ISSN 2093-0542

ICONI

KOREAN SOCIETY FOR INTERNET INFORMATION

The 17th International Conference on Internet

(ICONI 2025)

Dec. 14-17, 2025 Okinawa Convention Center, Okinawa, Japan

http://www.iconi.org

Proceedings of ICONI 2025

| Organized by |

Korean Society for Internet Information (KSII)

| Sponsored by |

OCVB

Contents

Enhancing Reasoning Performance of LLMs based on Metacognition with Naïve Evaluation Results

Jinmo Yang, Jaeho Kim, Hyeon-uk Jeong, Kidu Kim, and R. Young Chul Kim*
SE Lab., Hongik University
Sejong, South Korea
[e-mail: {yjmd2222, jaehokim1005}@g.hongik.ac.kr,
junghyunwook@lotte.net, kdkim@tta.or.kr, bob@hongik.ac.kr]
*Corresponding author: R. Young Chul Kim

Abstract

Nowadays, many large language models (LLMs) are equipped with a reasoning mode, enabling the model to reason from the given input and produce a more accurate output with an accompanying explanation. We have developed the reasoning content for the Korean Language Understanding Evaluation (KLUE) topic classification (TC) benchmark dataset, which consists of pure text-label pairs, for training an LLM, aiming to achieve better results than those obtained in non-reasoning mode. However, the F1-score in the reasoning mode was 0.662, much below that in the non-reasoning mode of 0.848. To solve this problem, we propose a metacognition approach with naïve evaluation results: the LLM is given the task of analyzing previous evaluation results to recognize the hidden human judgment criteria used in creating the original dataset. With the LLM's analysis incorporated in the input prompt, the F1-score was significantly boosted from 0.662 to 0.773. This demonstrates that LLMs have the potential to serve as a metacognitive agent that can analyze and compensate for inherent human bias in the original dataset, thereby addressing data quality issues.

Keywords: naïve evaluation (preliminary evaluation), classification, reasoning, human judgment

1. Introduction

Large language models (LLMs) have become indispensable in most workplaces, offering immediate and helpful responses from various kinds of queries. There are several strategies to enhance the quality of LLM outputs, including prompt engineering [1], data augmentation [2], retrieval-augmented generation (RAG) [3], and reasoning mode. Reasoning mode enables LLMs to think by having the models generate reasoning contents before producing the final outputs, thereby automatically following the chain-of-thought process for enhanced understanding of the task at hand [4].

We have incorporated the reasoning mode into the Korean Language Understanding Evaluation (KLUE) topic classification (TC) benchmark dataset [5] to achieve a higher score than our baseline F1-score of 0.848 (Qwen3 14B, non-reasoning mode). To prepare the dataset for the reasoning mode, we instructed an auxiliary LLM to generate the reasoning content for the text-label pairs of the original dataset. Contrary

This research was conducted with the support of the Korea Creative Content Agency (Project Name: Artificial Intelligence-Based Interactive Multimodal Interactive Storytelling 3D Scene Authoring Technology Development, Project Number: RS-2023-00227917, Contribution Rate: 100%) and the Korea Research Foundation's four, Brain Korea 21 (Project Name: Ultra-Distributed Autonomous Computing Service Technology Research Team, Project Number: 202003520005).

to our expectation, the performance decreased significantly, with the result of 0.662. This highlights that the basic reasoning mode was counterproductive to the objective.

To solve this problem, we introduce the metacognition approach, which incorporates naïve evaluation results. This involves prompting an LLM to conduct a deep analysis of the naïve (preliminary, previous) evaluation results to recognize the hidden human judgment criteria that influenced the creation of the original labels. The resulting meta-knowledge is then constructed into a corrective prompt, which is given to the evaluation LLM for a second evaluation to achieve increased benchmark results. The remainder of the paper is organized as follows: Section 2 presents related works, Section 3 outlines the methodology, Section 4 describes the experiments, Section 5 discusses the findings, and Section 6 draws conclusions.

