Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

Enhancing Alignment Between Large Language Models and Teacher in Open-Ended Assessment through In-Context Learning

Juyeong Lee[1*], Euna Lee[2], Kangrea Kim[3]

[1]Seoul National University

[2]Osan Wonil Middle School

[3]Sundong Elementary School

In-Context Learning을 통한 서·논술형 평가에서의 대형 언어 모델과 교사 간 채점 및 피드백 정합성 향상 방안

이주영[1*], 이은아[2], 김강래[3]

[1]서울대학교, [2]오산원일중학교, [3]선동초등학교

*Corresponding Author: Juyeong Lee, jy9307@snu.ac.kr

ABSTRACT

This study investigates the effectiveness of in-context learning (ICL) in enhancing the agreement between human teachers and large language models (LLMs) in the context of open-ended assessments. Using a dataset of 485 student responses to six open-ended questions from Korean, Technology, and Social Studies subjects administered in 2024, teacher-generated scores and feedback were collected alongside LLM-generated outputs under varying ICL conditions. Specifically, we provided GPT-4.1 with 0 to 20 examples in prompts to examine whether increasing example count improves agreement between the model and human raters. Quadratic Weighted Kappa (QWK) was used to assess score alignment, and BERTScore measured semantic similarity between teacher and model feedback. Regression and mixed-effects analyses revealed that increasing the number of examples generally improved alignment up to a certain threshold. The strongest improvements occurred with fewer than six examples, beyond which the benefits plateaued or even declined. Additionally, prompt length negatively moderated the effect of example count, suggesting that longer prompts may reduce the model's capacity to focus on relevant information. These results provide practical guidance for teachers using LLMs in open-ended assessments. Including teacher-generated examples in prompts helps models align more closely with human scoring and feedback. However, the optimal number of examples depends on the type of question and expected answer length: more examples benefit shorter responses, while fewer examples (five or fewer) are more effective for longer or more complex answers.

Keywords: In-context learning, few-shot learning, few-shot prompting, large language models, automated scoring, feedback alignment, ai-assisted assessment, reliability in educational assessment

Brain, Digital, & Learning
2025, Vol. 15, No. 3, 375-401
https://doi.org/10.31216/BDL.2025.15.3.4

Received: July 23, 2025
Revised: September 08, 2025
Accepted: September 10, 2025

Introduction

최근 교육 현장에서는 학습자의 고차 사고력과 자기표현 능력을 종합적으로 평가하기 위한 방안으로 서·논술형 평가의 중요성이 점차 강조되고 있다(Brookhart, 2010). Brookhart (2010)에 따르면, 서·논술형 평가는 단순한 선택형 문항과는 달리 서술형 문항에서의 사실적 지식 서술 및 개념 설명 능력부터 논술형 문항에서의 사고의 깊이, 논리적 구조, 표현력에 이르기까지 다층적인 인지 능력을 측정할 수 있다는 점에서 교육적 가치가 크다. 더 나아가 서·논술형 평가는 학생들이 복잡한 문제를 해결하고, 비판적으로 판단하며, 창의적으로 사고할 수 있는 능력을 기를 수 있도록 하는 효과적인 평가 도구로서 기능한다.

그러나 이러한 평가가 실제 학습에 긍정적인 영향을 미치기 위해서는 단순한 점수 부여를 넘어, 학생의 사고를 촉진하고 성찰을 유도할 수 있는 적절한 피드백과 일관된 평가 기준이 함께 제공되어야 한다(Hattie & Timperley, 2007; Nicol & Macfarlane‐Dick, 2006; Shute, 2008). 그러나 실제 교육 현장에서는 교육적 도구로 서·논술형 평가를 적극 활용하는 데 있어 여러 한계가 존재한다. Narciss (2013)는 시간 부족, 평가 역량의 차이, 인력의 한계 등으로 인해 서·논술형 평가에 대한 개별화된 피드백 제공이 어렵다는 사실을 확인한 바 있으며, Huang & Whipple (2023)은 서·논술형 평가에 대해 평가자마다 기준이 달라 채점의 신뢰도와 일관성을 유지하는 데에도 어려움이 있다고 지적하였다.

이러한 맥락에서, 최근 인공지능 기반의 자동 채점 및 피드백 시스템은 이러한 한계를 극복하고 서·논술형 평가가 지닌 학습의 질을 향상시킬 수 있는 도구로 주목받고 있다(Chiu et al., 2023; Huang et al., 2020). Dai et al. (2023)는 글쓰기 과제에서 ChatGPT가 유의미한 피드백을 제공할 수 있음을 확인한 바 있으며, Shin et al. (2024)과 Pack et al. (2024)은 LLM 기반 자동 채점 시스템이 실제 교사 평가와 유사한 수준의 평가 결과를 제공함을 실증하기도 했다.

그러나 Henderson et al. (2025)에 따르면, 학생들은 여전히 AI의 피드백보다 교사의 피드백을 더 도움이 되는 것으로 인식하고 있었으며, Baz & Hasırcı (2025)는 장기적인 관점에서 교사의 피드백이 학습자에게 더 효과적인 영향을 미친다는 사실을 확인하기도 했다. 또한 교사가 평가의 맥락을 가장 잘 이해하는 평가 설계자로서 다양한 계층에서 평가의 결과를 해석할 수 있는것에 비해, LLM은 일반적인 목적과 광범위한 사용자층을 전제로 설계되어 있어 특정 교육 맥락에 맞는 정교한 평가 판단이나 교사의 관점을 정밀하게 반영하는 데 한계를 보일 수밖에 없다(Lundgren, 2025).

이러한 문제를 보완하기 위한 대안으로 In-Context Learning (ICL)이 존재한다. 이는 별도의 파인튜닝(fine-tuning) 없이, 모델에게 예시 데이터를 프롬프트 내에서 제공함으로써 보다 특정한 양식과 평가 기준에 맞는 출력을 유도하는 방식이다(Brown et al., 2020). 특히 교육 현장에서는 별도의 모델 수정이나 학습 비용 없이 다양한 교사의 평가 및 피드백 관점을 반영할 수 있다는 점에서, ICL은 실용적인 접근법이라 할 수 있다(Wei et al., 2022).

이에 본 연구는 ICL을 통해 서·논술형 평가에서 LLM이 교사의 채점 및 피드백 작성 관점을 학습하고 모방할 수 있는지 탐색하고자 한다. 즉, 교사의 평가 예시를 프롬프트에 더 많이 포함시킴에 따라 교사와 LLM 간에 채점 결과 및 피드백의 정합성이 향상됨을 실험적으로 검증하고자 한다.

또한 서·논술형 평가에서는 각 문항의 특성에 따라 요구하는 답변의 길이 및 수준 등이 크게 달라질 수 있다. 예를 들어, 특정 문장의 의미를 해석할 것을 요구하는 문항은 비교적 짧은 길이로 답변이 가능하나, 두 작품의 주제 의식을 논리적으로 비교·평가할 것을 요구하는 논술형 문항의 경우 상대적으로 매우 긴 답변을 필

요로 한다. Levy et al. (2024)에 따르면 LLM은 동일한 과업을 수행함에 있어 제공받는 프롬프트의 길이(프롬프트가 포함하는 토큰의 수)가 길어질수록 핵심 정보를 효과적으로 추출하지 못하거나, 주의가 분산되어 추론 성능이 저하되는 경향을 보인다. 이는 LLM이 ICL을 통해 채점과 피드백 제공을 수행한다고 했을 때, 동일한 수의 예시를 제공받더라도 문항별 특성에 따른 예시의 길이 차이로 인해 성능 차이가 발생할 수 있음을 시사한다. 따라서 본 연구에서는 각 예시의 토큰 수를 조절 변수로 고려하여, 예시 수가 정합성 향상에 미치는 영향을 분석하고자 한다.

이상의 분석을 위해 다음 두 가지 연구문제를 설정하였다. 첫째, ICL을 활용하였을 때 프롬프트에 포함되는 예시의 수는 교사와 LLM 간 채점 결과 및 피드백의 정합성에 유의미한 영향을 미치는가? 둘째, ICL을 활용하였을 때 프롬프트가 포함하는 토큰의 수는 예시의 수가 교사와 LLM의 정합성에 미치는 영향을 조절하는가?

Materials and Methods

Materials

본 연구에서는 2024년 A중학교에서 실시된 국어, 기술·가정, 사회 교과의 서·논술형 평가 문항 6개(각 교과별 2문항)와 이에 대한 중학교 1학년 학생의 답안을 활용하였다(Appendix A). 본 연구를 수행한 연구자들은 교직 경력 6년, 16년, 5년의 교사로서 각 학생 응답에 대해 사전에 정립된 평가 기준에 따라 채점을 실시하고 개별 피드백을 작성하였다.

최초에 수집한 문항들은 2~3개의 평가요소를 기반으로 하는 분석적 채점기준을 가지고 있었으나, 다면적 평가요소를 모두 적용할 경우 교사의 채점과 LLM의 결과 간 불일치 발생 시 그 원인이 특정 평가요소에 기인한 것인지 혹은 요소 간 상호작용에서 비롯된 것인지 명확히 구분하기 어려운 한계가 있다. 이에 본 연구에서는 각 문항별로 가장 핵심적이고 대표성을 지닌 평가요소 하나를 선정하여 분석의 초점을 명료화하였다. 선정된 각 평가요소의 5개 수준 기준은 연구자 3인의 논의를 거쳐 명료화하였다(Appendix B). 분석의 일관성을 위해 각 평가에는 약어를 부여하여 표기하였는데, 국어 평가는 'KR 1'과 'KR 2', 기술·가정 평가는 'TH 1'과 'TH 2', 사회 평가는 'SS 1'과 'SS 2'로 명명하였다.

6개의 평가에 대해 각 평가마다 최소 70명에서 최대 101명의 학생 답안을 수집하여 총 485개의 학생 답안을 연구에 활용하였다. 연구자 3인은 사전에 설정한 평가 기준에 따라, 학생 답안을 직접 채점하고 피드백을 작성하였다. 이 때, 일반 교사들의 피드백 작성 경향을 반영하되, 연구자 간 편차를 최소화하기 위해 선행연구를 참고하여 통일된 작성 지침을 도출하였다. Parr & Timperley(2010)에 따르면, 서·논술형 문항에 대한 교사들의 피드백은 주로 오류를 교정하거나, 학생의 노력을 격려하고, 답안의 전반적인 강점과 개선점을 요약하는 경향을 보인다.

이러한 내용에 기반하여 본 연구에서는 세 명의 연구자가 통일된 기준에 따라 피드백을 작성하기 위해 다음과 같은 작성 지침을 따랐다. 모든 피드백은 완결된 문장으로 작성하되 느낌표, 이모티콘, 물결표 등은 사용하지 않았으며 평가 기준에 근거하여 드러나는 오류와 그에 대한 개선점을 제시하였다. 이 과정에서 평가 기준과 무관한 피드백은 일체 배제하였다.

