-specific sensitivity to prompting strategies.

Third, as demonstrated by the performance of SAAS (RoBERTa), incorporating the RoBERTa-based segment-level agent consistently improves stance detection performance across all LLMs, yielding gains of up to +0.071 in accuracy and +0.1 in F1 compared to the in-context learning baselines. In this configuration, Gemini-2.0-flash once again achieves the strongest results when combined with CoT prompting and six-shot augmentation. However, the overall performance remains constrained by the quality of segment-level predictions, as evidenced by the persistent and substantial performance gaps between this setting and the oracle variant across all LLMs. The largest gap of +0.236 in accuracy and +0.238 in F1 was observed for the fine-tuned Exaone-2.4b, which is the highest performance under the oracle setting despite its relatively small model size. This result highlights the potential of SAAS for efficient implementation in practical, resource-constrained scenarios.

Ablation study on segment labels We conduct an ablation study to assess the contribution of stance labels from individual news segments to overall article-level stance prediction. Specifically,

RoBERTa		CoT Embeddings		LKI-BART		PT-HCL	
Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
0.594±0.029	0.577±0.046	0.582±0.028	0.562±0.039	0.545±0.011	0.538±0.014	**0.617±0.007**	**0.618±0.008**

(a) Existing methods

Method	+ CoT	+6-shot	GPT-4o-mini		Gemini-2.0-flash		Claude-3-haiku		Exaone-2.4b (finetuned)	
			Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
Baseline			0.594±0.002	0.577±0.002	0.636±0.001	0.628±0.002	**0.568±0.002**	**0.538±0.002**	**0.554±0.003**	**0.544±0.003**
	√		**0.597±0.002**	**0.581±0.003**	**0.661±0.002**	**0.657±0.002**	0.561±0.002	0.533±0.003	0.54±0.001	0.539±0.003
		√	0.565±0.003	0.54±0.003	0.635±0.003	0.631±0.004	0.545±0.004	0.523±0.005	0.445±0.004	0.431±0.003
	√	√	0.526±0.003	0.494±0.003	0.641±0.005	0.635±0.004	0.541±0.004	0.51±0.004	0.456±0.002	0.443±0.002
SAAS (Oracle)			0.724±0.001	0.706±0.001	0.758±0.001	0.753±0.001	0.791±0.002	0.789±0.002	**0.837±0.004**	**0.837±0.004**
	√		0.716±0.002	0.692±0.002	**0.778±0.001**	**0.776±0.001**	0.74±0.001	0.732±0.001	0.813±0.003	0.821±0.002
		√	**0.796±0.003**	**0.798±0.003**	0.772±0.003	0.769±0.003	**0.815±0.001**	**0.816±0.001**	0.344±0.004	0.308±0.004
	√	√	0.747±0.001	0.74±0.002	0.777±0.003	0.773±0.003	0.794±0.002	0.797±0.002	0.338±0.005	0.3±0.006
SAAS (RoBERTa)			0.571±0.002	0.53±0.001	0.633±0.003	0.619±0.004	0.591±0.004	0.577±0.004	0.584±0.002	0.581±0.001
	√		0.553±0.001	0.509±0.001	0.633±0.001	0.626±0.001	0.579±0.003	0.552±0.004	**0.601±0.006**	**0.599±0.006**
		√	**0.607±0.006**	**0.602±0.007**	0.662±0.001	0.657±0.001	**0.639±0.002**	**0.638±0.004**	0.354±0.003	0.331±0.003
	√	√	0.591±0.002	0.57±0.002	**0.678±0.002**	**0.672±0.002**	0.61±0.004	0.608±0.004	0.332±0.001	0.308±0.001

(b) LLM in-context learning methods

Table 3: Performance for predicting overall stance of news articles

we evaluate model variants in which the stance label of a particular segment is omitted from the prompt. Table 4 reports the results for two settings: the oracle variant of SAAS using Exaone-2.4b, and the best-performing RoBERTa-agent variant, which uses Gemini-2.0-flash as the backbone LLM with CoT prompting and six-shot sample augmentation. The first row reports the performance of SAAS in its best configuration, while the subsequent rows show the results of the ablated variants. In the oracle setting, removing the lead segment results in the largest performance drop, while removing the headline or quotations yield the smallest. These findings highlight the importance of incorporating multiple segments in context, rather than focusing on a single segment, such as news headlines, as commonly addressed in prior stance detection research (Ferreira and Vlachos, 2016; Bourgonje et al., 2017). When predicted segment-level labels are used, the performance gaps from ablation are reduced. In this setting, removing direct quotations leads to the smallest drop, suggesting that quote-level stances are more difficult to interpret due to their brevity and subtlety. In Section A.4, we present the results of an additional ablation experiment that evaluates the effectiveness of using journalism-guided segments, compared to label augmentations based on random segments.

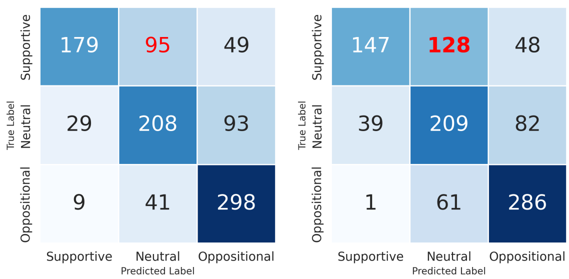
Model	ACC	F1
SAAS	0.837±0.004	0.837±0.004
w/o Headline	0.827±0.002	0.827±0.001
w/o Lead	0.763±0.001	0.75±0.001
w/o Conclusion	0.767±0.003	0.77±0.003
w/o Quotations	0.828±0.004	0.826±0.003

(a) SAAS (Oracle)

Model	ACC	F1
SAAS	0.678±0.002	0.672±0.002
w/o Headline	0.672±0.003	0.67±0.004
w/o Lead	0.61±0.001	0.63±0.002
w/o Conclusion	0.663±0.003	0.658±0.005
w/o Quotations	0.676±0.002	0.669±0.002

(b) SAAS (RoBERTa)

Table 4: Performance by ablating segment labels



(a) SAAS (b) Baseline

Figure 2: Confusion matrix

Error analysis Figure 2 presents the confusion matrices for the best-performing variant of SAAS using the RoBERTa agent and the baseline in-context learning methods, both of which use Gemini-2.0-flash as backbone. The results show that the proposed method achieves higher accuracy across all three target classes. Both models exhibit the greatest difficulty in correctly classifying articles with the gold label *supportive*, followed by *neutral* and *oppositional*. For 323 supportive-labeled articles, the baseline method frequently misclassified them as *neutral* (128 instances) or *oppositional* (48 instances). While SAAS performs better overall, it follows a similar error pattern, highlighting the challenge of identifying cues indicative of supportive stances—an area for future investigation.

6 Case studies

We conduct two case studies to highlight potential applications of SAAS. We additionally collected recent news data for six randomly selected issues from July 2024 to April 2025. The stance labels were manually annotated by the same annotators involved in the primary dataset.

Diversity in news recommendation The first case study investigates whether stance predictions by SAAS can enhance political diversity in news recommendations. We assume a scenario in which ten different users are each recommended a set of news articles after reading an initial article. As a baseline recommender based on content similarity, we use a multilingual version of Contriever (Izacard et al., 2022) to retrieve the top-20 most similar articles for each user from the newly collected article pool covering four distinct issues.

We then apply two versions of the Maximum Marginal Relevance (MMR) re-ranking method (Carbonell and Goldstein, 1998) to these initial recommendation. The first is the standard MMR approach, which ranks articles based on embedding similarity. The second, denoted as MMR (SaaS), incorporates predicted stance labels from SAAS, encoded as one-hot vectors, to promote stance diversity during re-ranking.

Table 5 presents the evaluation results, reporting the average values of Diversity and Precision@K for varying values of K (from 5 to 10). Diversity is measured as the entropy of the political preference distribution among the recommended articles, with higher entropy indicating greater ideological diversity. Since the political leaning associated with each stance label can vary by issue, we manually map each stance label to one of three political categories: progressive, moderate, and conservative.

The results indicate that re-ranking with stance predictions from SAAS leads to higher diversity scores, with only a slight reduction in precision compared to the baseline. Furthermore, the proposed re-ranking approach achieves a comparable level of precision to the standard MMR while yielding higher diversity. These findings demonstrates the potential of SAAS for promoting politically diverse news recommendations.