2. Related Works

2.1 Data Augmentation

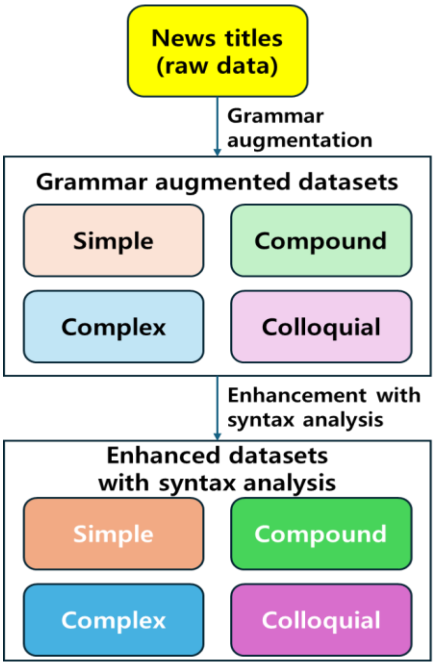


Fig. 1. Grammar augmentation of KLUE TC [5]

To enrich the outputs and increase the predictive performance of the LLMs, the input data can be augmented. Our previous research involved grammar-based augmentation [6] on the KLUE TC dataset [5]. Instructing an auxiliary LLM to expand the dataset with grammar and undergo syntax analysis, the datasets of the sentence types simple, compound, complex, and colloquial were generated, as shown in **Fig. 1**. The corresponding F1-scores were 0.819, 0.724, 0.409, and 0.695, indicating that clear and concise sentences are better understood than long and complex-structured sentences.

This paper is a direct extension of the above work, where reasoning content is added and the metacognition is applied.

2.2 Reasoning mode for classification

A similar task of reasoning generation was done by Henrichsen and Krebs [7]. These researchers challenged the notion that classification mode is often performed without any reasoning from the LLMs, and therefore, created a pipeline for generating the reasoning content of a training dataset and then outputting the final classified emotion. Their experiments showed a significant improvement of 0.087 points in accuracy compared to when the task was done without any reasoning. In the discussion, they mentioned that the quality of the reasoning generation may be questionable and that humans should evaluate the result. Additionally, they briefly mentioned the interpretability of the model, which our paper extends to the misalignment between the model's reasoning and human judgment criteria.

3. Methodology

This section consists of two parts: reasoning generation and metacognition on the naïve evaluation.

3.1 Reasoning Generation Mechanism

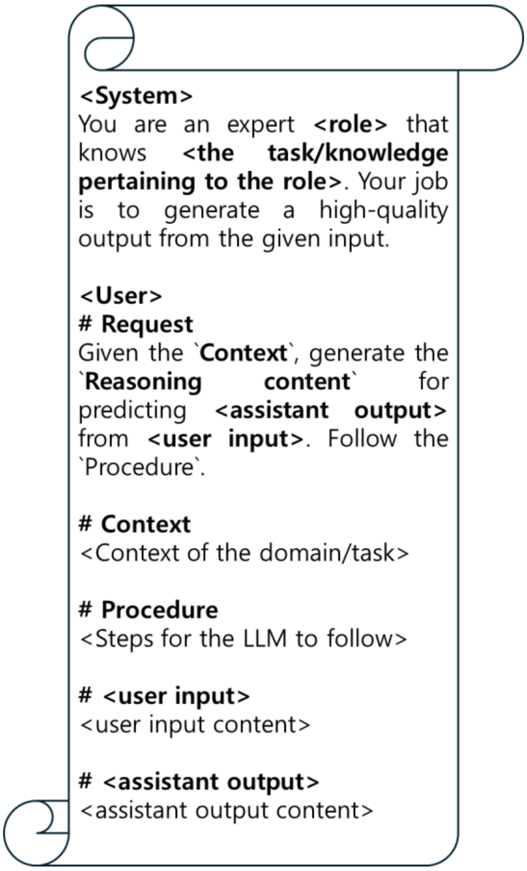
The reasoning content generation is done by instructing an auxiliary LLM to follow the procedure and generate the appropriate reasoning content for the prediction from text to label. The steps are as follows.