작성된 채점 결과와 피드백은 일관성과 신뢰도를 위해 연구자 간 상호 비교 및 검토를 실시하였다. 채점

시에는 3인이 독립적으로 채점을 실시하였으며, 신뢰도 검증을 위해 Fleiss's Kappa를 산출하였다. 분석 결과(Table 2), 전체 평가에 대한 채점자 간 신뢰도는 평균 $\kappa = .82$로 높은 수준의 일치도를 보였다. 평가 문항별로는 SS 1 ($\kappa = .79$), SS 2 ($\kappa = .81$), TH 1 ($\kappa = .87$), TH 2 ($\kappa = .82$), KR 1 ($\kappa = .84$), KR 2 ($\kappa = .80$)로 모든 문항에서 .75 이상의 높은 신뢰도를 확보하였다. 채점 결과가 연구자별로 상이한 경우에는 충분한 논의를 거쳐 평가 기준에 가장 부합하는 수준을 합의하여 최종 채점 결과로 확정하였다. 작성된 피드백은 평가 기준과의 일관성을 중심으로 연구자 간 검토 과정을 거쳐 최종 확정하였다(Appendix C).

Procedure

In-Context Learning

본 연구는 ICL기법을 활용하였다. ICL은 사전 학습된 언어 모델의 파라미터를 별도로 조정하지 않고, 프롬프트 내에 소량의 예시를 포함시켜 모델이 해당 작업의 양식과 논리를 일시적으로 모방하도록 유도하는 방식이다(Brown et al., 2020).

서·논술형 평가에서 LLM 기반 채점 및 피드백 모델이 실제 교육 현장에서 효과적으로 활용되기 위해서는, LLM이 교사 개개인의 평가 관점과 피드백 방식을 모방할 수 있어야 한다. 이에 본 연구에서는 교사와 LLM이 산출한 채점 결과의 유사도 및 피드백의 평가 기준 일치 정도를 정합성(Agreement)으로 정의하고 이를 향상시킬 수 있는 방법을 탐구하였다.

Mazzullo & Bulut (2025)는 파인튜닝을 통해 교육적 맥락에서 LLM이 유의미한 피드백을 제공할 수 있음을 보인 바 있으나, 각 평가 문항의 특성과 교사의 관점에 맞추어 매번 모델을 파인튜닝하는 것은 시간, 비용, 기술적 자원의 측면에서 매우 비효율적이다. 이러한 관점에서 ICL은 교사의 평가 사례를 프롬프트에 포함하는 것만으로 모델이 일시적으로 교사의 평가 관점을 따르도록 유도할 수 있기에 파인튜닝의 효율적인 대안이 될 수 있으며, Wan & Chen (2024)는 소수의 피드백 예시를 활용한 ICL를 통해 GPT-3.5가 교육적으로 유용한 피드백을 제공할 수 있음을 실증하기도 했다. Brown et al. (2020)에 따르면 모델의 파라미터 수가 늘어날수록 ICL의 효과는 뚜렷하게 높아지기에, 본 연구에서는 OpenAI에서 제공하는 모델 중 가장 파라미터 수가 많은 것으로 예측되는 GPT-4.1을 활용하여 서·논술형 평가에서 예시의 제공이 LLM과 교사 간 채점 결과 및 피드백 정합성 확보에 미치는 영향을 실험적으로 분석하였다.

Score & Feedback Generation

본 연구에서는 6개의 평가 문항별로 학생 답안-채점-피드백 데이터를 추출하고자 하였는데, 이때 6개 문항 중 답안 수가 가장 적은 SS 1 문항이 70개의 답안을 보유하고 있었으므로 성능 평가의 신뢰성을 확보하기 위해 예시 데이터는 최대 20개로 제한하고, 예시 데이터로 사용되지 않은 나머지 답안들은 교사와 AI의 정합성 검증을 위한 분석 데이터로 활용하였다.

또한 예시 선정 방식의 타당성을 검증하기 위해 파일럿 연구를 사전에 수행하였다. 파일럿 연구에서는 무작위 예시 선정 방식과 데이터셋 내의 점수대별 비율을 고려한 층화(Stratified) 추출 방식을 비교하였다. 모든 예시 제공이 더 효과적일 것으로 예상하였으나, 실제 결과에서는 무작위 선정 방식이 교사-LLM 간 정합성 측면에서 더 안정적이고 우수한 성과를 보였다. 이러한 파일럿 결과를 바탕으로 본 연구에서는 무작위 예시 선정 방식을 최종 채택하였다.

프롬프트는 Yoshida (2024)의 연구를 참고하여 구성하였다. 채점을 위한 프롬프트는 문항 설명, 평가 기준,

교사 채점 예시, 그리고 학생의 실제 응답을 포함하였다(Fig. 1a). 모델은 이러한 프롬프트를 바탕으로, 1점에서 5점 사이의 정수를 생성하도록 지시받았다. 이때 교사의 채점 예시 데이터를 0-20개 범위에서 한 개씩 늘려가며 LLM의 채점 결과가 교사의 피드백과 얼마나 유사해지는지를 검증하였다. 피드백을 위한 프롬프트는 문항 설명, 평가 기준, 교사의 채점 결과 및 피드백 예시, 그리고 학생 응답을 포함하였다(Fig. 1b). 모델은 이를 바탕으로 학생의 답안에 적절한 피드백을 생성하도록 지시받았다. 마찬가지로 교사의 피드백 예시 데이터를 0~20개 범위에서 한 개씩 늘려가며 LLM의 피드백이 교사의 피드백과 얼마나 유사해지는지를 검증하였다.

```
{ "role": "system", "content":
  """
  You are an assistant that evaluates students' essay-style responses.

  ## The following is the question:
  {question}

  ## Scoring criteria :
  {scoring criteria}

  ## The following is the examples :
  {examples}

  Please assign a score based on the criteria above. The score should be
    an integer between 1 and 5.

  ## Respond with only the score number:
  """
},
{ "role": "user", "content": "student response: {student_answer}" }
```

(a)

```
{ "role": "system", "content":
  """
  You are an assistant that provides feedback on students' essay-style
    responses.

  ## The following is the question:
  {question}

  ## Scoring criteria :
  {scoring_criteria}

  ## The following is the examples :
  {examples}

  Please provide feedback based on the evaluation perspective, tone, and
    length of the example feedback above.

  ## Student response has been scored {human_score}/5.
  ## Provide feedback in Korean:
  """
},
{ "role": "user", "content": "student response: {student_answer}" }
```

(b)

Fig. 1. Templates of prompts for LLM : (A) A scoring prompt, (B) A feedback prompt

LLM의 일관된 답변 생성을 유도하기 위해 토큰 생성 확률과 관계된 하이퍼파라미터를 고정하여 활용하였다. temperature는 0.001로 매우 낮게 설정하여 출력의 일관성을 확보하였으며, top-p은 0.9로 설정하여 모델이 상위 누적 확률 90% 내에서만 토큰을 선택하게 함으로써 안정적인 다양성을 유지하도록 하였고 seed 값을 42로 고정함으로써 동일한 입력에 대해 동일한 출력을 재현할 수 있도록 하였다. 또한 LLM이 답변 생성에 활용한 프롬프트의 토큰 수를 함께 로깅하여 추후 분석에 활용할 수 있도록 하였다.

최종적으로 학생 답안 485개 중 ICL을 위한 답안 120개(평가별 20개)를 제외한 365개의 답안에 대해 예시 수를 0개에서 20개까지 증가시키며 LLM이 채점과 피드백 작성을 수행한 결과를 수집하여 7,665개의 데이터를 가진 데이터셋을 구축하였다. 각 데이터는 평가 문항, 평가 기준, 학생 답안, 교사 피드백, LLM 피드백, 교사 채점 결과, LLM 채점 결과, 피드백 프롬프트 토큰 수, 채점 프롬프트 토큰 수를 포함하였다.

Measurements

Agreement Between LLM-Generated and Human Scores

LLM이 생성한 채점 결과와 교사의 채점 결과 간 정합성을 정량적으로 비교하기 위해 본 연구에서는 Quadratic Weighted Kappa (QWK) 점수를 활용하였다. QWK는 서로 다른 평가자 간의 일치도를 측정하는 대표적인 지표로, 단순한 일치도와는 달리 채점 결과 간의 상대적 거리를 가중치로 반영할 수 있다(Cohen, 1968). 부분 일치와 완전 불일치를 구분하고 등간 척도 기반의 채점 체계를 정확히 반영할 수 있다는 점에서 QWK는 적절한 신뢰도 분석 도구로 간주되며, 실제로 교육 평가 및 자동 채점 시스템의 성능 검증에 있어 QWK는

표준적인 지표로 널리 사용되어 왔다(Doewes et al., 2023; Yoshida, 2024). 이에 본 연구는 채점 결과의 예시 수를 0개에서 20개까지 점차 증가시키며 LLM과 교사 채점 결과 간 QWK 점수의 변화 추이를 분석하였다.

Agreement Between LLM-Generated and Human Feedback

LLM이 생성한 피드백과 교사가 작성한 피드백 간의 의미적 유사성을 정량적으로 비교하기 위해 본 연구는 BERTScore를 활용하였다. BERTScore는 사전 학습된 BERT 언어모델의 임베딩 벡터를 기반으로 두 문장 간의 의미적 정렬 정도를 계산하며, 단어 수준의 정렬만을 평가하는 기존 BLEU나 ROUGE 등의 지표에 비해 보다 정교한 평가가 가능하다(Zhang et al., 2020). 특히 서·논술형 피드백은 표현 방식이 다양하더라도 핵심 평가 관점이나 전달하고자 하는 내용이 유사할 수 있는데, BERTScore는 이러한 의미상의 유사성을 정량적으로 포착할 수 있는 강점을 가진다. 다만, BERTScore는 의미적 유사성에만 초점을 두기 때문에 이를 근거로 LLM이 교사 피드백의 교육적 효용성이나 구체성 같은 질적 특성까지 충실히 재현한다고 보기는 어렵다. 그럼에도 불구하고 문장 간 의미적 유사도를 정량적으로 평가할 수 있다는 점에서, 본 연구의 목적에 해당하는 교사와 LLM 간 피드백 정합성 측정에는 충분히 유의미하게 활용될 수 있다고 판단하였다. 본 연구에서는 피드백 예시 수를 0개에서 20개까지 점차 증가시키며 LLM과 교사 채점 결과 간 BERTScore의 변화 추이를 분석하였다.

Data Analysis

본 연구에서는 Python의 statsmodels패키지를 사용하여 ICL을 활용하였을 때 예시 개수가 LLM과 교사의 채점 결과 및 피드백 정합성에 미치는 영향과 이에 대한 프롬프트의 토큰 수의 조절 효과를 분석하였다.

이를 위해 첫째, 선형회귀 분석(Linear regression analysis)을 실시하였다. 독립변수 X는 프롬프트 내에 포함된 예시의 개수이며, 종속변수 Y는 각각 QWK 점수와 BERTScore로 설정하였다. 각 평가 항목별로 개별적인 선형회귀분석을 우선 수행한 뒤, 이러한 분석 결과를 종합하여 평가 간 변이를 고려한 선형혼합효과모형 분석(Linear mixed-effects model analysis)을 실시하였다. 이를 통해 평가의 개별 특성의 영향을 통제한 상태에서 예시 수 증가에 따른 정합성 변화 경향을 확인하였다. 이어 예시 수 증가에 따른 효과가 전 구간에 걸쳐 단일한 선형 경향으로 수렴하지 않을 가능성을 고려하여 예시 수 범위를 0–5, 6–10, 11–15, 16–20의 네 개 구간으로 나누어 구간별로 선형혼합효과모형을 구성하였다. 이를 분석함으로써 예시 수의 구간에 따라 효과가 변화하거나 포화되는 경향이 있는지를 확인하고자 하였다.