Political bias in news outlets The second case study demonstrates the utility of SAAS as an analytical tool for identifying media bias. Figure 3 presents a scatterplot in which each point reflects the distribution of predicted *supportive* and *oppositional* stances across news articles published by six major Korean news outlets. The analysis centers on two salient social issues tied to the 2025 presidential election[1]. Outlets are categorized as either progressive or conservative based on ideological classifications established in prior literature (Han et al., 2023; Song, 2007; Jo, 2003). The resulting clusters reveal clear differences in stance patterns that align with each outlet's known editorial stance. These findings underscore the potential of SAAS to map partisan bias in news coverage and support large-scale analyses of the media bias landscape.

Method	k =5		k =10	
	Diversity	Precision	Diversity	Precision
Contriever	0.535	**1**	0.723	**0.983**
+ MMR	0.622	0.975	0.764	0.969
+ MMR (SaaS)	**0.647**	0.983	**0.793**	0.971

Table 5: Effects of incorporating predicted stances for prompting diversity in news recommendation

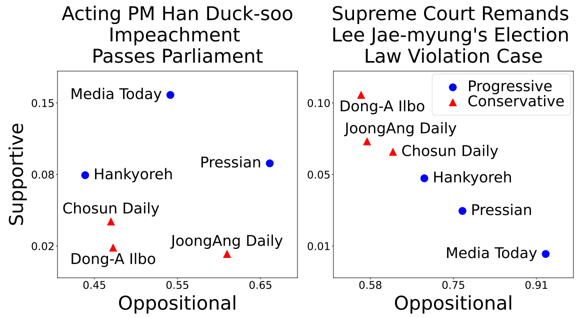


Figure 3: Predicted stance label distribution, grouped by the political leaning of six major news outlets

7 Conclusion

This paper presented a novel dataset for news stance detection in Korean, which includes stance annotations for both whole articles and journalism-guided news segments. Building on this resource, we proposed an agentic in-context learning method that improves article-level stance detection by LLMs through the augmentation of segment-level stance labels generated by an LM agent. Two case studies demonstrate the broader applicability of the proposed dataset and method beyond benchmarking, supporting efforts toward pluralistic and credible media environments.

[1] Ongoing at the time of submission.

Limitations

While our framework is evaluated on a single dataset, this reflects the fact that, to our knowledge, it is the first resource to pair journalism-guided segment-level labels with article-level stances. The consistent gains observed across five different LLMs—including sizable improvements under the oracle-setting—suggest that the core ideas are not confined to this particular corpus. Nevertheless, validation across additional languages and media ecosystems will be essential. To facilitate, we will publicly release our annotation protocol to lower the barrier to adoption.

Our approach currently incurs additional inference costs due to the invocation of an LM agent at the segment level. Encouragingly, under the oracle setting, Exaone-2.4b achieves the highest performance, suggesting the potential for lightweight deployment. Further, techniques such as post-training quantization (Lin et al., 2024a) may help reduce inference overhead without compromising accuracy.

Finally, the two case studies serve as proof of concept rather than exhaustive benchmarks. They demonstrate how the proposed method can benefit politically diverse recommendations and large-scale analyses of the media bias landscape, but do not yet cover the full spectrum of real-world news genres and delivery platforms. Future work could replicate these studies across broader scenarios to validate their impact more comprehensively.

Ethics Statement

We constructed K-NEWS-STANCE for training and evaluating article-level stance detection, based on publicly available news articles retrieved via API. Since these articles are produced under strict journalistic standards, the use of this data raises minimal privacy concerns. While the primary purpose of the dataset is to support article-level stance detection, it is also suitable for segment-level stance detection, as demonstrated in our training of a fine-tuned RoBERTa model for segment-level prediction. Beyond benchmarking, K-NEWS-STANCE has broader applicability for developing and evaluating stance detection models that contribute to pluralistic and credible media environments, as illustrated in our two case studies. The dataset will be released exclusively for academic purposes—such as benchmarking and media research—to respect the intellectual property rights of the original news publishers. Two graduate students (one female and one male) in an author's institution were recruited for manual annotation. In compliance with local wage regulations, they were compensated at a rate of approximately USD 7 per hour. Parts of this manuscript were proofread using ChatGPT. This research was approved by the Institutional Review Board at an author's institution.

References

Mehwish Alam, Andreea Iana, Alexander Grote, Katharina Ludwig, Philipp Müller, and Heiko Paulheim. 2022. Towards analyzing the bias of news recommender systems using sentiment and stance detection. In *Companion Proceedings of the Web Conference 2022*, WWW '22, page 448–457, New York, NY, USA. Association for Computing Machinery.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

Tariq Alhindi, Smaranda Muresan, and Daniel Preotiuc-Pietro. 2020. Fact vs. opinion: the role of argumentation features in news classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6139–6149, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4982–4991, Online. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Dimitar Alexandrov, James Glass, and Preslav Nakov. 2018. Predicting factuality of reporting and bias of news media sources. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3528–3539. Association for Computational Linguistics.

Amanda Bertsch, Maor Ivgi, Emily Xiao, Uri Alon, Jonathan Berant, Matthew R. Gormley, and Graham Neubig. 2025. In-context learning with long-context models: An in-depth exploration. In *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12119–12149.

Manudeep Bhuller, Tarjei Havnes, Jeremy McCauley, and Magne Mogstad. 2024. How the internet changed the market for print media. *American Economic Journal: Applied Economics*, 16(2):318–58.

Toby Bolsen and James N Druckman. 2015. Counteracting the politicization of science. *Journal of Communication*, 65(5):745–769.

Luís Borges, Bruno Martins, and Pável Calado. 2019. Combining similarity features and deep representation learning for stance detection in the context of checking fake news. *J. Data and Information Quality*, 11(3).

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Jaime Carbonell and Jade Goldstein. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dallas Card, Justin Gross, Amber Boydstun, and Noah A. Smith. 2016. Analyzing framing through the casts of characters in the news. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1410–1420, Austin, Texas. Association for Computational Linguistics.

Heyan Chai, Siyu Tang, Jinhao Cui, Ye Ding, Binxing Fang, and Qing Liao. 2022. Improving multi-task stance detection with multi-task interaction network. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2990–3000, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wei-Fan Chen, Khalid Al Khatib, Henning Wachsmuth, and Benno Stein. 2020. Analyzing political bias and unfairness in news articles at different levels of granularity. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 149–154, Online. Association for Computational Linguistics.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. STANDER: An expert-annotated dataset for news stance detection and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4086–4101, Online. Association for Computational Linguistics.

Iain J. Cruickshank and Lynnette Hui Xian Ng. 2023. Prompting and fine-tuning open-sourced large language models for stance classification. *arXiv preprint arXiv:2309.13734*.

Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. 2020. Unsupervised user stance detection on twitter. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 141–152.

James N Druckman and Michael Parkin. 2005. The impact of media bias: How editorial slant affects voters. *The Journal of Politics*, 67(4):1030–1049.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*, pages 3988–3994. International Joint Conferences on Artificial Intelligence.

Kayla Duskin, Joseph S Schafer, Jevin D West, and Emma S Spiro. 2024. Echo chambers in the age of algorithms: An audit of twitter's friend recommender system. In *Proceedings of the 16th ACM Web Science Conference*, WEBSCI '24, page 11–21, New York, NY, USA. Association for Computing Machinery.

Marc Evrard, Rémi Uro, Nicolas Hervé, and Béatrice Mazoyer. 2020. French tweet corpus for automatic stance detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6317–6322, Marseille, France. European Language Resources Association.

Chong Feng, Muzammil Khan, Arif Ur Rahman, and Arshad Ahmad. 2020. News recommendation systems-accomplishments, challenges & future directions. *IEEE Access*, 8:16702–16725.

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*. ACL.

Seth Flaxman, Sharad Goel, and Justin M Rao. 2016. Filter bubbles, echo chambers, and online news consumption. *Public opinion quarterly*, 80(S1):298–320.

Joseph Gatto, Omar Sharif, and Sarah Preum. 2023. Chain-of-thought embeddings for stance detection on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4154–4161, Singapore. Association for Computational Linguistics.

Matthew Gentzkow and Jesse M Shapiro. 2010. What drives media slant? evidence from us daily newspapers. In *Econometrica*, volume 78, pages 35–71. Wiley Online Library.

Bilal Ghanem, Paolo Rosso, and Francisco Rangel. 2018. Stance detection in fake news a combined feature representation. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. Association for Computational Linguistics.

gogamza. 2023. kobart-base-v2. `https://huggingface.co/gogamza/kobart-base-v2`. Accessed: 2025-05-14.