1. **Load the dataset.** The dataset, which consists of user input and assistant output, is loaded.
2. **Construct the template prompt.** The template prompt, which contains the

procedure for generating the reasoning content, is created.

3. **Append the user input and assistant output.** The template prompt is populated with the user's input data point and the assistant's output.
4. **Generate the reasoning content.** The LLM generates reasoning contents based on the filled prompt for the current datapoint.
5. **Repeat Steps 3 and 4.** These steps are repeated until reasoning content is generated for all datapoints in the dataset.

The abstract template prompt for Step 2 is provided in Fig. 2. This template should be completed with relevant domain and task information.



<System>
You are an expert **<role>** that knows **<the task/knowledge pertaining to the role>**. Your job is to generate a high-quality output from the given input.

<User>
# Request
Given the `**Context**`, generate the `**Reasoning content**` for predicting **<assistant output>** from **<user input>**. Follow the `Procedure`.

# Context
<Context of the domain/task>

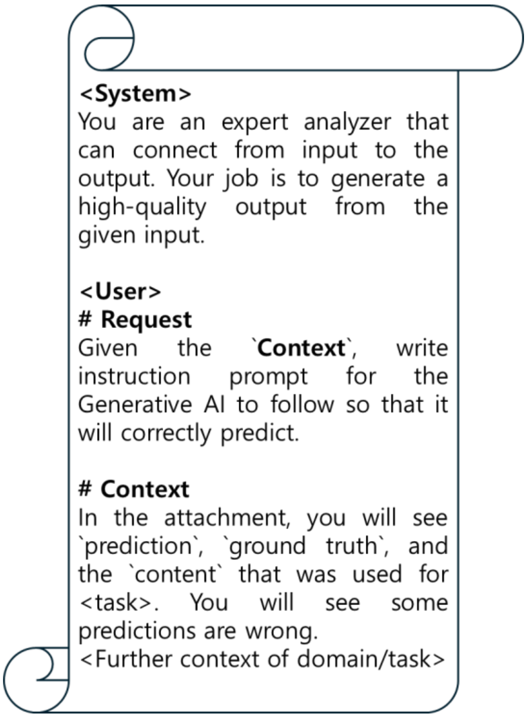
# Procedure
<Steps for the LLM to follow>

# <user input>
<user input content>

# <assistant output>
<assistant output content>

Fig. 2. Abstract template prompt for reasoning content generation

3.2 Mechanism for Metacognition on Naïve Evaluation



<System>
You are an expert analyzer that can connect from input to the output. Your job is to generate a high-quality output from the given input.

<User>
# Request
Given the `**Context**`, write instruction prompt for the Generative AI to follow so that it will correctly predict.

# Context
In the attachment, you will see `prediction`, `ground truth`, and the `content` that was used for <task>. You will see some predictions are wrong.
<Further context of domain/task>

Fig. 3. Abstract template prompt for evaluation procedure generation

To recognize the hidden human judgment criteria in creating the assistant output from the user input, an auxiliary LLM is given the results of naïve evaluation on the evaluation LLM. The auxiliary is instructed to generate the procedure for the evaluator to follow, ensuring that the latter reasons in the appropriate direction. The steps are as follows.

1. **Load the naïve evaluation results.** This includes the evaluation of LLM's prediction results with corresponding ground truths.
2. **Construct the template prompt.** The template prompt, which contains instructions for generating the procedure for the evaluator, is created.
3. **Append the naïve evaluation results.** The previous evaluation results are appended to the template prompt.
4. **Generate the evaluator procedure prompt.** From the completed prompt, the auxiliary LLM generates the procedure prompt for the evaluator to follow.