둘째, 예시 수가 채점 결과 및 피드백 정합성에 미치는 영향을 프롬프트 길이가 조절하는지 확인하기 위해 프롬프트의 길이(프롬프트가 포함하는 토큰 수)를 Z로 설정하여 조절 분석(Moderation analysis)을 설계하였다. 조절 분석은 예시 수(X)와 프롬프트 길이(Z) 간의 상호작용 항(X×Z)을 포함하는 형태로, 전체 구간을 대상으로 한 선형혼합효과모형에 조절 항을 추가하여 구성하였다. 이 때, 예시 수와 프롬프트 길이 간의 다중공선성을 완화하기 위해 두 변수를 평균 중심화한 후 분석을 수행하였으며 이를 통해 프롬프트 길이가 예시 수와 정합성 간의 관계에 미치는 조절 효과를 정량적으로 확인하였다.

Results

Effect of Example Count on the agreement Between LLM-Generated and Human Scores

Effect of Example Count on QWK Scores

프롬프트에 포함된 예시 수가 LLM의 채점 결과와 교사의 채점 결과 간의 정렬도에 미치는 영향을 확인하기 위해 6개의 서·논술형 평가 결과 데이터셋으로 선형 회귀 분석을 실시하였다.

우선 각 평가에서 학생 답안에 대한 교사의 채점 결과와 이에 대한 LLM의 채점 결과를 제공된 예시의 숫자에 따라 집단화 한 후 각 집단별로 QWK 점수를 산출하였다. 이를 통해 도출된 QWK 점수를 종속변수로, 예시의 숫자를 독립변수로 설정하여 회귀 분석을 진행하였다.

Table 1에 따르면, 전체 구간에 대해 여섯 개의 서·논술형 평가를 대상으로 예시 수가 QWK 점수에 미치는 영향을 개별적으로 분석한 결과, SS 2와 TH 1를 제외한 모든 평가에서는 예시 수에 따라 QWK 점수가 향상되는 경향이 나타났다. SS 2와 TH 1는 전체 구간에서 통계적으로 유의미한 경향을 확인할 수 없었다. Fig 2에 따르면 시각적으로 보았을 때, SS 2는 예시 수가 많아짐에 따라 초반에 QWK 점수가 급격하게 상승하여 예시

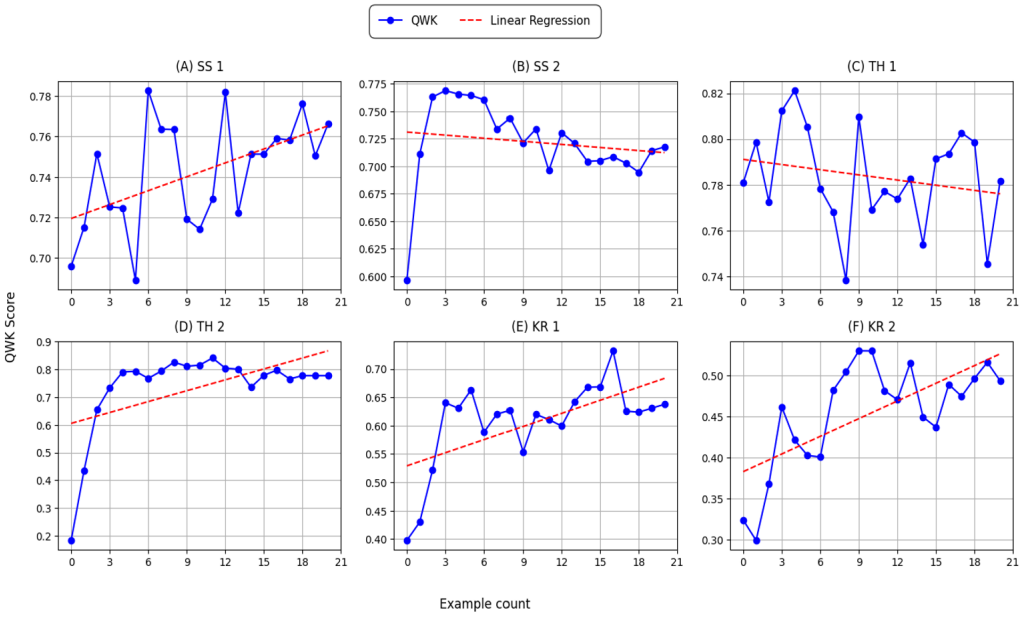

Fig. 2. Effect of example count on qwk score across six datasets (with linear regression trend)

Table 1. Regression results for example count on qwk score across six datasets

Datasets	Coef(B)	SE	t	p	Lower 95%	upper 95%	ΔR^2
SS 1	0.002	0.000	2.658	< 0.001	0.005	0.004	0.232
SS 2	-0.001	0.001	-0.68	0.504	-0.003	0.001	-0.027
TH 1	-0.001	0.001	-0.952	0.352	-0.002	0.001	-0.004
TH 2	0.013	0.004	2.743	< 0.001	0.003	0.023	0.246
KR 1	0.007	0.002	3.479	< 0.001	0.003	0.012	0.357
KR 2	0.007	0.001	4.130	< 0.001	0.003	0.010	0.445

Note. SE = Standard error, p = significance

수 3개에서 가장 높은 점수를 달성한 이후로 점차 QWK 점수가 감소하는 모습을 보였으며 TH 1은 전체적으로 특정한 경향이 존재한다고 보기 어려웠다.

이와 같이 평가에 따라 일부 다른 경향성이 확인되었기에 전체적인 경향성을 분석함에 있어 평가 간 차이를 고려하지 않은 단순 회귀모형은 분석의 타당성을 저해할 수 있다. 실제로 우도비 검정 결과, 평가별 임의 효과를 포함한 혼합효과모형이 고정효과모형보다 통계적으로 유의미하게 우수한 적합도를 보였다($\chi^2 = 114.866$, $df = 2$, $p < 0.001$). 이에 따라, 각 평가를 집단화 변수로 설정하고 평가별 편차는 임의효과로 반영한 혼합효과모형을 통해 집단 간 차이를 통제한 상태에서 전체 데이터의 경향을 분석하였다. 분석 결과(Table 2)에 따르면, 예시 수는 QWK 점수에 통계적으로 유의미한 정적 영향을 미치고 있었다. 이는 평가별 차이를 통제한 상황에서 예시 수가 증가할수록 QWK 점수, 즉 LLM의 채점 정확도가 일관되게 상승함을 의미한다.

Table 2. Mixed effects regression results for example count on qwk score across six datasets

	Coef(B)	SE	t	p	Lower 95%	upper 95%
Intercept	0.626	0.051	12.218	< 0.001	0.005	0.004
example count	0.004	0.001	4.505	< 0.001	0.002	0.006
Random effect(SD)	0.136					

Note. SE = Standard error, SE = Standard Deviation, p = significance

하지만 예시 수에 따른 평가별 QWK 점수의 경향 그래프(Fig. 3)에서 확인할 수 있듯이, 예시 수의 증가에 따른 QWK 점수의 변화는 전체 구간에서 일관된 선형 경향으로 설명되기 어려운 양상을 보였다. 일부 평가 문항(SS 2, TH 1)에서는 급격한 초기 상승 이후 점차 하향하는 형태가 관찰되며, 다른 문항에서는 불규칙한 진폭을 보이기도 했다. 이러한 비선형적이고 구간별로 상이한 경향성은 단일 선형 모형으로는 충분히 설명되기 어려운 측면이 있다. 이에 본 연구에서는 예시 수를 네 개의 구간(Q1: 0–5, Q2: 6–10, Q3: 11–15, Q4: 16–20)으로 나누어 4개의 혼합효과 모형을 구성하고 이를 각각 분석함으로써 예시 수의 변화가 QWK 점수에 미치는 영향을 보다 정밀하게 분석하였다.

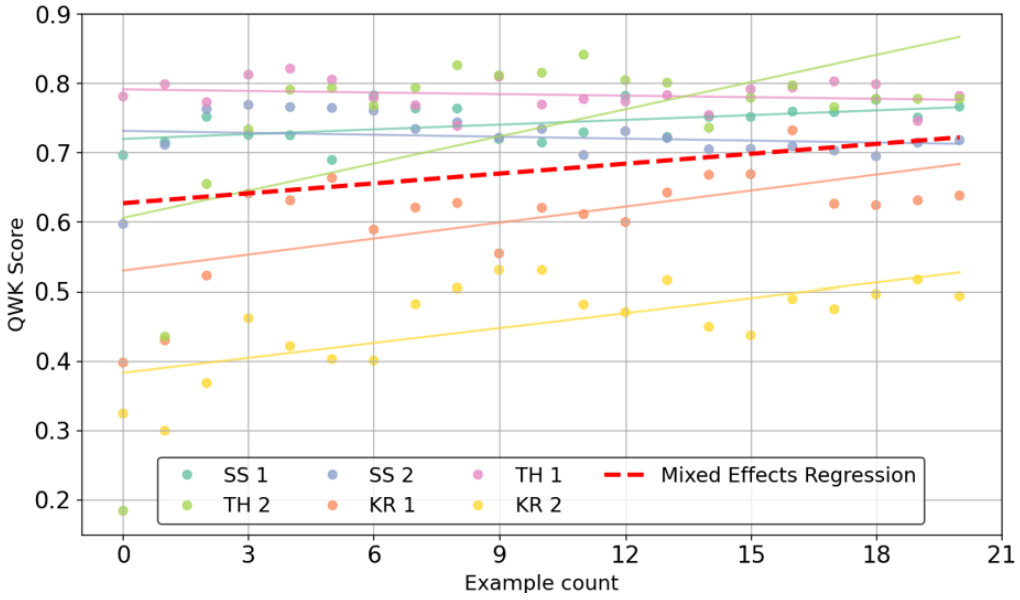

Fig. 3. Effect of example count on qwk score across six datasets (with mixed effects regression)

분석 결과(Table 3)에 따르면 Q1 구간에서 예시 수의 증가가 QWK 점수에 유의미한 정적 영향을 미친 반면, Q2 구간부터는 예시 수의 효과가 통계적으로 유의하지 않았으며, Q3와 Q4 구간에서도 유사한 결과가 나타났다.

Table 3. Mixed effects regression results for the effect of example count on qwk score across binned example ranges

Bin		Coef(B)	SE	p	lower 95%	upper 95%
Q1 (0-5)	Intercept	0.529	0.066	<0.001	0.399	0.658
	example count	0.039	0.008	<0.001	0.022	0.057
	Random effect(SD)					0.174
Q2 (6-10)	Intercept	0.666	0.059	<0.001	0.549	0.783
	example count	0.003	0.004	0.439	-0.004	0.011
	Random effect(SD)					0.318
Q3 (11-15)	Intercept	0.711	0.067	<0.001	0.581	0.843
	example count	-0.001	0.003	0.621	-0.008	0.005
	Random effect(SD)					0.368
Q4 (16-20)	Intercept	0.758	0.071	<0.001	0.620	0.895
	example count	-0.003	0.003	0.245	-0.009	0.002
	Random effect(SD)					0.378

Note. SE = Standard error, SE = Standard Deviation, p = significance

이는 예시 수가 5개 이하로 적은 상황에서는 예시 수가 증가할수록 QWK 점수가 빠르게 향상되는 데 반하여, 예시 수가 6개 이상으로 많은 상황에서는 예시 수가 증가함에 따라 QWK 점수가 유의미한 영향을 받지 못했음을 의미한다.