Jiyoung Han and Christopher M Federico. 2018. The polarizing effect of news framing: Comparing the mediating roles of motivated reasoning, self-stereotyping, and intergroup animus. *Journal of Communication*, 68(4):685–711.

Jiyoung Han, Youngin Lee, Junbum Lee, and Meeyoung Cha. 2023. News comment sections and online echo chambers: The ideological alignment between partisan news stories and their user comments. *Journalism*, 24(8):1836–1856.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77, Dublin, Ireland. Association for Computational Linguistics.

Jiwoo Hong, Yejin Cho, Jiyoung Han, Jaemin Jung, and James Thorne. 2023. Disentangling structure and style: Political bias detection in news by inducing document hierarchy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5664–5686, Singapore. Association for Computational Linguistics.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Dong-si Jo. 2003. Hyeon yi-seop, president of media today - confident in achieving a profitable first year with page increases and the internet. *Newspaper and Broadcasting*, pages 86–87.

Bill Kovach and Tom Rosenstiel. 2021. *The elements of journalism, revised and updated 4th edition: What newspeople should know and the public should expect*. Crown.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Alapan Kuila, Somnath Jena, Sudeshna Sarkar, and Partha Pratim Chakrabarti. 2024. "news has various shades": Quantifying and analyzing media bias at aspect-level granularity. *ACM Journal on Computing and Sustainable Societies*, 2(4):1–51.

Haewoon Kwak, Jisun An, and Yong-Yeol Ahn. 2020. A systematic media frame analysis of 1.5 million new york times articles from 2000 to 2017. In *Proceedings of the 12th acm conference on web science*, pages 305–314.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2023. Stance detection with collaborative role-infused llm-based agents. *arXiv preprint arXiv:2310.10467*.

LG AI Research. 2024. Exaone 3.5: Series of large language models for real-world use cases. *arXiv preprint arXiv:2412.04862*.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717, Singapore. Association for Computational Linguistics.

Yingjie Li and Cornelia Caragea. 2021. Target-aware data augmentation for stance detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1850–1860, Online. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2738–2747, New York, NY, USA. Association for Computing Machinery.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024a. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

Luyang Lin, Lingzhi Wang, Xiaoyan Zhao, Jing Li, and Kam-Fai Wong. 2024b. IndiVec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1038–1050, St. Julian's, Malta. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yujian Liu, Xinliang Frederick Zhang, David Wegsman, Nicholas Beauchamp, and Lu Wang. 2022. POLITICS: Pretraining with same-story article comparison for ideology prediction and stance detection. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1354–1374, Seattle, United States. Association for Computational Linguistics.

Lauri Lüüsi, Uku Kangur, Roshni Chakraborty, and Rajesh Sharma. 2024. Political stance detection in Estonian news media. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 12–28, Helsinki, Finland. Association for Computational Linguistics.

Bertin Martens, Luis Aguiar, Estrella Gomez-Herrera, and Frank Mueller-Langer. 2018. The digital transformation of news media and the rise of disinformation and fake news.

Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. 2021. Stance detection in German news articles. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 66–77, Dominican Republic. Association for Computational Linguistics.

Matthew S McGlone. 2005. Quoted out of context: Contextomy and its consequences. *Journal of Communication*, 55(2):330–346.

Karen McIntyre. 2019. Solutions journalism: The effects of including solution information in news stories about social problems. *Journalism Practice*, 13(1):16–34.

Melvin Mencher and Wendy P Shilton. 1997. *News reporting and writing*. Brown & Benchmark Publishers Madison, WI.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2024. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *Plos one*, 19(4):e0302380.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Thomas E Nelson, Rosalee A Clawson, and Zoe M Oxley. 1997. Media framing of a civil liberties conflict and its effect on tolerance. *American Political Science Review*, 91(3):567–583.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, et al. 2021. Klue: Korean language understanding evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge. `http://www.fakenewschallenge.org/`. Accessed: 2025-04-14.

Myrthe Reuver, Suzan Verberne, and Antske Fokkens. 2024. Investigating the robustness of modelling decisions for few-shot cross-topic stance detection: A preregistered study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9245–9260, Torino, Italia. ELRA and ICCL.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7698–7716, Online. Association for Computational Linguistics.

Jonathon P Schuldt, Sara H Konrath, and Norbert Schwarz. 2011. "global warming" or "climate change"? whether the planet is warming depends on question wording. *Public opinion quarterly*, 75(1):115–124.

Adam F Simon and Jennifer Jerit. 2007. Toward a theory relating political discourse, media, and public opinion. *Journal of communication*, 57(2):254–271.

Seonyeong Song, Hyeonho Song, Kunwoo Park, Jiyoung Han, and Meeyoung Cha. 2023. Detecting contextomized quotes in news headlines by contrastive learning. *arXiv preprint arXiv:2302.04465*.

Yonghoi Song. 2007. Internet news media and issue development: A case study on the roles of independent online news services as agenda-builders for anti-us protests in south korea. *new media & society*, 9(1):71–92.

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.

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, and Xing Xie. 2023. Personalized news recommendation: Methods and challenges. *ACM Trans. Inf. Syst.*, 41(1).

Seunghyun Yoon, Kunwoo Park, Joongbo Shin, Hongjun Lim, Seungpil Won, Meeyoung Cha, and Kyomin Jung. 2019. Detecting incongruity between news headline and body text via a deep hierarchical encoder. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 791–800.

Bowen Zhang, Genan Dai, Fuqiang Niu, Nan Yin, Xiaomao Fan, Senzhang Wang, Xiaochun Cao, and Hu Huang. 2024a. A survey of stance detection on social media: New directions and perspectives. *arXiv preprint arXiv:2409.15690*.

Ruike Zhang, Yuan Tian, Penghui Wei, Daniel Dajun Zeng, and Wenji Mao. 2024b. An LLM-enabled knowledge elicitation and retrieval framework for zero-shot cross-lingual stance identification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12253–12266, Miami, Florida, USA. Association for Computational Linguistics.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024c. LLM-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378, Mexico City, Mexico. Association for Computational Linguistics.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

Lynn M Zoch and Judy VanSlyke Turk. 1998. Women making news: Gender as a variable in source selection and use. *Journalism & Mass Communication Quarterly*, 75(4):762–775.

Segment	Supportive (%)	Neutral (%)	Oppositional (%)
Headline	24	45	31
Lead	22.5	49.1	28.4
Conclusion	27	43.1	29.9
Quotations	29.5	38.6	31.9
Article	31.6	35.1	33.2

Table A1: Stance label distribution

A Appendix

A.1 Dataset Details

Label distribution Table A1 summaizes the distribution of stance labels at both article and segment levels. While the article-level stance labels are relatively balanced across classes, neutral stances appear more frequently at the segment level. Figure A1 visualizes the relationship between segment-level and article-level stance labels using Cramer's V. Article-level stance shows strong associations with the stances expressed in the headline, lead, and conclusion, each yielding Cramer's V values of approximately 0.7. In contrast, stance labels derived from quotations exhibit weaker associations, with Cramer's V values around 0.3. We also observe a strong correlation between the headline and lead stances, suggesting a shared rheotorical framing established early in the article.

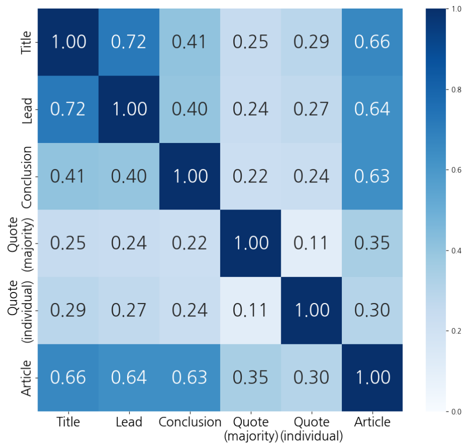


Figure A1: Stance label associations

List of target issues and media outlets Table A2 presents a comprehensive list of the target issues in K-NEWS-STANCE. The dataset contains articles from 31 media outlets, including the following top-10 news agencies: Kyunghyang Shinmun (경향신문), Segye Ilbo (세계일보), Korea JoongAng Daily (중앙일보), Kookmin Ilbo (국민일보), Seoul Shinmun (서울신문), Chosun Daily (조선

일보), Seoul Economic Daily (서울경제), Korea Economic Daily (한국경제), MoneyToday (머니투데이), and Hankook Ilbo (한국일보). Table A3 provides the list of target issues used in the newly collected dataset for two case studies.