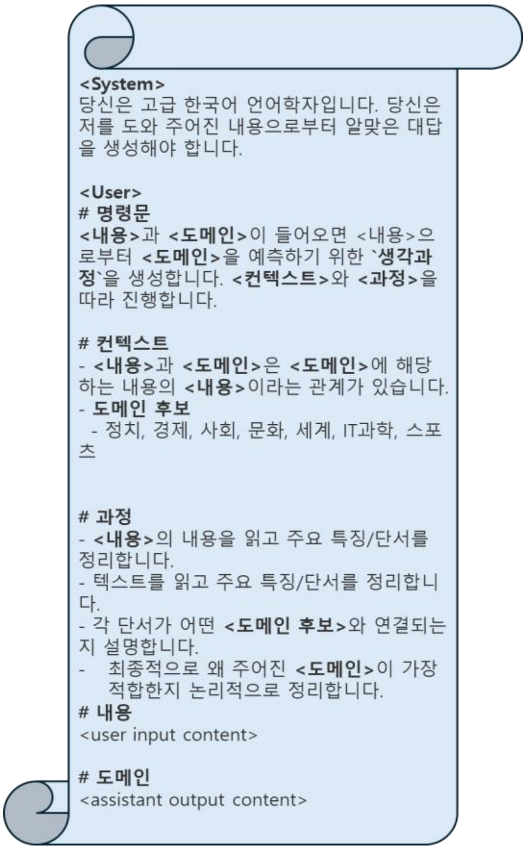
5. **Evaluate.** The corrective evaluation is conducted according to the procedure prompt.
6. **Repeat steps 3-5**. These steps are to be repeated until the evaluation reaches the desired level.

The abstract template prompt for Step 2 is provided in Fig. 3. Similarly, this template should be completed with relevant domain and task information.

4. Experiments

This section comprises experiments on reasoning content generation and metacognition in naïve evaluation.

4.1 Reasoning Generation Experiment



```
<System>
당신은 고급 한국어 언어학자입니다. 당신은
저를 도와 주어진 내용으로부터 알맞은 대답
을 생성해야 합니다.

<User>
# 명령문
<내용>과 <도메인>이 들어오면 <내용>으
로부터 <도메인>을 예측하기 위한 `생각과
정`을 생성합니다. <컨텍스트>와 <과정>을
따라 진행합니다.

# 컨텍스트
- <내용>과 <도메인>은 <도메인>에 해당
하는 내용의 <내용>이라는 관계가 있습니다.
- 도메인 후보
  - 정치, 경제, 사회, 문화, 세계, IT과학, 스포
츠


# 과정
- <내용>의 내용을 읽고 주요 특징/단서를
정리합니다.
- 텍스트를 읽고 주요 특징/단서를 정리합니
다.
- 각 단서가 어떤 <도메인 후보>와 연결되는
지 설명합니다.
-   최종적으로 왜 주어진 <도메인>이 가장
    적합한지 논리적으로 정리합니다.
# 내용
<user input content>

# 도메인
<assistant output content>
```

Fig. 4. Filled template prompt for reasoning content generation on KLUE TC [5]

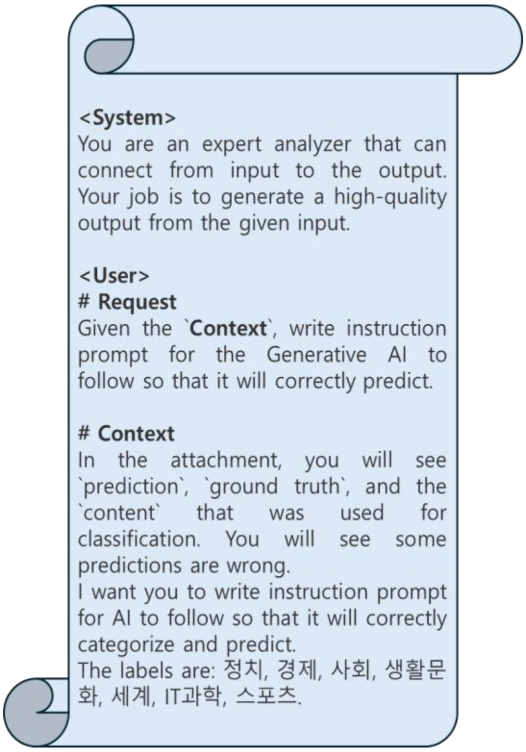
The dataset for generating reasoning content is the KLUE TC dataset [5]. We select the Qwen 3 32B AWQ model for its medium size, which allows for local parallel execution with sufficient memory, and its multilingual support for the Korean language. The machine specification includes an Xeon Gold 6416H (18-core, 2.2 GHz) CPU, 2 x 256 GB RAM, and an NVIDIA H200 14 GB GPU.