Moderation Analysis of the Effect of Example Count on QWK Scores

앞서 구간별 분석에서 예시 수의 효과가 일정 수준 이후 감소하는 경향이 확인됨에 따라, 예시가 포함된 프롬프트의 길이(토큰 수)가 이러한 효과를 조절하는지를 검토하기 위해 조절효과 분석을 실시하였다. 통계 분석에 앞서, 프롬프트 길이와 예시 수를 활용한 회귀분석의 적합성을 확인하기 위해 두 변수 간 VIF(Variance inflation factor)를 계산하였다. 그 결과 VIF가 2.3으로 나타나 유의미한 다중공선성은 없는 것으로 판단하였다. 이후 예시 수와 프롬프트 길이를 평균 중심화한 후, 상호작용 항을 생성하여 혼합효과모형에 포함시켰다. 이때 평가 간 차이는 집단화 변수로 통제하였고, 평가별 편차는 임의효과로 반영하였다.

분석 결과(Table 4)에 따르면, 프롬프트 길이의 영향을 통제한 상태에서 예시 수의 주효과는 통계적으로 유의하지 않았으나, 예시 수와 프롬프트 길이 간의 상호작용 효과는 유의미한 부적 효과를 나타냈다. 이는 프롬프트가 길어질수록 예시 수의 증가가 QWK 점수에 미치는 긍정적 영향이 약화될 수 있음을 시사한다.

Table 4. Results of the moderation analysis for the effect of example count on QWK score by prompt length

Variables	Coef(β)	SE	t	p	lower-95%	upper-95%
Con	0.694	0.052	13.196	< 0.001	0.591	0.797
example count	0.001	0.003	0.528	0.598	-0.004	0.008
prompt length	1.7×10^{-5}	1.6×10^{-5}	1.044	0.029	-1.5×10^{-5}	4.8×10^{-5}
example count × prompt length	-2.9×10^{-6}	9.0×10^{-7}	-3.1897	0.002	-4.7×10^{-6}	-1.1×10^{-6}
Random effect(SD)						0.157

Note. SE = Standard error, SE = Standard Deviation, p = significance

이러한 상호작용 효과를 보다 구체적으로 확인하고자, 프롬프트 길이를 평균에서 –1SD(706 tokens), 평균(2264 tokens), +1SD(3822 tokens)로 고정한 조건부 분석을 추가로 수행하였다. 그 결과(Table 5), 토큰 수가 적은 조건에서는 예시 수의 증가가 QWK 점수에 유의한 정적 영향을 미치는 것으로 나타났으나 평균 및 높은 토큰 조건에서는 통계적으로 유의미한 효과가 확인되지 않았다. 이는 프롬프트가 짧을 때에는 예시 수 증가가 정합성 향상에 기여하지만, 프롬프트가 평균 이상으로 길어질 경우 그 효과가 미미하거나 반대로 약화될 가능성이 있음을 의미한다.

Table 5. Results of the conditional analysis for the effect of example count on QWK score by prompt length

Length of tokens	Coef(β)	SE	t	p	lower-95%	upper-95%
706	0.006	0.002	2.414	0.017	0.001	0.011
2264	0.001	0.003	0.528	0.598	-0.004	0.007
3822	-0.002	0.004	-0.659	0.511	-0.011	0.005

Note. SE = Standard error, p = significance

Effect of Example Count on the agreement Between AI-Generated and Human Feedback

Effect of Example Count on BERTScore

프롬프트에 포함된 교사 피드백 예시 수가 LLM이 생성한 피드백과 교사의 피드백 간의 정합성에 미치는 영향을 확인하기 위해, 6개의 서·논술형 평가 결과 데이터를 대상으로 선형 회귀 분석을 실시하였다.

피드백 간의 의미적 정합성은 BERTScore를 활용하여 측정하였으며, 이때 비교 대상인 피드백이 모두 한국어로 작성되었기 때문에 한국어 표현의 의미적 정밀도를 더 잘 반영할 수 있는 다국어 사전학습 언어 모델인 XLM-RoBERTa-large를 활용하여 이를 계산하였다.

평가별 선형회귀 분석 결과 표(Table 6)에 따르면, 전체 구간에 대해 여섯 개의 서·논술형 평가를 대상으로 예시 수가 BERTScore에 미치는 영향을 개별적으로 분석한 결과, KR 2를 제외한 모든 평가에서 예시 수가 BERTScore를 증가시키는 데 있어 통계적으로 유의미한 영향을 미치는 것으로 나타났다. Fig 4에 따르면, 통계적으로 유의미한 결과를 관찰할 수 없었던 KR 2에서는 예시 6개 이상의 구간부터 평균 정합성이 감소하거나 불안정한 추세를 보이는 등 비선형적 경향이 일부 관찰되는 등 특정한 경향이 존재한다고 보기 어려웠다. 이러한 경향은 실제 예시 수에 따라 생성된 피드백 결과의 예(Table 7)를 보면 더욱 명료하게 드러난다. 대부분의 평가와 비슷한 양상을 보였던 TH 1의 예를 살펴보면, 예시가 전혀 없는 상황에서 LLM은 교사의 평가 관점이나 표현방식과는 관계없이 피드백을 제공함으로써 낮은 BERTScore를 얻었으나, 예시 수가 점차 증가함에 따라 교사의 평가 관점 및 표현방식을 반영한 피드백을 제공함으로써 높은 BERTScore를 획득하였다.

Table 6. Regression results for example count on bertscore across six datasets

Datasets	Coef(B)	SE	t	p	Lower 95%	upper 95%	ΔR^2
SS 1	0.001	0.0000	7.7172	< 0.001	0.0004	0.0006	0.0539
SS 2	0.001	0.001	10.8404	< 0.001	0.001	0.001	0.101
TH 1	0.002	0.001	18.1816	< 0.001	0.001	0.002	0.17
TH 2	0.003	0.001	26.1227	< 0.001	0.003	0.004	0.34
KR 1	0.001	0.001	4.5435	< 0.001	0.0002	0.001	0.018
KR 2	0.001	0.001	0.6536	0.5135	-0.0001	0.001	-0.001

Note. SE = Standard error, p = significance

이에 반해 비선형적인 경향을 보였던 KR 2의 예를 살펴보면, LLM은 예시 수가 5개일 때 교사의 평가 관점이나 표현방식을 반영한 피드백을 제공하여 높은 BERTScore를 획득하였으나, 15개의 예시를 제공했을 때는 오히려 교사의 피드백과 멀어지는 것을 확인할 수 있다.

이처럼 평가별로 일부 다른 경향성이 관찰되므로 전체적인 경향성을 분석함에 있어 평가 간 차이를 고려하지 않은 단순 회귀모형은 분석의 타당성을 저해할 수 있다. 실제로 우도비 검정 결과, 평가별 임의 효과를 포함한 혼합효과모형이 고정효과모형보다 통계적으로 유의미하게 우수한 적합도를 보였다($\chi^2 = 4557.343$, $df = 2$, $p < 0.001$). 이에 따라, 각 평가를 집단화 변수로 설정하고 평가별 편차는 임의효과로 반영한 혼합효과모형을 통해 집단 간 차이를 통제한 상태에서 전체 데이터의 경향을 분석하였다.

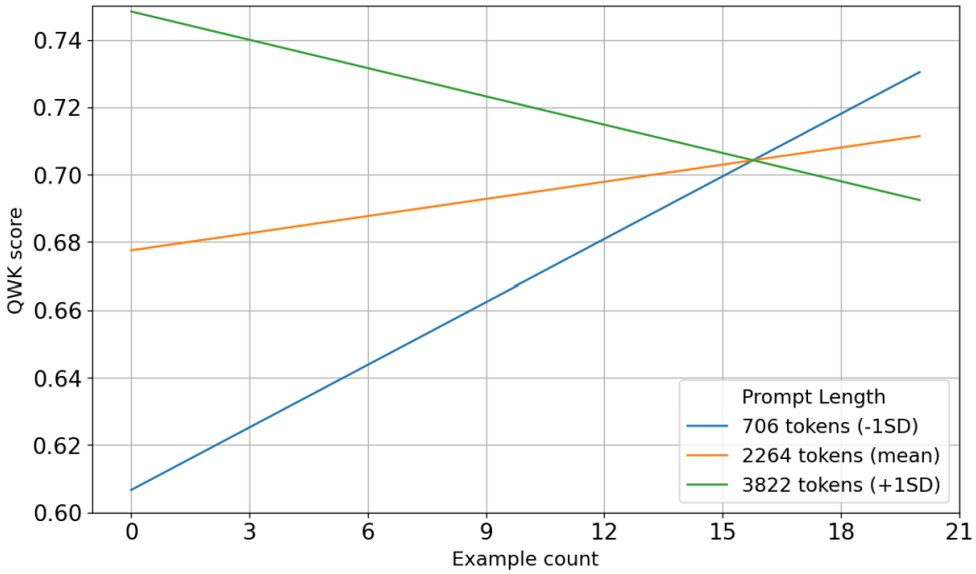

Fig. 4. Effect of example count on qwk score at different prompt lengths

Table 7. Examples of changes in llm-generated feedback and corresponding bertscore by example count

Code	Human feedback	Example count	LLM generated feedback	BERTScore
TH 1	It is necessary to clearly explain how this challenge contributes to reducing carbon dioxide emissions. Additionally, it should be explained how reducing CO_2 emissions helps support a sustainable Earth system.	0	This response mentions reducing single-use products and carbon dioxide emissions as ways to protect the environment, but it lacks specific explanations or examples. It needs further elaboration on why each method is important and how it can be put into practice. For example, suggesting concrete actions like using a tumbler or a reusable shopping bag to reduce single-use items would strengthen the answer. Additionally, including specific behaviors to reduce CO_2 emissions—such as using public transportation or conserving energy—would make the response more comprehensive.	0.896
		5	It is necessary to explain how reducing single-use products is connected to the reduction of carbon dioxide emissions. Additionally, it should be explained how such changes impact the sustainability of the Earth's systems.	0.958
		15	It is necessary to clearly explain which specific activities of the challenge lead to a reduction in carbon dioxide emissions. Additionally, it should be described how these changes contribute to the sustainability of the Earth's systems	0.958

Table 7. Examples of changes in llm-generated feedback and corresponding bertscore by example count(continued)

Code	Human feedback	Example count	LLM generated feedback	BERTScore
KR 1	There is no clear statement of the creative intention or the reasons behind it, so the response could be improved by clearly stating the intention at the beginning and supporting it with a specific reason.	0	This response does not clearly state the purpose of the writing, nor does it provide sufficient reasons or supporting evidence. While the reason for choosing the creative experience and the intention behind the piece should be explained separately, the two are mixed together, making it unclear what the actual creative intention is. Although symbolic expressions based on personal experience are mentioned, there is a lack of specific explanation or justification for why such expressions were used or what message they were meant to convey. According to the evaluation criteria, both the statement of creative intention and its supporting reasons are insufficient, corresponding to a score of 1.	0.886
		5	The creative intention and the reason for choosing the experience are not clearly presented. To strengthen the response, the writer should clearly express the purpose at the start and provide a concrete reason to support it.	0.946
		15	The creative intention and the reason for selecting the experience are not clearly presented. The intention behind the writing—what message the student aimed to convey through this experience—was not explained separately, and the reason for selection is limited to merely listing the experience itself. In addition, there is a lack of supporting reasons or evidence to reinforce the creative intention. In future writing, please clearly distinguish between the creative intention and the reason for selection, and provide specific explanations or justifications that support the intended message.	0.887