A.2 Experimental Setups

For evaluation, we use macro F1 and accuracy, which are standard metrics for multi-class classification. We report the average performance over ten runs, along with the standard error, by varying the random seed from 42 to 51. The training split is used to train the fine-tuned baseline models and the segment-level stance prediction agents. Few-shot samples are selected from the training set via similarity search, using KLUE-RoBERTa-large as the dense retriever.

Experiments were conducted using a machine equipped with three Nvidia RTX A6000 GPUs (48GB per each) and 128GB of RAM. All experiments were run in a software environment configured with Python 3.9.19, PyTorch 2.5.1, Transformers 4.52.0, and vLLM 0.8.5. For the RoBERTa-based models—including the segment-level stance agent and three fine-tuned baselines (RoBERTa, CoT Embeddings, PT-HCL)—we used KLUE-RoBERTa-large, a pretrained checkpoint trained on a Korean corpus (Park et al., 2021). Based on validation experiments, we set the learning rate as 3×10^{-5}, with a batch size of 32 for CoT Embeddings and 16 for all other models. AdamW was used for optimizer and froze the bottom seven layers. GPT-4o-mini was used for CoT Embeddings and LKI-BART. We employed KoBART-base-v2 (gogamza, 2023) for LKI-BART. For training KoBART, we set the learning rate as 3×10^{-5}, batch size of 16, and used AdamW optimizer. We accessed GPT-4o-mini, Claude-3-haiku, and Gemini-2.0-flash via API. We set the temperature as 1.0, and max tokens as 1000 for chain-of-thought prompting and 100 for others for all LLM API calls. For the full fine-tuning of Exaone-3.5-2.4B, we used the AdamW optimizer with a learning rate of 5e-5, weight decay of 0.01, and 100 warmup steps. Training was conducted for 10 epochs with a per-device batch size of 6 for both training and evaluation. We used $\lambda = 0.3$ as the diversity hyperparameter in MMR. The hyperparameters were selected based on the settings reported in the original studies that introduced these methods.

The model ids and parameter sizes used in the experiments are provided below.

- GPT-4o-mini: `gpt-4o-mini-2024-07-18` (Parameter size: unknown)
- Claude-3-haiku: `claude-3-haiku-20240307` (Parameter size: unknown)
- Gemini-2.0-flash: `gemini-2.0-flash` (Parameter size: unknown)
- EXAONE-3.5-2.4b: https://huggingface.co/LGAI-EXAONE/EXAONE-3.5-2.4B-Instruct (Parameter size: 2.14B)
- Klue-RoBERTa-large: https://huggingface.co/klue/roberta-large (Parameter size: 337M)
- Klue-RoBERTa-base: https://huggingface.co/klue/roberta-base (Parameter size: 111M)
- KoBART-base-v2: https://huggingface.co/gogamza/kobart-base-v2 (Parameter size: 124M)
- mContriever: https://huggingface.co/facebook/mcontriever (Parameter size: 178M)

A.3 Used Prompts

In Figure A2 and A3, we present the English-translated prompt in English and its original Korean version, respectively, used to prompt an LLM for article-level stance detection. For training the RoBERTa model used as the segment-level stance prediction agent, we used the following input template: `[CLS] issue [SEP] segment [SEP]`.

A.4 Supplementary Results

Using an LLM as the segment-level agent Table A4 presents the performance of two additional variants of SAAS for predicting the overall stance of news articles. Specifically, these variants use Gemini-2.0-flash as the segment-level stance detection agent, with the first using an instruction-only prompt and the second incorporating six-shot samples. Gemini-2.0-flash was selected due to its strong performance on segment-level stance detection (Table A5). We present the best performance by SAAS(RoBERTa) by varying the LLM backbones for article-level stance detection for reference.

The results show that SAAS with a fine-tuned RoBERTa agent generally outperforms the LLM-based variant, with the sole exception being GPT-4o-mini, where the LLM-based approach performs comparably. Since the best performance across different LLM backbones is achieved by the RoBERTa-based variant, we report its results in Table 3, demonstrating SAAS's effectiveness over baseline method. Nevertheless, given that LLM-based segment agent require few or no labeled examples, the competitive performance of SAAS using LLM agents highlights its potential for zero- and few-shot stance scenarios.

Segment-level stance prediction We evaluate the performance of language model agents inn predicting stance labels for individual news segments. Specifically, we compare three approaches: fine-tuning a MLM, zero-shot inference with an LLM, and six-shot in-context learning with an LLM. For MLM fine-tuning and few-shot selection, we use the segment-level stance labels and corresponding news text from the training split of K-NEWS-STANCE. Table A5a reports the accuracy and macro F1 scores for eight models.

We find that LLMs generally outperform fine-tuned RoBERTa models. Given that LLMs require few or no labeled examples, these results highlights their effectiveness for stance detection in short texts. However, despite their strong performance at the segment level, SAAS with an LLM agent underperforms in article-level stance detection compared to the variant using the RoBERTa agent, as observed in Table A4.

To better understand this discrepancy, we analyze class-wise performance across segment types, as shown in Tables A5b to A5e. We observe that the RoBERTa model generally performs better in classifying *neutral*-labeled segments, as reflected in higher F1 scores for the *neutral* class. We hypothesize that accurately identifying neutral segments is a key factor contributing to the effectiveness of a segment-level stance agent.

Effects of journalism-guided segments Table A6 present the results of an ablation experiment to understand the impact of journalism-guided segments, adopted in SAAS. The first row indicates the best performance of the proposed method using Gemini 2.0 flash as a backbone LLM with CoT prompting and six-shot sample augmentation. The second is the performance by its counterpart, where randomly selected sentences become the target of label augmentation. We sample the sentences without replacement until it reaches the total segment length to control for the length effect. Results show decreases in performance with an accuracy of 0.029 and F1 of 0.027, demonstrating the advantage of adopting journalism-guided segments.

Qualitative error analysis From a qualitative analysis of incorrect predictions made by SAAS using RoBERTa—our best-performing model—we identified two primary error patterns.

The first stems from the model's failure to interpret positive descriptions as indicative of a supportive stance, often resulting in the misclassification of supportive articles as neutral. This is the most prominent error type observed in the quantitative error analysis (Figure 2). For example, a news article expresses a favorable view on the issue of "Han Dong-hoon: Cutting National Assembly Seats to 250 is on the Table" by outlining the benefits of the policy proposed by Han. However, the segment-level agent fails to capture this supportive framing, which may subsequently cause the LLM to predict a neutral stance.

The second error pattern arises during the orchestration of segment-level stance labels. Even when segment-level predictions are accurate, as simulated in the oracle setting, the LLM sometimes fails to infer the correct overall stance. This issue is especially pronounced in articles that contain multiple quotations expressing divergent or conflicting viewpoints.

These two patterns point to potential future directions for improving the article-level stance detection: enhancing the segment-level detection model and selectively incorporating the most salient quotations, rather than considering all of them for label augmentation.