The abstract template prompt from Step 2 in Subsection 3.1 (Fig. 2) is populated with domain information and translated into Korean, as shown in Fig. 4. With this prompt, the reasoning content for the original KLUE TC dataset of 45,678 news title datapoints is created.

4.2 Metacognition on KLUE Evaluation

The metacognition on the hidden human judgment criteria in the labeling task of the KLUE TC dataset is repeated twice on ChatGPT5, with the evaluation LLM being the GPT-4.1 nano model fine-tuned with the KLUE TC dataset, and reasoning generated from Subsection 4.1 [8, 9].

The abstract template prompt from Step 2 in Subsection 3.2 (Fig. 3) is filled with domain information, as shown in Fig. 5.

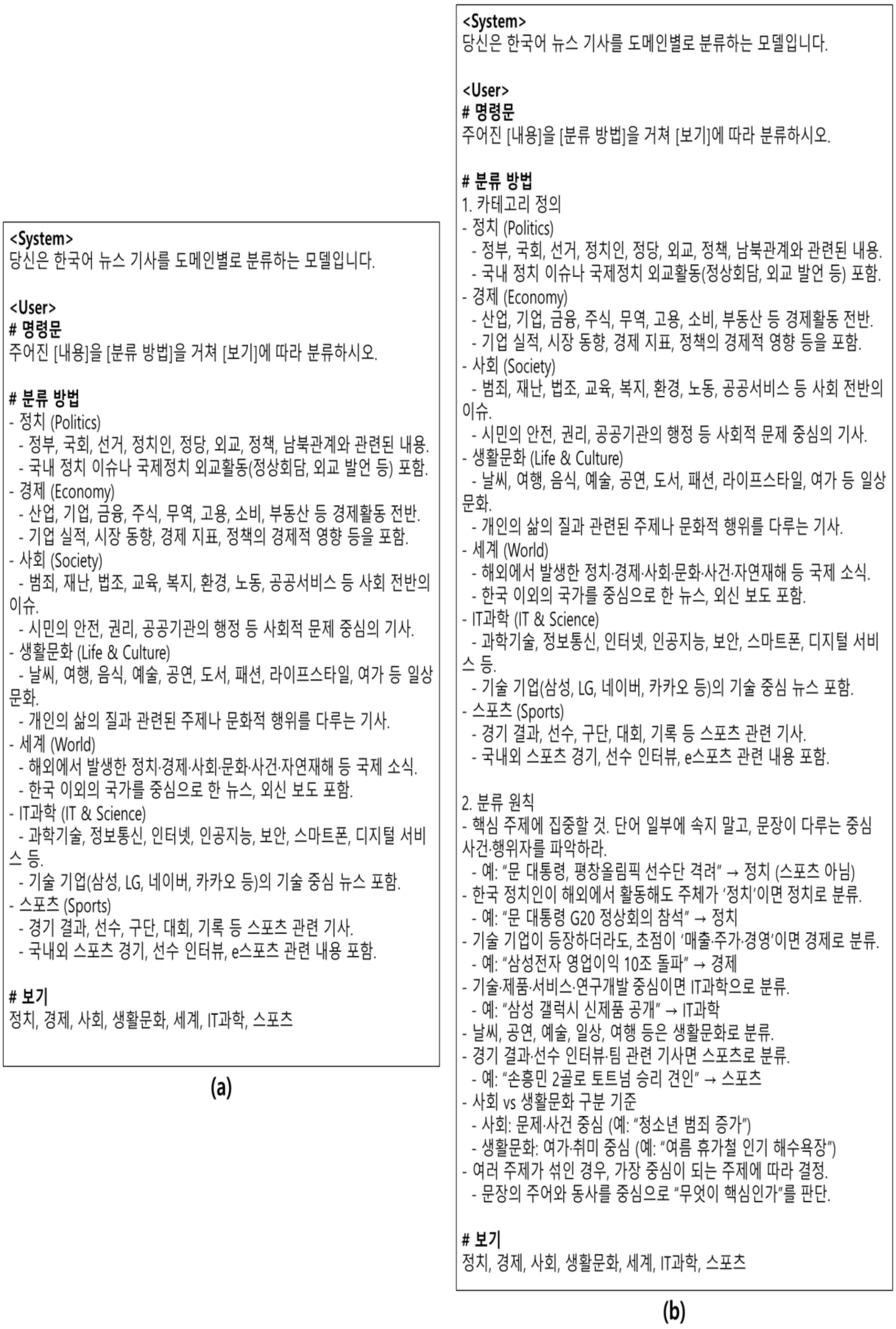


```
<System>
You are an expert analyzer that can
connect from input to the output.
Your job is to generate a high-quality
output from the given input.

<User>
# Request
Given the `Context`, write instruction
prompt for the Generative AI to
follow so that it will correctly predict.

# Context
In the attachment, you will see
`prediction`, `ground truth`, and the
`content` that was used for
classification. You will see some
predictions are wrong.
I want you to write instruction prompt
for AI to follow so that it will correctly
categorize and predict.
The labels are: 정치, 경제, 사회, 생활문
화, 세계, IT과학, 스포츠.
```