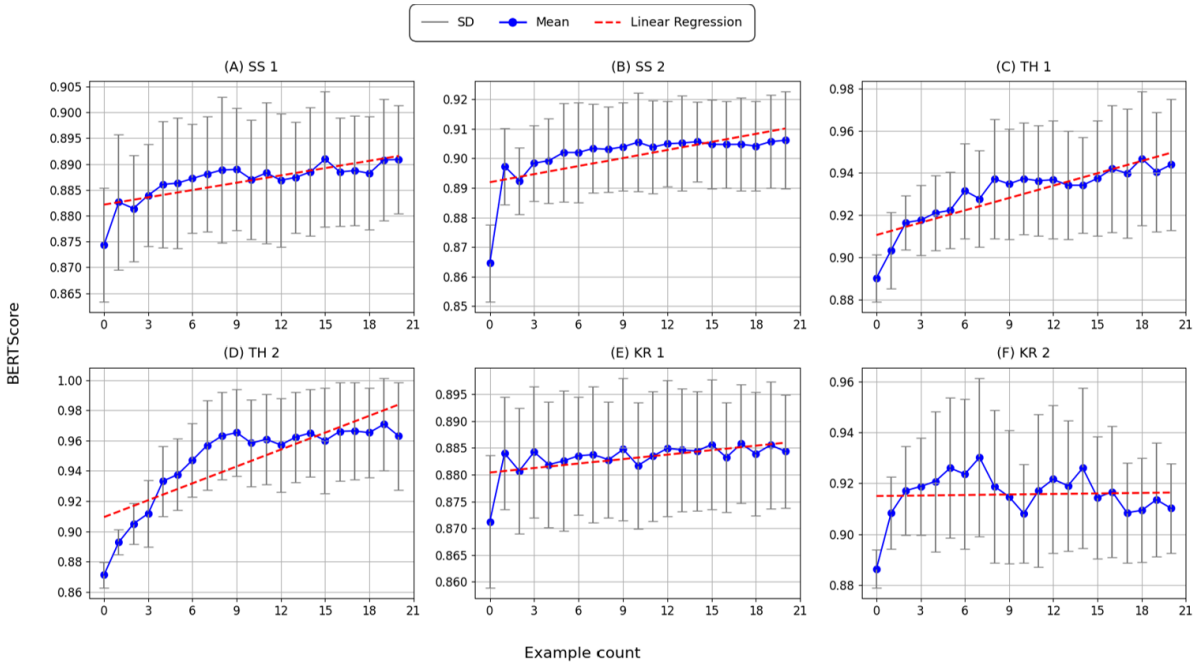

Fig. 5. Effect of example count on bertscore across six datasets (with linear regression trend)

그 결과(Table 8)에 따르면 예시 수는 BERTScore에 통계적으로 유의한 정적 영향을 미치는 것으로 나타났다. 즉, 평가 간 차이를 통제한 상태에서도 예시 수가 증가할수록 BERTScore가 일관되게 상승하는 경향이 확인되었다.

하지만 Fig. 6에 제시된 시각화 결과에서 알 수 있듯이, 예시 수의 증가에 따른 BERTScore의 변화는 전체 구간에서 일관된 선형 경향으로 설명되기 어려운 양상을 보였다. 일부 평가에서는 급격한 초기 상승 이후 점차 평준화되는 곡선 형태가 관찰되며, 다른 문항에서는 불규칙한 진폭을 보이기도 했다. 이러한 비선형적이고 구간별로 상이한 경향성은 단일 선형 모형으로는 충분히 설명되기 어려운 측면이 있다.

Table 8. Mixed effects regression results for example count on bertscore across six datasets

	Coef(B)	SE	t	p	Lower 95%	upper 95%
Intercept	0.897	0.01	88.227	< 0.001	0.877	0.917
example count	0.001	0.0001	27.899	< 0.001	0.001	0.001
Random effect(SD)						0.016

Note. SE = Standard error, SE = Standard Deviation, p = significance

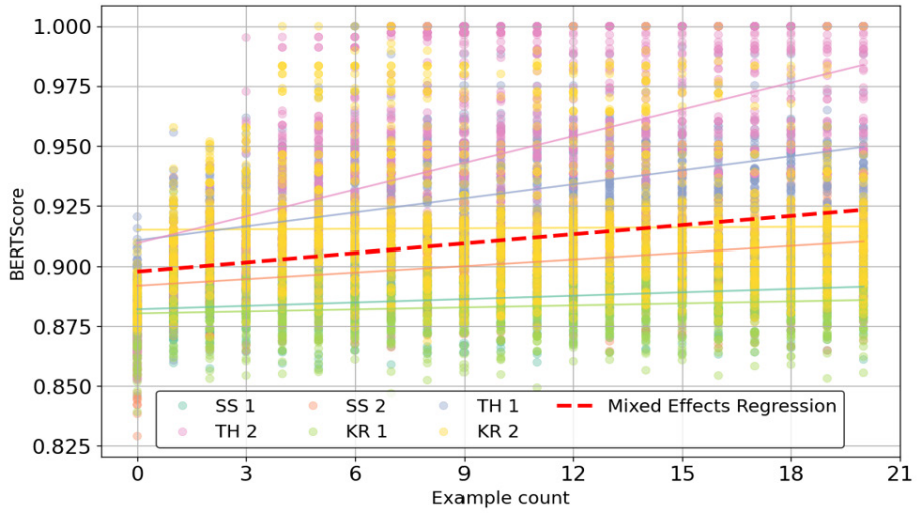

Fig. 6. Effect of example count on bertscore across six datasets (with mixed effects regression)

이러한 경향을 보다 정밀하게 분석하기 위해, 예시 수를 네 개의 구간(Q1: 0–5, Q2: 6–10, Q3: 11–15, Q4: 16–20)으로 나누어 각각에 대해 별도의 혼합효과모형을 구성하여 분석하였다. 그 결과(Table 9), Q1 구간에서 예시 수는 BERTScore에 유의미한 정적 영향을 미쳤으나, Q2 이후의 모든 구간에서는 예시 수의 효과가 통계적으로 유의하지 않은 것으로 나타났다. 이는 예시 수가 5개 이하로 적은 상황에서는 예시 수가 증가할수록 BERTScore가 향상되는 데 반하여, 예시 수가 6개 이상으로 많은 상황에서는 예시 수의 증가가 BERTScore에 유의미한 영향을 주지 못했음을 의미한다.

Table 9. Mixed effects regression results for the effect of example count on bertscore across binned example ranges

Bin		Coef(B)	SE	p	lower 95%	upper 95%
Q1	Intercept	0.882	0.006	< 0.001	0.87	0.894
(0-5)	example count	0.006	0.001	< 0.001	0.005	0.006
	Random effect(SD)					0.008
Q2	Intercept	0.914	0.012	< 0.001	0.891	0.938
(6 -10)	example count	-0.0001	0.001	0.89	-0.001	0.001
	Random effect(SD)					0.023
Q3	Intercept	0.912	0.013	< 0.001	0.886	0.937
(11 -15)	example count	0.0003	0.001	0.505	-0.001	0.001
	Random effect(SD)					0.023
Q4	Intercept	0.915	0.014	< 0.001	0.886	0.944
(16 -20)	example count	0.0001	0.001	0.886	-0.001	0.001
	Random effect(SD)					0.028

Note. SE = Standard error, SE = Standard Deviation, p = significance

Moderation Analysis of the Effect of Example Count on BERTScore

예시 수가 BERTScore에 미치는 영향에서 프롬프트 길이의 조절 효과를 확인하기 위해 평균 중심화된 예시 수와 프롬프트 길이를 변수로 설정하고, 이들의 상호작용 항을 포함한 혼합효과모형 분석을 실시하였다. 통계 분석에 앞서, 프롬프트 길이와 예시 수를 활용한 회귀분석의 적합성을 확인하기 위해 두 변수 간 VIF를 계산하였다. 그 결과 VIF가 2.65로 나타나 두 변수간에 유의미한 다중공선성은 없는 것으로 판단하였다. 이후 예시 수와 프롬프트 길이를 평균 중심화한 후, 상호작용 항을 생성하여 혼합효과모형에 포함시켰다. 이때 각 평가 항목을 집단화 변수로 포함하여 평가 간 차이를 통제하고, 평가별 편차는 임의효과로 반영하였다.

분석 결과(Table 10), 프롬프트 길이의 영향을 통제한 상태에서 예시 수가 BERTScore에 미치는 영향은 통계적으로 유의하였다 반면, 예시 수와 프롬프트 길이 간의 상호작용 효과는 통계적으로 유의한 부적 효과를 보였다. 이는 예시 수가 BERTScore에 미치는 긍정적 효과가 프롬프트 길이가 길어짐에 따라 점차 감소함을 시사하며, 지나치게 긴 프롬프트는 예시의 효과를 약화시킬 수 있음을 의미한다.

이러한 상호작용 효과를 보다 정밀하게 확인하기 위하여, 프롬프트 길이(토큰 수)를 고정한 조건부 분석을 추가로 실시하였다. 분석은 토큰 수를 각각 평균에서 -1SD(754 tokens), 평균값(2447 tokens), 평균에서 +1SD(4139 tokens) 수준으로 설정한 뒤, 해당 조건별로 예시 수가 BERTScore에 미치는 영향을 선형 회귀 모형을 통해 추정하는 방식으로 진행되었다.

그 결과(Table 11), 프롬프트 길이에 관계없이 예시 수는 BERTScore 향상에 모두 통계적으로 유의미한 정적 효과를 보였으나 프롬프트가 길어짐에 따라 이러한 정적 효과가 점차 감소하는 경향을 보였다.

이는 예시를 포함한 전체 프롬프트 길이가 짧을수록 예시 수가 BERTScore 향상에 보다 크게 기여하는 반면, 프롬프트 길이가 길어질수록 예시 수의 효과는 점진적으로 감소함을 시사한다.