Date	Issue (in Korean)	Issue (in English)
2022-06-15	화물연대 파업 8일만에 철회, 안전운임제는 지속 추진	Truckers' Strike Ends After 8 Days, Safety Freight Rates to Continue
2022-06-16	내년 최저임금 업종 구분 없이 동일 적용	Next Year's Minimum Wage to be Applied Uniformly Regardless of Industry
2022-06-22	尹정부 탈원전 폐기 공식화	Yoon Administration Officially Ends Nuclear Phase-Out Policy
2022-07-17	이재명 당대표 선거 출마 선언	Lee Jae-myung Announces Candidacy for Party Leadership Election
2022-07-31	초등 입학연령 하향 추진	Push for Lowering Elementary School Entry Age
2022-08-10	박민영 대통령실 청년대변인 발탁	Park Min-young Appointed as Presidential Spokesperson for Youth Affairs
2022-10-06	정부 여성가족부 폐지 정부조직개편안 확정	Government Finalizes Plan to Abolish Ministry of Gender Equality and Family
2022-10-11	정부 학업성취도 자율평가 확대 추진	Government Expands Voluntary Academic Achievement Assessment
2022-10-26	법무부, 촉법소년 기준 만 13세로 1년 하향	Ministry of Justice Lowers Age Threshold for Juvenile Offenders to 13
2022-11-07	文 대통령, 키우던 풍산개 2마리 정부 반환	Moon Jae-in Hands Over Two Pungsan Dogs to the State
2022-12-21	정부, 다주택자 부동산 규제 대폭 완화	Government Eases Real Estate Regulations for Multiple Homeowners
2022-12-21	윤석열 대통령 "노조부패, 척결해야할 3대 부패"	President Yoon Calls "Union Corruption is One of Three Major Evils to Be Eradicated"
2023-01-12	정부 '강제징용 배상 일본 기업 대신 지급' 공식화	Korea Announces Plan to Compensate Forced Labor Victims On Behalf of Japanese Companies
2023-02-15	노란봉투법 국회 상임위 통과	Labor-Friendly 'Yellow Envelope Act' Passes National Assembly Standing Committee
2023-02-21	법원 동성 부부 배우자도 건강보험 피부양자 인정	Court Recognizes Same-Sex Partners as Legal Dependents for Health Insurance
2023-03-06	주 52시간 근로시간 개편 최대 69시간 가능	Government Proposes Overhaul of 52-Hour Workweek, Allowing Up to 69 Hours
2023-03-23	헌재 '검수완박' 절차는 위헌, 법안은 유효	Constitutional Court: Prosecutorial Reform Law Upheld, Procedure Unconstitutional
2023-03-23	'양곡관리법 개정안' 민주당 주도 본회의 통과	Democratic Party Leads Passage of Revised Grain Management Act in National Assembly Plenary Session
2023-04-05	학폭 가해기록 보존기간 연장 · 입시반영 검토	School Bullying Records May Be Kept Longer, Reflected in College Admissions
2023-05-22	등록 재산에 가상자산 포함법안 통과	Bill Passed to Include Cryptocurrency in Public Asset Declarations
2023-06-18	당정, 중대 범죄자 신상공개 특별법 추진	Ruling Party and Government Push Special Bill to Disclose Identities of Serious Criminals
2023-06-28	'출생통보제' 법사위 소위 통과	'Birth Notification System Bill' Advances in Parliament
2023-07-04	IAEA 보고서 일본 오염수 방류 문제 없다	IAEA Report Says Japan's Fukushima Water Release Poses No Safety Concerns
2023-07-12	정부 · 여당, 실업급여 하한액 낮추거나 폐지까지 검토	Government and Ruling Party Consider Lowering or Abolishing Minimum Unemployment Benefit
2023-07-20	환경부 4대강 16개보 모두 존치	All 16 Weirs on Four Rivers to Remain Intact, Says Environment Ministry
2023-07-31	정부, 외국인 가사도우미 100여명 시범 도입	Seoul to Test Foreign Domestic Worker Program with Initial 100 Hires
2023-08-22	새 대법원장 후보에 이균용 서울고법 부장판사	Lee Kyun-yong, Nominated as New Chief Justice Candidate
2023-08-23	한덕수 총리 "흉악 범죄 예방 위해 의경 재도입 검토"	Prime Minister Han Duck-soo Considers Reintroducing Conscripted Police to Prevent Heinous Crimes
2023-09-26	헌재 '대북 전단 금지법' 위헌 결정	Constitutional Court Invalidates Law Banning Leaflets to North Korea
2023-10-12	검찰 이재명 '백현동 배임혐의' 불구속 기소	Prosecution Indicts Lee Jae-myung Without Detention over Baekhyeon-dong Scandal
2023-11-02	국민의힘 '김포 서울 편입' 특위 발족	PPP Launches Special Committee for Gimpo-Seoul Integration
2023-11-05	내년 6월까지 공매도 전면 금지	Short Selling Fully Banned Until June 2025
2023-12-07	민주당, 전당대회서 권리당원 표 비중 확대	DP Boosts Role of Rank-and-File Members in Party Convention Votes
2023-12-21	금리 4% 넘는 자영업자 최대 300만 원 환급	Business Owners Paying Over 4% Interest Eligible for Government Refunds
2023-12-27	이준석 국민의힘 탈당 후 신당 창당 돌입	Lee Jun-seok Leaves PPP and Begins Forming New Party
2024-01-16	한동훈 "국회의원 250명으로 줄이겠다"	Han Dong-hoon: Cutting National Assembly Seats to 250 is on the Table
2024-01-29	이준석 "경찰 · 소방관 되려는 여성 군 복무해야"	Lee Jun-seok: Women Should Serve in Military to Join Police or Firefighting Forces
2024-01-09	'개 식용 금지법' 국회 본회의 통과	National Assembly Passes Bill Banning Dog Meat Consumption
2024-02-06	내년 의대 정원 2천 명 증원	Medical School Admissions to Increase by 2,000 Next Year
2024-02-13	조국 전 법무부 장관 신당 창당 선언	Former Justice Minister Cho Kuk Declares Launch of New Party
2024-03-27	한동훈 국회 세종시로 완전 이전 공약 발표	Justice Minister Han Dong-hoon Proposes Relocating the National Assembly to Sejong City
2024-04-24	한강에서 먹고 자고 일한다 '리버시티 서울' 발표	Seoul Announces 'River City' Project to Create Living, Working, and Leisure Spaces on the Han River
2024-04-10	경기 화성을 개혁신당 이준석 대표 당선	Reform Party Leader Lee Jun-seok Wins Parliamentary Seat in Hwaseong, Gyeonggi
2024-05-08	'외국 면허 의사' 국내 진료 허용	Foreign-Licensed Doctors Allowed to Practice in Korea
2024-05-30	법원 하이브 '민희진 해임' 제동	Court Rejects Hive's Attempt to Dismiss Min Hee-jin
2024-06-03	윤석열 대통령 "동해 140억배럴 석유 · 가스 매장 추정"	President Yoon: East Sea May Hold 14 Billion Barrels of Oil and Gas

Table A2: A comprehensive list of target issues in K-NEWS-STANCE

Study	Date	Issue (in Korean)	Issue (in English)
Recommendation	2024-07-21	군, 북한 오물풍선 잇단 살포에 대북 방송 전면 시행	Military to Resume Full-Scale Propaganda Broadcasts to North Following Repeated Trash Balloon Incidents
	2024-08-12	윤석열 대통령, 검찰총장 후보자 심우정 법무장관 지명	President Yoon Nominates Vice Justice Minister Shim Woo-jung for Prosecutor General
	2024-12-03	사상초유 감사원장 탄핵안 내일 표결	Unprecedented Impeachment Motion Against Board of Audit and Inspection Chairman Faces Vote Tomorrow
	2024-12-27	민주, 한 권한대행 탄핵안 발의...27일 표결	Minjoo Party Files Impeachment Motion Against Acting Prime Minister Han; Vote Set for 27th
Media bias	2024-12-26	한덕수 권한대행 탄핵안 본회의 통과	Acting PM Han Duck-soo Impeachment Passes Parliament
	2025-04-30	대법원, 이재명 대선후보 공직선거법 사건 파기환송	Supreme Court Remands Lee Jae-myung's Election Law Violation Case

Table A3: List of target issues covered in the two case studies.

Method	+ CoT	+6-shot	GPT-4o-mini		Gemini-2.0-flash		Claude-3-haiku		Exaone-2.4b (finetuned)	
			Accuracy	F1	Accuracy	F1	Accuracy	F1	Accuracy	F1
SAAS (RoBERTa)			0.607±0.006	0.602±0.007	**0.678±0.002**	**0.672±0.002**	0.639±0.002	0.638±0.004	0.601±0.006	0.599±0.006
SAAS (Gemini-2.0-flash)			0.545±0.002	0.549±0.002	0.572±0.001	0.549±0.001	0.55±0.004	0.526±0.006	0.536±0.002	0.529±0.001
	√		0.56±0.001	0.538±0.002	0.598±0.001	**0.583±0.002**	0.56±0.004	0.531±0.005	**0.541±0.002**	**0.536±0.002**
		√	0.558±0.003	0.539±0.004	**0.597±0.004**	0.581±0.005	**0.596±0.004**	**0.584±0.004**	0.325±0.006	0.31±0.007
	√	√	**0.572±0.002**	**0.554±0.002**	0.592±0.003	0.576±0.003	0.589±0.002	0.579±0.003	0.329±0.002	0.304±0.001
SAAS (Gemini-2.0-flash w/ 6-shot)			0.548±0.016	0.553±0.016	0.609±0.001	0.587±0.001	0.575±0.001	0.559±0.001	**0.566±0.001**	**0.575±0.001**
	√		0.583±0.008	0.571±0.009	0.625±0.003	0.613±0.004	0.573±0.004	0.547±0.003	0.558±0.002	0.556±0.002
		√	0.608±0.004	0.598±0.004	0.617±0.003	0.605±0.003	**0.599±0.004**	**0.59±0.004**	0.343±0.005	0.322±0.004
	√	√	**0.611±0.003**	**0.602±0.004**	**0.63±0.003**	**0.617±0.003**	0.592±0.006	0.587±0.007	0.331±0.004	0.3±0.003

Table A4: Performance for predicting overall stance of news articles by SAAS with LLM agents

[System Prompt]
Stance detection is the task of determining the expressed or implied opinion, or stance, of a statement toward a certain, specified target. You are given an issue and a news article about that issue. Your task is to classify the article's stance toward the given issue as one of the following: supportive, neutral, or oppositional.