Fig. 5. Filled prompt for metacognition analysis on naïve evaluation on KLUE TC [5]



```
<System>
당신은 한국어 뉴스 기사를 도메인별로 분류하는 모델입니다.

<User>
# 명령문
주어진 [내용]을 [분류 방법]을 거쳐 [보기]에 따라 분류하시오.

# 분류 방법
- 정치 (Politics)
  - 정부, 국회, 선거, 정치인, 정당, 외교, 정책, 남북관계와 관련된 내용.
  - 국내 정치 이슈나 국제정치 외교활동(정상회담, 외교 발언 등) 포함.
- 경제 (Economy)
  - 산업, 기업, 금융, 주식, 무역, 고용, 소비, 부동산 등 경제활동 전반.
  - 기업 실적, 시장 동향, 경제 지표, 정책의 경제적 영향 등을 포함.
- 사회 (Society)
  - 범죄, 재난, 법조, 교육, 복지, 환경, 노동, 공공서비스 등 사회 전반의 이슈.
  - 시민의 안전, 권리, 공공기관의 행정 등 사회적 문제 중심의 기사.
- 생활문화 (Life & Culture)
  - 날씨, 여행, 음식, 예술, 공연, 도서, 패션, 라이프스타일, 여가 등 일상 문화.
  - 개인의 삶의 질과 관련된 주제나 문화적 행위를 다루는 기사.
- 세계 (World)
  - 해외에서 발생한 정치·경제·사회·문화·사건·자연재해 등 국제 소식.
  - 한국 이외의 국가를 중심으로 한 뉴스, 외신 보도 포함.
- IT과학 (IT & Science)
  - 과학기술, 정보통신, 인터넷, 인공지능, 보안, 스마트폰, 디지털 서비스 등.
  - 기술 기업(삼성, LG, 네이버, 카카오 등)의 기술 중심 뉴스 포함.
- 스포츠 (Sports)
  - 경기 결과, 선수, 구단, 대회, 기록 등 스포츠 관련 기사.
  - 국내외 스포츠 경기, 선수 인터뷰, e스포츠 관련 내용 포함.

# 보기
정치, 경제, 사회, 생활문화, 세계, IT과학, 스포츠
```