Table 10. Results of the moderation analysis for the effect of example count on BERTScore by prompt length

Variables	Coef(β)	SE	t	p	lower-95%	upper-95%
Con	0.915	0.007	125.181	< 0.001	0.901	0.929
example count	0.003	0.0001	21.186	< 0.001	0.002	0.003
prompt length	-8.2×10^{-6}	6.4×10^{-7}	-12.8795	< 0.001	-9.5×10^{-6}	-7.0×10^{-6}
example count × prompt length	-4.9×10^{-7}	3.3×10^{-8}	-14.6189	< 0.001	-5.5×10^{-7}	-4.2×10^{-7}
Random effect(SD)						0.009

Note. SE = Standard error, SE = Standard Deviation, p = significance

Table 11. Effect of example count on BERTScore at different prompt lengths

Length of tokens	Coef(β)	SE	t	p	lower-95%	upper-95%
754	0.003	0.0001	31.886	< 0.001	0.003	0.004
2447	0.003	0.0001	21.188	< 0.001	0.002	0.003
4139	0.002	0.0001	12.335	< 0.001	0.001	0.002

Note. SE = Standard error, p = significance

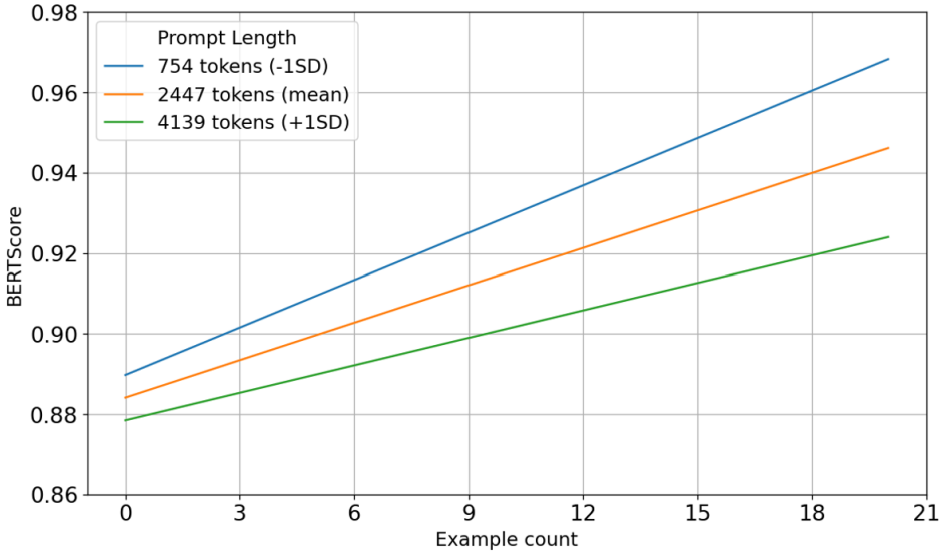

Fig. 7. Effect of example count on bertscore at different prompt lengths

Discussions

본 연구는 서·논술형 평가에서 채점과 피드백 작성을 수행함에 있어 예시의 제공이 교사와 LLM 간의 채점 및 피드백 정합성 향상에 유의미한 영향을 미칠 수 있음을 실증하였다. 먼저 프롬프트에 포함된 예시의 수가 정합성에 미치는 영향을 선형회귀 분석 및 혼합효과 회귀분석을 통해 확인하였으며, 주요 결과는 다음과 같다.

첫째, 프롬프트에 포함될 채점 결과 및 피드백 예시의 수가 0개에서 20개까지 점차 증가함에 따라 채점과 피드백 모두에서 교사와 LLM 간의 정합성이 향상되는 경향이 확인되었다. 이 결과는 프롬프트에 포함된 예시가 많아질수록 LLM이 교사의 채점 및 피드백 관점을 더 잘 모방하게 되었음을 의미하며, ICL이 LLM의 응답 방향을 사용자 의도에 맞추는 데 효과적이라는 기존 연구들의 결과와도 일치한다(Brown et al., 2020; Cai et

al., 2024; Ma et al., 2023; Min et al., 2022). 그러나 일부 문항에서는 예시 수가 일정 수준 이상 증가했을 때 오히려 정합성이 감소하거나 불안정한 경향도 나타났는데, 이는 문항 구성, 평가 기준의 복잡성 등 문항별 특성에 기인할 가능성이 높다(Yoshida, 2024).

둘째, 예시 수 증가의 효과는 구간별로 차이가 있었다. 예시 수가 5개 이하(Q1)일 때는 정합성이 급격히 향상되었지만, 이후(Q2 이상)의 구간에서는 효과가 크게 둔화되거나 거의 소실되었다. 특히 11개 이상의 예시(Q3, Q4)에서는 회귀계수가 0에 수렴하고 통계적 유의성도 사라져, 일정 수준 이상에서는 적은 수의 예시 제공이 성능 향상에 큰 기여를 하지 못함을 시사했다. 이러한 증가 추이의 변화는 Agarwal et al.(2024)에서 LLM이 여러 과제에서 예시 2~8개를 제공받았을 때 가장 급격하게 성능 증가를 보이던 것과 흡사하며, 이는 LLM이 소수의 예시만으로도 교사의 핵심적인 평가 관점을 빠르게 학습할 수 있음을 보여준다. 하지만 동시에 일정 개수 이상부터는 단순히 예시를 더 추가하는 방식으로 교사의 평가 관점을 더 명확하게 이해시키는 것에 한계가 있음을 드러내기도 한다. 예시의 추가만으로 더 이상 성능이 향상되지 않는 포화 지점(Plateau)에 대해서는 LLM에게 부여된 과제의 종류 또는 ICL의 세부 적용 방식에 따라 차이가 존재한다(Agarwal et al. 2024; Brown et al., 2020; Chandra et al., 2025; Zhang et al. 2023). 따라서 평가 루브릭이나 평가 문항 또는 답변의 질적 차이에 근거한 ICL의 효과에 대한 추가 연구가 필요하며, 루브릭 점수대별 대표성에 근거한 예시 배치 전략과 같은 ICL의 방법론적 개선에 대한 연구 또한 이루어질 필요가 있다.

이어서 프롬프트에 포함된 토큰의 길이가 예시 수의 정합성 향상 효과를 조절할 것이라는 가정 하에 조절분석을 수행하였으며, 이를 통해 채점 결과와 피드백 모두에서 예시 수와 프롬프트 길이 간에 부정적인 상호작용 효과를 확인하였다. 이는 프롬프트의 길이의 증가가 동일 과제를 수행함에 있어 LLM의 성능에 부정적인 영향을 미친다는 기존 연구들의 결과와 일맥상통한다(Levy et al., 2024).

이상의 결과는 실제 교육 현장에서 교사가 LLM을 활용하여 서·논술형 평가에 대한 채점과 피드백 작성을 수행할 때, 교사가 취해야 할 프롬프트 작성 전략에 시사점을 제공한다. 우선 교사는 일부 학생 답안에 대한 자신의 채점 결과와 피드백을 예시로서 프롬프트에 포함시킴으로써 LLM이 교사의 채점 기준과 피드백 방식을 모방하도록 할 수 있다. 이를 통해 서·논술형 평가가 가지던 시간적·인적 자원의 한계(Narciss, 2013)를 극복하고, 일관된 관점의 채점 및 피드백 작성(Huang & Whipple, 2023)을 수행할 수 있다. 또한 교사가 프롬프트에 포함시켜야 하는 예시의 수는 평가 문항의 특성과 요구되는 답안의 길이에 따라 차별적으로 설계될 필요가 있다. 짧은 답안이 요구되는 문항의 경우 예시 수를 충분히 늘려 교사의 평가 기준을 명확하게 전달하는 것이 효과적인 반면, 길고 복잡한 답안을 요구하는 문항에서는 오히려 5개 이내로 예시 수를 제한하는 것이 더 효과적일 수 있다.

이러한 시사점에도 불구하고, 본 연구에는 몇 가지 한계가 존재한다. 첫째, 사용된 문항과 평가 기준이 특정 과목과 서·논술형 문항 유형에 국한되어 있다는 점에서, 다양한 과목이나 과제 유형으로 결과를 일반화하는 데에는 제한이 있다. 둘째, 교사의 채점 결과와 피드백을 단순히 프롬프트에 나열하는 방식으로 예시를 제공하였기 때문에, 보다 정교한 프롬프트 설계나 예시 선정 전략의 효과를 확인하기는 어려웠다. Zhao et al.(2021)과 Kumar & Talukdar(2021)의 연구에 따르면, ICL의 성능은 예시의 구성 방식이나 위치, 표현 구조 등에 따라 민감하게 달라질 수 있으므로, 향후에는 이러한 요소들을 통제하거나 비교하는 연구가 요구된다. 셋째, 본 연구에서는 데이터 제약으로 인해 최대 20개의 예시까지만 실험을 수행하였으나, Agarwal et al. (2024)의 Many-Shot ICL 연구에서는 수백 개에서 수천 개의 예시를 활용한 경우 일부 과제에서 추가적인 성능 향

상이 관찰되었다. 따라서 충분한 데이터가 확보된 환경에서 Many-Shot ICL을 적용한 서·논술형 평가 연구를 통해 본 연구 결과의 확장 가능성을 탐구할 필요가 있다. 넷째, 본 연구에서는 교사와 LLM 피드백 간의 정합성을 측정하기 위해 BERTScore를 활용하였으나 BERTScore는 의미적 유사성에만 초점을 두기 때문에 이를 근거로 LLM이 교사 피드백의 질적 특성까지 충실히 재현한다고 보기는 어렵다. 따라서 추후 연구에서는 BERTScore와 같은 의미 기반 지표를 보완할 수 있는 평가 방법을 모색할 필요가 있으며, 하나의 지표에만 의존하기보다 여러 지표를 종합적으로 활용함으로써 ICL이 LLM의 피드백에 미치는 효과에 대해 보다 다층적으로 검토할 필요가 있다.

따라서 추후 연구를 통해 다양한 과목과 문항 유형을 포함하여 LLM과 교사 간 정합성 향상 효과의 일반화 가능성을 검토할 필요가 있다. 특히 서·논술형 평가의 문항 특성과 평가 기준에 따라 정합성 향상 효과가 어떻게 달라지는지에 대한 체계적인 분석이 요구된다. 더 나아가, 본 연구에서는 예시의 개수에만 초점을 두었으나, 예시의 선정 방식이 결과에 영향을 줄 가능성도 존재한다. 예컨대 점수대별로 고르게 예시를 제공하거나 특정 평가 기준을 강조하는 방식은 무작위 추출과는 다른 양상을 보일 수 있다. 따라서 향후 연구에서는 예시의 개수뿐만 아니라 선정 방식까지 함께 고려하여, LLM이 교사의 채점 및 피드백 관점을 더욱 정밀하게 모방할 수 있는 최적의 전략을 탐색할 필요가 있다. 또한 Chandra et al.(2025)이 제안한 바와 같이 문항 유형이나 답안의 길이에 따라 최적의 예시 수를 자동으로 제안하는 방법론에 대한 연구도 이루어질 필요가 있다.

Conclusions

본 연구는 교사의 채점 및 피드백 예시를 활용한 ICL이 서·논술형 평가에서 LLM의 정합성 향상에 유의미한 효과를 가질 수 있음을 실증적으로 확인하였다. 예시 수가 증가할수록 교사와 LLM 간의 정합성은 향상되었으나, 그 효과는 5개 예시 이후 급격히 둔화되었으며, 프롬프트의 길이가 길어질수록 정합성 향상 효과는 감소하였다. 이는 서·논술형 평가에서 ICL을 활용함에 있어 예시 수와 프롬프트 길이의 균형이 핵심 요소임을 시사한다. 따라서 교육 현장에서 LLM 기반 자동 채점 및 피드백 시스템을 효과적으로 활용하기 위해서는 문항의 특성과 답안의 예상 길이에 따라 최적의 예시 수와 프롬프트 구성을 설계할 필요가 있다.

Conflicts of Interest

The authors declared no conflicts of interest.

References

Agarwal, R., Singh, A., Zhang, L. M., Bohnet, B., Rosias, L., Chan, S. C. Y., ... Larochelle, H. (2024). Many-shot in-context learning. arXiv preprint arXiv:2404.11018

Baz, M. & Has ı rc ı , A. S. (2025). The effect of feedback on informative text writing: ai or teacher? Open Praxis, 17, 611-631. https://doi.org/10.55982/openpraxis.17.3.871

Brookhart, S. M. (2010). How to Assess Higher-Order Thinking Skills in Your Classroom. ASCD.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, 33, 1877–1901.

Cai, S., Mensah-Boateng, T., Kuksov, X., Yuan, J., & Tang, S. (2024). The power of adaptation: Boosting in-context learning through adaptive prompting. arXiv preprint arXiv:2403.08516.