The criteria for each label are as follows:
- Supportive: The article shows a favorable tone toward the issue, emphasizes quotes in support of the issue, and predominantly uses positive or optimistic language.
- Neutral: The article maintains an objective tone, balances quotes from both supportive and critical perspectives, and uses neutral language.
- Oppositional: The article shows a skeptical tone toward the issue,
emphasizes quotes that criticize the issue, and predominantly uses negative or pessimistic language.

Additional information is provided on the stance of the headline, lead, conclusion, and quotes regarding the issue. Each segment is marked with XML tags, and the final stance should be determined by taking into account the detailed stance labels of each part.

[User Prompt]
Issue: *Government confirms organizational restructuring plan to abolish Ministry of Gender Equality and Family*
Headline: *<Headline stance="Oppositional">MOGEF downgraded to a department... Concerns "Gender equality policies will be buried in the giant MOHW"</Headline>*
Article: *<Lead stance="Oppositional">Under the government restructuring plan announced on the 6th, the Ministry of Gender Equality and Family (MOGEF) faces demotion to a department under the MOHW after 21 years as an independent ministry. The government emphasizes that MOGEF's functions will be retained and may create synergy with the MOHW's welfare policy capabilities. Even experts who support the reorganization question whether the enormous MOHW can respond quickly to gender equality issues.</Lead>*
MOGEF highlights that integrating its youth policies with MOHW's child welfare functions can yield synergistic effects. On this day, Minister Kim Hyunsook announced a 'Support Plan for In- and Out-of-School Youth' and said, <Quotation stance="Neutral">If we became a well-authorized department under the MOHW, we could have included more in today's announcement</Quotation> According to the Ministry of the Interior and Safety's restructuring plan, functions like support for career-interrupted women will be transferred to the Ministry of Employment and Labor, The four core functions—youth, family, women and gender equality, and rights (e.g., support for victims of sexual/domestic violence)—will be transferred to the Population, Family, and Gender Equality Bureau under the MOHW. Some argue that organizations led by ministers and those led by department heads have significantly different authority within government. Huh Min-sook, a legislative researcher at the National Assembly, said, <Quotation stance="Oppositional">MOGEF already lacked budget and authority, making cooperation difficult — demoting it will only weaken it further</Quotation>
There are concerns that the control tower responsible for formulating gender equality policies and overseeing their implementation across all government agencies will disappear.Park Sun-young, senior researcher at the Korean Women's Development Institute, stated, <Quotation stance="Oppositional">Gender equality policy is about coordination across all ministries — that's why MOGEF was created,</Quotation> and pointed out that <Quotation stance="Oppositional">putting it under the implementation-focused MOHW would undermine its effectiveness.</Quotation> Even experts who criticize MOGEF's performance say transferring and downsizing its functions to the MOHW would hinder gender equality policy. Jung Jae-hoon, professor at Seoul Women's University, noted <Quotation stance="Supportive">a "gender ghetto" phenomenon had emerged in which gender issues were discussed only among women within MOGEF</Quotation> and stated that <Quotation stance="Supportive">a presidential committee on gender equality should be established to elevate the issue to the level of the President's agenda</Quotation>. Hong Sung-geol, professor of public administration at Kookmin University, stated that <Quotation stance="Supportive">in the case of family policy, separating it from the MOHW's welfare agenda had led to fragmented policy momentum</Quotation>and viewed the restructuring positively. However, he also stated, <Quotation stance="Oppositional">A gender equality committee that can evaluate all ministries' policies is the ideal approach</Quotation>
<Conclusion Stance="Oppositional">Within MOGEF, concerns are rising that the policies will be treated as secondary if placed under the MOHW. One MOGEF official said, <Quotation stance="Oppositional">Our role is to protect those who cannot raise their voices, based on awareness of diversity and gender</Quotation>and added, <Quotation stance="Oppositional">These duties are bound to become insignificant within the vast MOHW</Quotation></Conclusion>

Figure A2: The English-translated prompt used in SAAS, shown with an illustrative input. Blue italic text highlights the user input.

[System Prompt]
입장 분류는 특정 대상에 대한 텍스트의 명시적 또는 묵시적인, 의견이나 입장을 결정하는 작업입니다.

이슈와 뉴스 기사가 제공되며, 당신의 임무는 주어진 이슈에 대한 뉴스 기사의 입장을 지지적, 중립적 혹은 비판적 중 하나로 분류하는 것입니다.

각 라벨의 판단 기준은 다음과 같습니다:
- 지지적: 이슈에 대해 호의적인 논조, 옹호하는 입장의 인용문을 중심으로 배치하며, 긍정적 · 낙관적 어조가 지배적인 경우
- 중립적: 이슈에 대해 객관적인 논조, 옹호하거나 비판하는 입장의 인용문을 균형 있게 배치하며, 중립적 어조를 사용하는 경우
- 비판적: 이슈에 대해 회의적인 논조, 비판하는 입장의 인용문을 중심으로 배치하며, 부정적 · 비관적 어조가 지배적인 경우

추가 정보로 이슈에 대한 제목, 도입부, 결론부, 직접인용구의 입장 정보가 각각 제공됩니다.
각 위치는 XML 태그로 표시되며, 세부 라벨 정보를 함께 고려하여 최종 입장을 결정하세요.

[User Prompt]
Issue: *정부 여성가족부 폐지 정부조직개편안 확정*
Headline: *<제목 입장="비판적">'본부'로 격하된 여가부..."공룡 복지부에서 성평등 정책 묻힐 것" 우려</제목>*
Article: *<도입부 입장="비판적"> 6일 발표된 정부조직 개편안에 따라 여성가족부가 출범 21년 만에 독립부처에서 보건복지부 산하 본부로 격하될 위기에 처했다. 정부는 여가부 기능은 유지되고, 보건복지부의 복지 정책 기능과 시너지를 낼 수 있다는 점을 강조한다. 그러나 여가부 개편에 찬성하는 전문가들도 '공룡 부처'가 되는 보건복지부가 성평등 문제에 기민하게 대처할 수 있을지 의문을 제기하고 있다.</도입부>*
여가부는 복지부의 아동복지 기능과 여가부의 청소년 정책 등이 통합되면 시너지 효과를 낼 수 있다는 점을 강조하고 있다. 김현숙 여가부 장관은 이날 '학교 안팎 청소년 지원 강화 대책'을 발표하며 <직접인용구 입장="중립적">복지부 산하의 상당한 권한을 가진 본부가 된다면 오늘 여가부가 발표한 내용보다 더 많은 내용이 담길 수 있었을 것</직접인용구>이라고 말했다. 행정안전부의 정부조직 개편안에 따르면 기존 여가부 기능 중 경력단절여성 지원 등 여성고용 기능은 고용노동부로 이관하고, △청소년 △가족 △여성 및 성평등 △권익(성폭력, 가정폭력 등 피해자 지원) 등 4대 기능은 복지부 산하 인구가족양성평등본부로 이관된다.
그러나 장관이 이끄는 조직과 본부장이 이끄는 조직이 정부 부처 내에서 갖는 위상이 다르다는 주장이 나온다. 허민숙 국회 입법조사관은<직접 인용구 입장="비판적">기존의 여가부도 예산과 권한이 작아 다른 부처와 협력하기 어려웠는데 본부로 격하시킨다면 힘을 더 빼는 것</직접 인용구>이라고 했다. 성평등 정책을 수립하고 정부조직 전반에서 성평등이 지켜지는지 점검할 컨트롤타워가 사라진다는 우려가 나온다. 박선영 한국여성정책연구원 선임연구위원은<직접 인용구 입장="비판적">양성평등 정책은 전 부처에 대한 조정 업무이고, 그래서 여성가족부가 만들어졌던 것</직접 인용구>이라며<직접 인용구 입장="비판적">집행 기능 중심의 복지부에 집어넣으면 실효성이 떨어질 것</직접 인용구>이라고 지적했다. 기존 여성가족부가 정책 기능을 제대로 수행하지 못했다고 평가하는 전문가들도 기능을 복지부로 축소 · 이관해선 성평등 정책을 펴기 어렵다고 지적했다. 정재훈 서울여대 사회복지학과 교수는<직접 인용구 입장="지지적">성차별 문제와 관련해 그동안 정부에선 여성가족부에서 여성끼리 모여 논의해도 된다는 '성게토화(Gender Ghetto)' 현상이 나타났다</직접 인용구>며<직접 인용구 입장="지지적">대통령 직속 성평등위원회를 설치해 성평등 문제를 대통령 의제로 끌어올리는 게 필요하다</직접 인용구>고 지적했다. 홍성걸 국민대 행정학과 교수는<직접 인용구 입장="지지적">가족정책의 경우 그동안 복지부의 복지 정책과 분리되면서 정책 동력이 분산되는 부정적인 측면이 있었다</직접 인용구>면서 여가부 개편의 시너지 효과를 긍정적으로 평가했다. 다만 홍 교수도 <직접 인용구 입장="비판적">성평등 정책은 전 부처의 모든 정책을 테스트할 수 있는 양성평등위원회를 만들어 맡기는 게 바람직하다</직접 인용구>고 말했다.
<결론부 입장="비판적">여가부 내부에서도 복지부 산하로 들어가면 정책들이 '곁가지 취급' 받을 거라는 우려가 나온다. 한 여가부 관계자는<직접 인용구 입장="비판적">여가부의 일은 다양성, 젠더에 대한 인식을 갖고서 스스로 목소리를 낼 수 없는 이들을 보호하는 것</직접 인용구>이라며<직접 인용구 입장="비판적">방대한 복지부로 갔을 때 이런 업무의 비중은 미미해질 수밖에 없을 것</직접 인용구>이라고 했다.</결론부>