(a)

```
<System>
당신은 한국어 뉴스 기사를 도메인별로 분류하는 모델입니다.

<User>
# 명령문
주어진 [내용]을 [분류 방법]을 거쳐 [보기]에 따라 분류하시오.

# 분류 방법
1. 카테고리 정의
- 정치 (Politics)
  - 정부, 국회, 선거, 정치인, 정당, 외교, 정책, 남북관계와 관련된 내용.
  - 국내 정치 이슈나 국제정치 외교활동(정상회담, 외교 발언 등) 포함.
- 경제 (Economy)
  - 산업, 기업, 금융, 주식, 무역, 고용, 소비, 부동산 등 경제활동 전반.
  - 기업 실적, 시장 동향, 경제 지표, 정책의 경제적 영향 등을 포함.
- 사회 (Society)
  - 범죄, 재난, 법조, 교육, 복지, 환경, 노동, 공공서비스 등 사회 전반의 이슈.
  - 시민의 안전, 권리, 공공기관의 행정 등 사회적 문제 중심의 기사.
- 생활문화 (Life & Culture)
  - 날씨, 여행, 음식, 예술, 공연, 도서, 패션, 라이프스타일, 여가 등 일상 문화.
  - 개인의 삶의 질과 관련된 주제나 문화적 행위를 다루는 기사.
- 세계 (World)
  - 해외에서 발생한 정치·경제·사회·문화·사건·자연재해 등 국제 소식.
  - 한국 이외의 국가를 중심으로 한 뉴스, 외신 보도 포함.
- IT과학 (IT & Science)
  - 과학기술, 정보통신, 인터넷, 인공지능, 보안, 스마트폰, 디지털 서비스 등.
  - 기술 기업(삼성, LG, 네이버, 카카오 등)의 기술 중심 뉴스 포함.
- 스포츠 (Sports)
  - 경기 결과, 선수, 구단, 대회, 기록 등 스포츠 관련 기사.
  - 국내외 스포츠 경기, 선수 인터뷰, e스포츠 관련 내용 포함.

2. 분류 원칙
- 핵심 주제에 집중할 것. 단어 일부에 속지 말고, 문장이 다루는 중심 사건·행위자를 파악하라.
  - 예: "문 대통령, 평창올림픽 선수단 격려" → 정치 (스포츠 아님)
- 한국 정치인이 해외에서 활동해도 주체가 '정치'이면 정치로 분류.
  - 예: "문 대통령 G20 정상회의 참석" → 정치
- 기술 기업이 등장하더라도, 초점이 '매출·주가·경영'이면 경제로 분류.
  - 예: "삼성전자 영업이익 10조 돌파" → 경제
- 기술·제품·서비스·연구개발 중심이면 IT과학으로 분류.
  - 예: "삼성 갤럭시 신제품 공개" → IT과학
- 날씨, 공연, 예술, 일상, 여행 등은 생활문화로 분류.
- 경기 결과·선수 인터뷰·팀 관련 기사면 스포츠로 분류.
  - 예: "손흥민 2골로 토트넘 승리 견인" → 스포츠
- 사회 vs 생활문화 구분 기준
  - 사회: 문제·사건 중심 (예: "청소년 범죄 증가")
  - 생활문화: 여가·취미 중심 (예: "여름 휴가철 인기 해수욕장")
- 여러 주제가 섞인 경우, 가장 중심이 되는 주제에 따라 결정.
  - 문장의 주어와 동사를 중심으로 "무엇이 핵심인가"를 판단.

# 보기
정치, 경제, 사회, 생활문화, 세계, IT과학, 스포츠
```

(b)

Fig. 6. Procedures generated from metacognition analysis from **(a)** first evaluation results and **(b)** second results

The procedures generated by ChatGPT5 are shown in **Fig. 6**. With the second procedure, there was a significant increase in the F1-score, from 0.662 on the naïve evaluation with reasoning to 0.773 on the corrective prompt.