Chandra, M., Ganguly, D., & Ounis, I. (2025). One size doesn't fit all: Predicting the number of examples for in-context learning. Lecture Notes in Computer Science, 67–84. https://doi.org/10.1007/978-3-031-88708-6_5

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin, 70, 213–220. https://doi.org/10.1037/h0026256

Dai, W., Liu, Y., Xia, M., & Wang, C. (2023). Evaluating ChatGPT as a feedback generator for data science writing. International Journal of Educational Technology in Higher Education, 20, 24. https://doi.org/10.1186/s41239-023-00420-1

Doewes, A., Kurdhi, N. A., & Saxena, A. (2023). Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. Proceedings of the 16th International Conference on Educational Data Mining, 103–113. https://doi.org/10.5281/zenodo.8115784

Hattie, J., & Timperley, H. (2007). The power of feedback. Review of Educational Research, 77, 81–112. https://doi.org/10.3102/003465430298487

Huang, X., Tan, S., & Shi, Y. (2020). Automated essay scoring: A survey of the state of the art. IEEE Transactions on Learning Technologies, 13, 802–817. https://doi.org/10.24963/ijcai.2019/879

Huang J., & Whipple, P. (2023). Rater variability and reliability of constructed response questions in New York state high-stakes tests of English language arts and mathematics: implications for educational assessment policy. Humanities and Social Sciences Communications, 10, 1-11. https://doi.org/10.1057/s41599-023-02385-4

Kumar, S., & Talukdar, P. (2021). Reordering examples helps during priming-based few-shot learning. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. 4507-4518. https://doi.org/10.18653/v1/2021.findings-acl.395

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 1, 15339–15353. https://doi.org/10.18653/v1/2024.acl-long.818

Ma, H., Zhang, C., Bian, Y., Liu, L., Zhang, Z., Zhao, P., Zhang, S., Fu, H., Hu, Q., & Wu, B. (2023). Fairness-guided few-shot prompting for large language models. NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems, 1870, 43136-43155. https://doi.org/10.48550/arXiv.2303.13217

Mazzullo, E., & Bulut, O. (2025). Automated feedback generation for open-ended questions: Insights from fine-tuned LLMs. Proceedings of Machine Learning Research, 264, 103-120. https://proceedings.mlr.press/v264/mazzullo25a.html.

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). Rethinking the role of demonstrations: What makes in-context learning work? Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 11048–11064. https://doi.org/10.18653/v1/2022.emnlp-main.759

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. Digital Education Review, 7–26. https://files.eric.ed.gov/fulltext/EJ1013726.pdf

Nicol, D. J., & Macfarlane‐Dick, D. (2006). Formative assessment and self‐regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education, 31, 199–218. https://doi.org/10.1080/03075070600572090

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability, Computers and Education Artificial Intelligence, 6, 100234. http://doi.org/10.1016/j.caeai.2024.100234

Parr, J. M. & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. Assessing Writing, 15, 68–85. https://doi.org/10.1016/j.asw.2010.05.004

Shin, B., Lee, J., & Yoo, Y. (2024) Exploring automatic scoring of mathematical descriptive assessment using prompt engineering with the GPT-4 model: Focused on permutations and combinations. The Mathematical Education, 63, 187-207. https://doi.org/10.7468/mathedu.2024.63.2.187

Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78, 153–189. https://doi.org/10.3102/0034654307313795

Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. Physical Review Physics Education Research, 20, 010152. https://doi.org/10.1103/PhysRevPhysEducRes.20.010152

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le ,Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

Yoshida, L. (2024). The impact of example selection in few-shot prompting on automated essay scoring using GPT Models. Communications in Computer and Information Science, 61–73. https://doi.org/10.1007/978-3-031-64315-6_5

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of Artificial Intelligence in Education. Computers and Education: Artificial Intelligence, 4, 100118. https://doi.org/10.1016/j.caeai.2022.100118

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. International Conference on Learning Representations (ICLR), 1-41. https://openreview.net/forum?id=SkeHuCVFDr

Zhang, X., Hsu, J., Lu, Y., & Hsu, T. (2023). CODE4STRUCT: Code Generation for Few-Shot Event Structure. Proceedings of the Association for Computational Linguistics (ACL), 3640-3663. https://doi.org/10.18653/v1/2023.acl-long.202

Zhao, T. Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models. Proceedings of the 38th International Conference on Machine Learning, 139, 12697-12706, https://proceedings.mlr.press/v139/zhao21c.html

Authors Information

Juyeong Lee: Seoul National University, Graduate Student, 1st Author.
https://orcid.org/0009-0007-0374-360X

Euna Lee: Osan Wonil Middle School, Teacher, Co-Author.
https://orcid.org/0009-0003-7818-7258

Kangrea Kim: Sundong Elementary School, Teacher, Co-Author.
https://orcid.org/0009-0009-6514-7998

Appendix

Appendix A.

Code	Assessment item
SS1	Evaluate the effectiveness of the government's proposed measures to solve the problem described in the case you studied. In your answer, explain both the strengths and weaknesses of the measures, and suggest possible improvements if needed.
SS2	Choose one travel destination you would like to visit and describe its landscape characteristics, climate, and geographical features based on how its landforms were created.
TH1	Clearly describe how your group's challenge activity affects the sustainable Earth system, demonstrating a thorough understanding of its impact. - Identify the main aspects of the Earth system that are influenced. - Summarize both the positive and negative impacts.
TH2	Compare the social system in reality with the one shown in the movie where the nature of time as a resource is different. In your answer, analyze the differences in economy, social structure, jobs, and life system, explain what is considered valuable in the movie's social system and why, and reflect on what this comparison tells us about our own society.
KR1	Write a piece explaining your own creative work. <Conditions> - You must clearly state your creative intent and support it with appropriate and well-justified reasons or evidence.
KR2	Read the two texts, (A) and (B), which present opposing viewpoints. Then, logically express your own opinion by maintaining coherence and providing appropriate supporting reasons.

Appendix B.

Code	Criterion	Level	
SS1	Critical analysis of social issues with logical reasoning and evidence	5	Critically examines the issue and evaluates the validity and limitations of the government's measures in a balanced way. The reasoning is clear, logical, and expressed with strong clarity.
		4	Provides a logical evaluation of the government's measures, but the critical perspective is somewhat limited or the analysis is partially superficial.
		3	Mentions the government's measures but offers only a shallow critical analysis, with simple or underdeveloped reasoning.
		2	Provides little to no validity analysis, and the reasoning is unclear or weakly structured.
		1	Shows no critical perspective and remains at the level of simple description.
SS2	Comprehensive explanation of geographical characteristics through analysis of landform formation processes	5	Accurately analyzes landform formation processes and clearly explains causal relationships between landscape, climate, and geographical features with specific evidence.
		4	Generally explains landform formation and connects most geographical characteristics, but some relationships lack clarity or detail.
		3	Mentions landform formation but geographical explanations are superficial, or connections between elements are limited.
		2	Simple listing without no analytical approach.
		1	Does not meet any of the above levels.
TH1	Explanation of the causal relationship between the challenge activity and Earth system sustainability	5	Clearly explains the causal relationship between the group's challenge activity and its impacts on the sustainable Earth system. The explanation demonstrates clarity, logical coherence, and a thorough understanding of the interconnections.
		4	Describes the impacts of the group's challenge activity on the sustainable Earth system, but the causal relationship or interconnections are not clearly articulated.
		3	Provides a description of the impacts of the group's challenge activity on the sustainable Earth system, but without further elaboration of causal links.
		2	Mentions only the results of the group's challenge activity or only the impacts on the sustainable Earth system, without connecting the two
		1	Does not meet any of the above levels.

Appendix B. (continued)

Code	Criterion	Level	
TH2	Analysis of differences between real-world and film social systems from multiple perspectives	5	Writes about the differences between real-world and film social systems from two or more diverse perspectives (economic, social structure, occupational, biological systems, etc.) and writes about the value criteria of the social system within the film.
		4	Writes about the differences between real-world and film systems from two or more diverse perspectives (economic, social structure, occupational, biological systems, etc.), OR writes about the differences between real-world and film systems from one perspective and the value criteria of the social system within the film.
		3	Writes about the differences between real-world and film systems from one perspective.
		2	Fails to clearly write about the differences between real-world and film systems.
		1	Does not meet any of the above levels.
KR1	Explanation and justification of creative intent	5	Clearly states the creative intent and supports it with appropriate and well-justified reasons or evidence.
		4	States the creative intent and supports it with appropriate and well-justified reasons or evidence.
		3	States the creative intent and provides reasons or evidence,but some parts are not well-justified.
		2	States the creative intent of their work but provides little to no supporting reason or evidence.
		1	Does not meet any of the above levels.
KR2	Development and coherence of argumentative essay	5	Presents a clear stance and develops an argumentative essay with logical coherence and consistency throughout.
		4	Presents a clear stance and develops an essay, but the logical flow or consistency of the argument is somewhat lacking.
		3	Presents a stance and attempts an argumentative essay, but poor logic and coherence make the argument difficult to understand.
		2	Attempts an argumentative essay from a chosen stance but fails to complete it.
		1	Does not meet any of the above levels.

Appendix C.

Code	Students' response	Score	Feedback
SS1	Since teenagers are people with immature personalities who haven't become adults yet and need adult protection, I couldn't support teenage part-time jobs. But now my thinking has changed. The government used various laws like the Labor Standards Act, Youth Protection Act, and Industrial Accident Protection Act to make laws that protect teenagers in all areas including working hours, minimum working age, and how to deal with unfair labor and unfair dismissal. Because of this, teenagers can work fairly by writing work contracts and can solve problems even if they get hurt at work, which means they can work in a better environment. In this way, the government's policy worked well. But still quite many teenagers aren't getting protection from these laws. The sad reality is that not only the above case but over 50% of Incheon teenagers experienced violations of teenage labor rights, and work contracts often aren't signed either. This means we can't ignore that there are laws for teenage labor but they're not being followed. For this, the government should regularly and carefully investigate businesses that use teenage workers, and we should also consider having teenagers complete one hour of related education per month so they can know the Labor Standards Act better and make fair demands. If we pay more attention and monitor more thoroughly and also provide education, this law will shine even brighter.	5	Clearly presents both policy achievements and limitations while providing concrete improvement suggestions like regular inspections and monthly education programs.
	I think the government's policies about teenage part-time jobs have both good points and lacking points. The good point is that they try to protect us with laws like the Labor Standards Act and Youth Protection Act. For example, they decided we can't work night shifts and employers must follow minimum wage rules. Actually, when my older brother worked at a convenience store in high school and the boss treated him unfairly, he reported it to the Ministry of Employment and Labor and solved it. But the problem is that many places still don't follow these laws. Especially small stores and restaurants often don't write work contracts and make people work too many hours. So I think the government should check these places more often, and they should teach us teenagers more about our rights.	4	Evaluates government policies using specific laws like Labor Standards Act and Youth Protection Act with concrete examples, but some arguments lack depth. Shows logical flow but has limited critical perspective regarding comprehensive policy analysis.
	I was also shocked when I saw this news article. Because I thought school is a place for learning and teachers would naturally respect students' human rights, but I was really surprised that they do sexual harassment beyond swearing and insulting. But I think the victim student was really cool and handled it well by being brave in that scary and frightening situation to tell the world about the human rights violation. After telling the world, the school admitted how serious the problem was and said they would prepare education and monitoring systems to prevent it from happening again, so I think that's really good. After this situation, I think not just teachers but everyone should respect everyone's human rights and not violate them.	3	Mentions the school's preventive education and monitoring system as government measures, but lacks sufficient evidence or deeper analysis. The evaluation of policy validity is superficial and needs more discussion of limitations and alternatives.
	I also think Coupang is responsible. Because Coupang said "there's no direct connection between the job and death" but then admitted it and apologized when the media got stronger, I think that's a problem. Also, I thought they were too irresponsible thinking this problem would never happen. I thought it was bad from the beginning that this incident happened when it shouldn't have. Also, I really hated that Coupang tried to hide this incident. So I think the government should focus more and pay more attention to this problem from now on. And I hope they pay attention once a year to see if there are any problems like this.	2	Mentions the need for more government attention and annual monitoring, but fails to clearly present specific government policies. The analysis lacks concrete evidence and logical development regarding policy effectiveness.
	When unfair things like this happen, the court should manage so that part-time workers get fair wages and proper weekly holiday pay. If not, unfair things like this can happen again.	1	No specific government policy is presented and the content is just a simple statement. Need to identify concrete government measures and provide analysis of their effectiveness with supporting evidence.