Figure A3: The original Korean-language prompt used in SAAS, shown with an illustrative input. Blue italic text highlights the user input.

Type	Model	Accuracy	F1	F1 (*supportive*)	F1 (*neutral*)	F1 (*oppositional*)
Fine-tuned MLM	RoBERTa-base	0.582±0.004	0.559±0.006	0.442±0.017	0.649±0.005	0.587±0.011
	RoBERTa-large	**0.606±0.006**	**0.583±0.01**	**0.453±0.031**	**0.662±0.005**	**0.634±0.011**
LLM (zero-shot)	GPT-4o-mini	0.592 ± 0.002	0.578 ± 0.003	0.476 ± 0.001	0.570 ± 0.002	0.687 ± 0.004
	Gemini-2.0-flash	**0.647±0.004**	**0.638 ± 0.003**	**0.575±0.003**	**0.594±0.002**	**0.745±0.001**
	Claude-3-haiku	0.587 ±0.001	0.58±0.004	0.49 ±0.003	0.542±0.002	0.71 ±0.003
LLM (6-shot)	GPT-4o-mini	0.598±0.004	0.583±0.004	0.469 ±0.004	0.574±0.003	0.707±0.004
	Gemini-2.0-Flash	**0.671±0.004**	**0.664±0.004**	0.6±0.004	**0.623±0.003**	**0.771±0.004**
	Claude-3-haiku	0.637±0.002	0.636±0.002	**0.613±0.003**	0.534±0.003	0.761±0.001

(a) Performance on all segments

Type	Model	Accuracy	F1	F1 (*supportive*)	F1 (*neutral*)	F1 (*oppositional*)
Fine-tuned MLM	RoBERTa-base	0.636±0.005	0.576±0.008	0.428±0.021	0.719±0.005	0.581±0.011
	RoBERTa-large	**0.664±0.006**	**0.61±0.008**	**0.45±0.021**	**0.731±0.007**	**0.648±0.02**
LLM (zero-shot)	GPT-4o-mini	0.65±0.004	0.636±0.003	**0.577±0.004**	**0.683 ± 0.004**	0.647±0.001
	Gemini-2.0-flash	**0.658 ± 0.004**	**0.642 ± 0.005**	0.563 ± 0.002	0.68 ± 0.002	**0.682 ± 0.003**
	Claude-3-haiku	0.623 ± 0.004	0.597 ± 0.003	0.469 ± 0.003	0.668 ± 0.004	0.655 ± 0.004
LLM (6-shot)	GPT-4o-mini	0.656 ± 0.002	0.636 ± 0.004	0.554 ± 0.003	0.695 ± 0.003	0.659 ± 0.002
	Gemini-2.0-Flash	**0.702 ± 0.002**	**0.687 ± 0.004**	**0.603 ± 0.003**	**0.724 ± 0.003**	**0.733 ± 0.002**
	Claude-3-haiku	0.64 ± 0.003	0.635 ± 0.004	0.56 ± 0.001	0.619 ± 0.003	0.727 ± 0.001

(b) Performance on Headline

Type	Model	Accuracy	F1	F1 (*supportive*)	F1 (*neutral*)	F1 (*oppositional*)
Fine-tuned MLM	RoBERTa-base	0.675±0.007	**0.61±0.008**	0.465±0.014	0.752±0.008	**0.612±0.019**
	RoBERTa-large	**0.681±0.006**	0.608±0.012	**0.475±0.018**	**0.758±0.004**	0.592±0.029
LLM (zero-shot)	GPT-4o-mini	0.668 ± 0.005	0.62 ± 0.004	0.518 ± 0.004	**0.74 ± 0.002**	0.601 ± 0.003
	Gemini-2.0-flash	**0.677 ± 0.002**	**0.648 ± 0.002**	**0.554 ± 0.003**	0.727 ± 0.002	**0.662 ± 0.004**
	Claude-3-haiku	0.581 ± 0.003	0.548 ± 0.003	0.387 ± 0.002	0.641 ± 0.001	0.616 ± 0.003
LLM (6-shot)	GPT-4o-mini	**0.676 ± 0.004**	0.634 ± 0.002	0.499 ± 0.001	**0.741 ± 0.001**	**0.661 ± 0.005**
	Gemini-2.0-Flash	0.58 ± 0.003	0.575 ± 0.002	0.686 ± 0.002	0.541 ± 0.001	0.5 ± 0.003
	Claude-3-haiku	0.637±0.005	**0.636±0.004**	**0.761±0.007**	0.534±0.007	0.613±0.003

(c) Performance on Lead

Type	Model	Accuracy	F1	F1 (*supportive*)	F1 (*neutral*)	F1 (*oppositional*)
Fine-tuned MLM	RoBERTa-base	0.576±0.005	0.546±0.005	0.422±0.019	0.63±0.01	0.586±0.012
	RoBERTa-large	**0.594±0.006**	**0.559±0.01**	**0.431±0.025**	**0.65±0.005**	**0.597±0.018**
LLM (zero-shot)	GPT-4o-mini	0.634 ± 0.002	0.603 ± 0.005	0.447 ± 0.004	0.675 ± 0.004	0.686 ± 0.003
	Gemini-2.0-flash	**0.681 ± 0.003**	**0.661 ± 0.004**	**0.545 ± 0.004**	**0.703 ± 0.002**	**0.735 ± 0.004**
	Claude-3-haiku	0.592 ± 0.002	0.576 ± 0.003	0.457 ± 0.004	0.598 ± 0.002	0.671 ± 0.001
LLM (6-shot)	GPT-4o-mini	0.635 ± 0.005	0.602 ± 0.004	0.431 ± 0.002	**0.667 ± 0.005**	0.708 ± 0.004
	Gemini-2.0-Flash	**0.66 ± 0.005**	**0.65 ± 0.004**	**0.56 ± 0.002**	0.65 ± 0.005	**0.74 ± 0.004**
	Claude-3-haiku	0.6 ± 0.002	0.593 ± 0.005	0.56 ± 0.003	0.625 ± 0.002	0.593 ± 0.003

(d) Performance on Conclusion

Type	Model	Accuracy	F1	F1 (*supportive*)	F1 (*neutral*)	F1 (*oppositional*)
Fine-tuned MLM	RoBERTa-base	0.541±0.006	0.536±0.007	0.443±0.02	0.582±0.006	0.582±0.01
	RoBERTa-large	**0.571±0.01**	**0.563±0.013**	**0.452±0.037**	**0.591±0.006**	**0.647±0.006**
LLM (zero-shot)	GPT-4o-mini	0.57 ± 0.004	0.552 ± 0.002	0.465 ± 0.005	0.494 ± 0.005	0.697 ± 0.004
	Gemini-2.0-flash	**0.636 ± 0.002**	**0.622 ± 0.001**	**0.58 ± 0.004**	**0.524 ± 0.004**	**0.761 ± 0.004**
	Claude-3-haiku	0.582 ± 0.004	0.573 ± 0.003	0.504 ± 0.004	0.486 ± 0.002	0.729 ± 0.001
LLM (6-shot)	GPT-4o-mini	0.575 ± 0.003	0.557 ± 0.002	0.461 ± 0.002	0.494 ± 0.003	0.715 ± 0.002
	Gemini-2.0-Flash	**0.669 ± 0.003**	**0.658 ± 0.002**	0.608 ± 0.002	**0.581 ± 0.003**	0.786 ± 0.002
	Claude-3-haiku	0.649 ± 0.003	0.64 ± 0.003	**0.614 ± 0.005**	0.5 ± 0.003	**0.805 ± 0.003**