5. Discussions

The original KLUE TC dataset does not include the decision factors used in labeling news titles when labels overlapped, other than the fact that the majority vote from three people determined the final label [5], which is why metacognition was necessary. This mechanism was effective in enhancing evaluation performance from naïve reasoning because the auxiliary LLM recognized the hidden human judgment criteria—specific to these three people—that the evaluation LLM was supplied with.

There is a notable limitation on this mechanism. If the human judgment is biased in a single direction—the same hidden rule was applied in creating the text-answer pair—the auxiliary LLM is more likely to produce a single corresponding rule for the hidden human judgment criterion. However, if the bias is random—the choice among the overlapping labels is random—the auxiliary may not be able to differentiate the hidden rule for each datapoint. For example, “게시판 과천과학관 2016년 최우수 책임운영기관 선정” and “관장님 언제 오나요... 과천과학관 1년째 수장 공백” (translated as “Notice: Gwacheon National Science Museum Selected as the Best Responsible Administrative Agency of 2016” and “When Will the Director Arrive? Gwacheon National Science Museum Without a Chief for a Year”) both hint at the societal issue about Gwacheon National Science Museum, but the former was labeled as “society” and the latter “science.” This increases the difficulty for the auxiliary in recognizing the random nature of the human mind's rules.

The metacognition mechanism was applied to the text-label set of the classification task. Still, it can be extended to any dataset, e.g., the judgment behind the reasoning steps of a reasoning dataset. Suppose the mechanism is to be applied to more complex benchmark datasets. In that case, the template prompts will require more thoughtful and careful designs, as the basic template prompt for the classification task in our case was relatively simple.

6. Conclusions

To leverage the current capacity of generative AI, reasoning mode and prompts are often utilized to enhance the quality of the AI's output. We have demonstrated the mechanism for generating the reasoning content of a simple text-label pair benchmark dataset for KLUE TC and the metacognition involved in naïve evaluation, which informs the design of the reasoning steps. With the metacognition approach, the performance increased from 0.662 to 0.773 in F1-score. For future work, we will increase the repetition count of the metacognition analysis to enhance performance further and extend the application to static code analysis.

References

[1] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, “Unleashing the potential of prompt engineering for large language models,” *Patterns*, vol.6, no.6, pp.1-44, 2025.

[2] J. Kim, Y. Lee, Y. Han, S. Jung, and H. Choi, “Does Incomplete Syntax Influence Korean Language Model? Focusing on Word Order and Case Markers,” *arXiv preprint arXiv:2407.09184*, pp.1-19, 2024.

[3] M. Arslan, H. Ghanem, S. Munawar, and C. Cruz, “A survey on RAG with LLMs,” Procedia Computer Science, vol.246, pp.3781-3790, 2024.

[4] J. Wang, “A tutorial on LLM reasoning: Relevant methods behind ChatGPT o1,” arXiv preprint arXiv:2502.10867, pp.1-15, 2025.

[5] S. Park, et al., “KLUE: Korean Language Understanding Evaluation,” *arXiv preprint arXiv:2105.09680*, pp.1-76, 2021.

[6] J. Yang and R. Kim, “Enhancing performance of natural language understanding of AI with syntax augmentation,” accepted for publication, 2025.

[7] M. Henrichsen and R. Krebs, “Two-stage reasoning-infused learning: Improving classification with LLM-generated reasoning,” *arXiv preprint arxiv:2507.00214*, pp.1-17, 2025.

[8] *ChatGPT*, OpenAI. [Online]. Available: https://chatgpt.com, accessed Oct. 23, 2025.

[9] GPT-4.1-nano-2025-04-14, OpenAI. [Online]. Available: https://platform.openai.com/docs/models/gpt-4.1-nano, accessed Oct. 23. 2025.