(e) Performance on Quotation

Table A5: Segment-level stance detection performance

Label augmentation	Accuracy	F1
For journalism-guided segments (i.e., SAAS)	0.678±0.002	0.672±0.002
For randomly selected segments	0.649±0.001	0.645±0.002

Table A6: Performance by ablating journalism-guided segments labels

Target Issue: '개 식용 금지법' 국회 본회의 통과

Headline (*Supportive*): "역시 김건희"... 개식용금지법 통과에 개딸들 이례적 환호
Body Text
-Lead (*Supportive*): '김건희법' 별칭인 개식용금지법 국회 통과
개 식용 금지 주장해온 김건희 여사에개딸들 환호 "정말 다행, 감사합니다"
식용을 목적으로 개를 도살하거나 사육 · 증식하는 것을 금지하는 법안이 국회 본회의를 통과했다. 김건희 여사의 꾸준한 노력 끝에 '김건희법'이라 불리는 법안이 통과되며 개딸들 사이에서는 이례적인 환호가 터져나왔다.
...
-Conclusion (*Supportive*): 코리아리서치인터내셔널이 동물복지문제연구소 어웨어 의뢰로 전국 성인남녀 2000명을 대상으로 진행한 '2023 개 식용에 대한 국민인식조사'에 따르면 응답자의 94.5%는 지난 1년 동안 개고기를 먹은 적이 없다고 응답했다.
Overall Stance: *Supportive*

Headline (*Neutral*): '개 식용 금지법' 통과에 "기념비적 역사" vs "헌법소원 낼 것"
Body Text
-Lead (*Neutral*): 동물단체-육견협회, 엇갈린 반응
"동물권 승리" vs "먹을 권리 강탈"
이른바 '개 식용 금지법'이 9일 국회를 통과하자 동물단체들은 "기념비적인 역사가 쓰였다"며 일제히 환영했다. 반면 이를 줄곧 반대해 온 대한육견협회는 "직업선택의 자유를 빼앗았다"며 헌법소원을 내겠다는 뜻을 밝혔다.
...
-Conclusion (*Neutral*): 이날 국회 본회의에서 의결된 '개의 식용 목적의 사육 · 도살 및 유통 등 종식에 관한 특별법'은 식용 목적의 개 도살 · 사육 · 증식, 개나 개를 원료로 한 식품의 유통 · 판매를 금지하는 것을 골자로 한다.
Overall Stance: *Neutral*

Headline (*Oppositional*): 개 식용 금지법 통과... "20년 보신탕 팔았는데 살길 막막"
Body Text
-Lead (*Oppositional*): 관련업주 당혹, 2027년부터 처벌
처벌 수위 · 적정성 놓고 논란 제기
사육 · 유기견 급증 해결 급선무
개 식용 금지법이 9일 국회 본회의를 통과하면서 개 식용 논쟁이 다시 불거졌다. 정부는 이번 법안을 통해 더이상의 논란을 막겠다는 입장이지만 처벌 수위, 처벌의 적정성을 두고 또 다른 논란이 제기되고 있다.
...
-Conclusion (*Oppositional*): 춘천에서 20년 간 영양탕집을 운영한 A씨는 "20년 간 장사했는데 금지되면 어떤 업종을 해야 될 지 막막하다" "더 한 것도 먹으면서 왜 갑자기 보신탕을 금지시키느냐. 자꾸 우리를 공격하니까 답답하다"고 했다. 또다른 업주 B씨는 "작업장이 싹 없어져 고기 구할 곳도 없어 문을 닫고 쉰 적도 있다"며 "임대료가 한 달에 160만원 나가서 하루에 80만원은 팔아야 남는데 요즘 같은 경우는 20만원 겨우 팔고 있어 월세도 못 내는 형편"이라고 토로했다. 법안 통과로 인해 처벌 수위와 적정성에 대한 논란이 재점화되었고, 사육 중인 개의 급증과 유기견 문제 해결도 시급하다는 지적이 나왔다.
Overall Stance: *Oppositional*

Table A7: The original example in Korean. The colored highlight indicates the stance label for quotations (blue: supportive, red: oppositional).

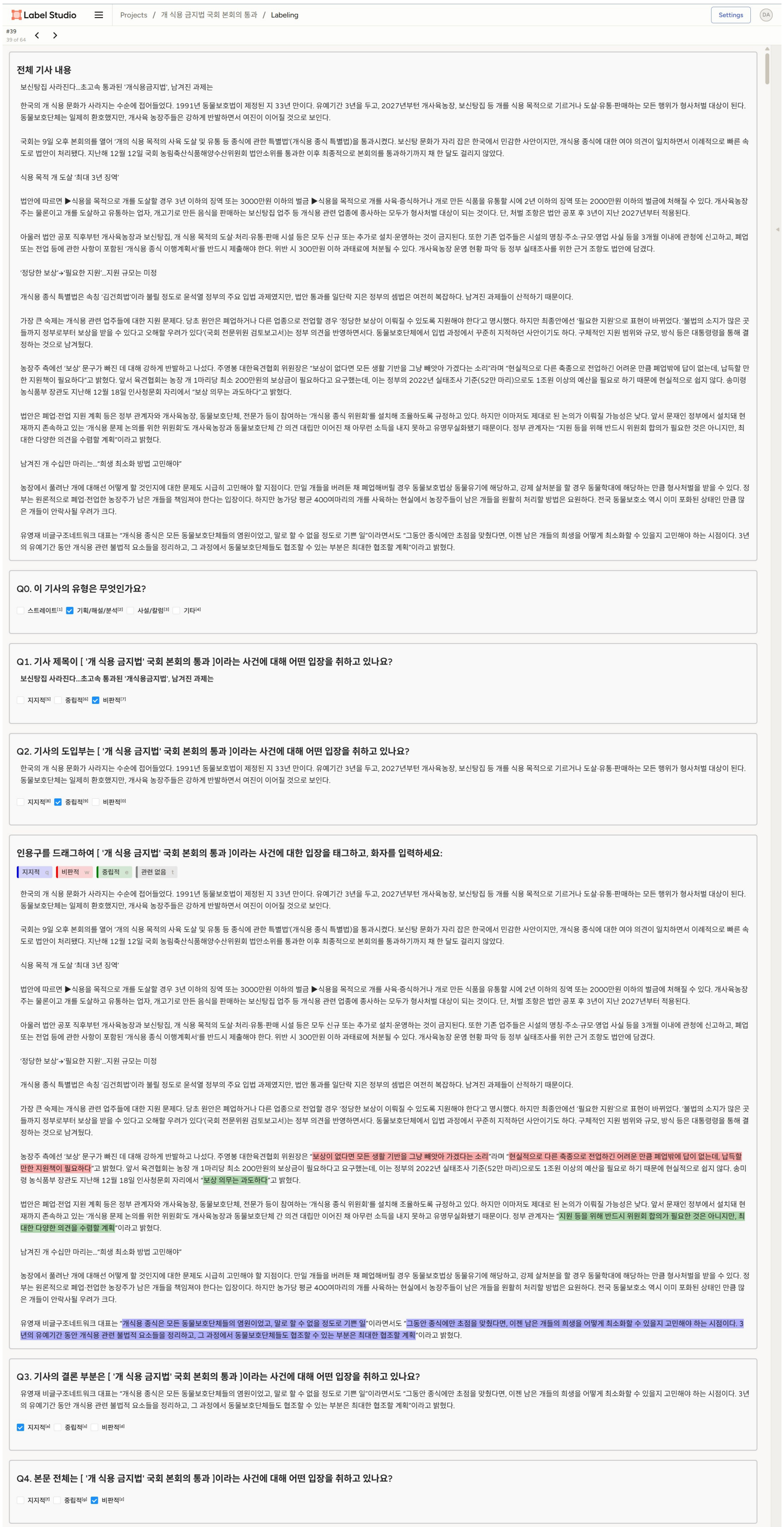


Figure A4: Labeling interface