Appendix C. (continued)

Code	Students' response	Score	Feedback
SS2	Dokdo - Dokdo is two volcanic islands created by underwater volcanic activity about 4.6 million years ago. Lava kept erupting from the sea floor and piled up to form today's Dongdo and Seodo, and 89 small rock islands were also created around them. Geographically, it is located in the East Sea, 87km southeast of Ulleungdo. The average temperature in winter is 1 degree and in summer is about 23 degrees. For landscape features, Seodo is 168m above sea level and has high cliffs, while Dongdo is relatively flat so a lighthouse is installed there. The cliffs created by volcanic activity show unique coastal scenery. Because of this volcanic terrain and oceanic climate, seabirds like black-tailed gulls and storm petrels live there, and various fish like pollack and squid live in the surrounding sea. In the past, many sea lions lived there too, but they can't be seen now because of overhunting during Japanese colonial period. Like this, Dokdo is our precious land where volcanic terrain and oceanic climate combine to create a unique ecosystem.	5	Accurately analyzes volcanic landform formation and systematically explains causal relationships between landscape, climate, and ecosystem with specific evidence.
	Rio de Janeiro - Rio de Janeiro has lakes formed because of sand dunes created by sand carried by waves. This place is in a tropical climate region, but it's relatively cool even in summer. White sandy beaches and mountains covered with thick tropical rainforests come together to create world-famous beautiful scenery. Because of these geographical features, it became a popular destination that many tourists visit.	4	Explains sand dune formation and connects most geographical features, but lacks specific explanation for why it's cool despite tropical climate. Need clearer details about trade winds or ocean influences on climate.
	Yonggul - Gyeongju Yonggul is a sea cave made by waves eroding rocks. So it has passages that look like a dragon went in, which is its special feature. In this area, the highest temperature in winter is less than 18 degrees and in summer it goes above 30 degrees. The cave creates beautiful coastal scenery in Gyeongju.	3	Mentions wave erosion forming caves but limited connections between landform formation, climate, and landscape. Need more specific analysis of how erosion and climate interact to create the current scenery.
	Dokdo - Dokdo is two islands. Seodo has steep slopes and Dongdo is flat, so the Dokdo guards are there. Dokdo has strong waves and lots of wind.	2	Missing landform formation processes and only lists individual elements separately. Need to explain how volcanic activity created Dokdo's terrain and connect this to climate and landscape features.
	Himalayas - It is the highest place on Earth. Mount Everest is there.	1	No analysis of landform formation processes. Need to explain how the Himalayas were formed and connect this to climate and geographical features.

Appendix C. (continued)

Code	Students' response	Score	Feedback
TH1	Soil plays a very important role on Earth, but we don't know much about it. Through the Soil Photo Challenge, people learn how precious soil is. When people take pictures of soil and post them, they learn that soil helps plants grow, cleans rainwater, and becomes home to many living things. When people learn these things, they try to buy eco-friendly products or use less chemical fertilizer. When other people see the photos and posts that individuals share, they also become interested and try to protect soil. This way, small actions by individuals spread to the whole society, and policies and activities to protect soil increase. In the end, this leads to a better Earth environment.	5	Clearly explains the causal relationship between the soil photo challenge and Earth system impacts, showing how individual awareness leads to behavioral changes and broader environmental protection.
	Our Challenge has a good effect on the environment. When we spend less money, we buy fewer things. When we buy fewer things, factories make fewer products too. When factories make fewer products, they use less resources and energy, and less trash is made. Even though one person saving money might seem small, if many people do it together, it can really help protect the environment.	4	Describes how saving money impacts the Earth system through reduced consumption and production, but the causal relationships could be explained more clearly. Consider strengthening the explanation of how individual actions connect to larger environmental systems.
	The 'Go-min-hae' Challenge is an activity that protects the environment by reducing unnecessary shopping. People post items they want to buy in group chats, and others give opinions about whether they really need it. This way, we can reduce impulse buying and make less trash, which helps the environment.	3	Describes the challenge activity and mentions environmental impacts like reducing trash, but lacks explanation of causal connections. Need to explain how reduced consumption leads to specific environmental improvements.
	Our Challenge is to reduce carbon dioxide. It can support a sustainable Earth system.	2	Necessary to clearly explain how this challenge contributes to reducing carbon dioxide emissions. Additionally, it should be explained how reducing CO_2 emissions helps support a sustainable Earth system.
	Good influence	1	Need to describe what the challenge is and explain its environmental effects.

Appendix C. (continued)

Code	Students' response	Score	Feedback
TH2	Real life and the movie show different systems from many perspectives like economics, social structure, jobs, and legal systems. Economically, money is used for exchange in real life, but time acts as currency in the movie. In terms of social structure, rich and poor people in real life are divided by wealth, but in the movie, social classes are divided by how much time people have. For jobs, the movie has time-related jobs like Timekeepers and Minutemen that don't exist in real life. In the legal system, robbers who steal money and police who stop them in real life correspond to Minutemen who steal time and Timekeepers who catch them in the movie. The value system in the movie's society is not based on effort or popularity like in real life, but on how much something is related to time. Since time is directly connected to life, everything's value is judged based on time, creating a unique value system.	5	Comprehensively addresses differences from multiple diverse perspectives and clearly explains the value criteria within the film's social system
	Real life and the movie show big differences in economic systems and job structures. Economically, money is used for exchange in real life, but time acts as currency in the movie. In terms of jobs, real life doesn't have special time-related jobs, but the movie has jobs like Timekeepers who manage and maintain time. Also, the movie's society has a class structure where people with lots of time live rich and comfortable lives, while people without enough time must live poor and rushed lives.	4	Successfully writes about differences from multiple perspectives including economic systems and occupational structures. Shows clear understanding of how the two worlds differ in various aspects, though could benefit from discussing the film's value criteria more explicitly.
	In real life, we use money to buy things, but in the movie, time is used as money. In real life, people work and get paid with money, but in the movie, people work and get paid with time. This shows how the economic system is completely different in terms of what we use as currency.	3	Clearly writes about the differences from one perspective - the economic system where money vs. time serves as currency. Well-focused comparison of the economic differences between real life and the movie.
	Things in real life like daily life and living seem like.....they would become more comfortable than before, but in the movie, there doesn't seem to be much change.	2	Mentions some changes but fails to clearly explain the differences between real-world and film systems. Need to specifically compare how the systems work differently in each setting.
	I don't know.	1	No comparison between real-world and film systems is provided. Need to identify and explain at least one difference between the two systems.

Appendix C. (continued)

Code	Students' response	Score	Feedback
KR1	If I had to pick the three most painful memories, I would definitely choose this as number one. It was the most emotionally painful memory and also an opportunity for me to reflect on myself. I used metaphors like: the teacher as a witch, the teacher's words as poison, my feelings when being scolded as a prisoner on death row, the sky I looked at during hard times, the emerald-colored Jeju sea I first saw during happy times as a beautiful healing sea, the opportunity for reflection as a key, my past reflection as going into the ocean, and praise as sweet rain falling on dry land. My creative intention was to let many people know that many things changed after I looked up at the sky and emptied my mind. As described in the writing, in the past I was lazy, didn't do my best, and postponed many things. Of course, I didn't notice this about myself, so whenever I got in trouble for vocabulary tests, I only blamed the teacher. It's really pathetic when I think about it. Actually, people around me had told me not to postpone things so much. The price for ignoring their advice was the teacher saying "If you can't even do this much, you should leave the academy." This is the part about the witch and the witch's poison in my writing. After hearing this three days a week, my personality became very timid and I walked looking down at the ground a lot. Then one day I looked up at the sky, and my mind felt refreshed and my head became clear, so I reflected on my past. The change from this reflection was huge. I received a lot of praise from people around me and became diligent. Like the title of this piece, I looked at the sky often and reflected on myself every day. As I reflected on myself, I gradually began to notice others in my memories too, and I realized there are quite a few people like me - people who can't look back at themselves and just blame others. I wrote this piece hoping that people who still don't reflect on themselves will read this and start reflecting on themselves.	5	Clearly presents the creative intent about self-reflection and fully supports it with detailed personal experience and well-justified metaphorical expressions.
	I chose this event because among the things I experienced recently, I think it was something that made me reflect on myself, grow, and change my behavior. The metaphorical expression I used was in the part where I wrote "Why didn't I know that life is like snow that suddenly falls and melts in an instant, and that I should enjoy this moment?" I compared snow to life because I thought snow and life have many things in common. Snow is limited and falls suddenly but melts quickly, and people enjoy those moments. This is similar to how new life is born and dies as time passes. Also, snow is usually white, but it quickly becomes dirty and turns black. This is also similar to life, which changes color as it faces different situations. That's why I compared snow and life. My creative intention in writing this was to show the reflection and growth I gained through the upcoming farewell with my grandfather. Through this event, I realized that I should be kind not only to my grandfather but also to people around me, and I learned how precious life is. Because it was an event that gave me so many realizations, I decided to write this piece.	4	Clearly states the creative intent and provides appropriate reasoning through the snow-life metaphor, though some connections could be stronger. Consider strengthening the link between your metaphors and your overall creative purpose.
	I chose studying because it's something all of us students can relate to, and studying is difficult right now. The metaphorical expressions are "like mountain ranges" and "like a mountain range" - that's a simile. The symbol is mountain ranges. My creative intention was to create empathy, because when people can relate to something, they can read it better.	3	Presents the intent to create empathy but lacks adequate explanation of how the mountain range metaphor achieves this goal. Need to explain more specifically why this metaphor effectively creates relatability among students.

Appendix C. (continued)

Code	Students' response	Score	Feedback
KR1	I used to have many conflicts with my friends and it was really hard. But I looked back at myself and reflected on my actions, and my relationships with my friends got better. That's why I wrote this poem - to tell people that if you want to stop conflicts, you need to reflect on yourself first. I described my past self, who couldn't reflect properly, as a baby. I also showed specifically what I realized.	2	There is no clear statement of the creative intention or the reasons behind it, so the response could be improved by clearly stating the intention at the beginning and supporting it with a specific reason.
	I chose the experience of enlightenment because this experience changed me, even if just a little. The metaphor is about awakening myself.	1	No creative intent is stated. Only describes the choice of subject and literary devices without explaining the purpose of the writing. Need to clearly state what message or meaning you want to convey through this piece.
KR2	Internet is useful in modern life. We can get information and talk with people by using websites, blogs, and cafes. Internet info is fast. If we go to a website and search what we want to know, it comes out right away. Internet also tells us new things that happened. If we search news in websites, we can see famous news nowadays. There is a lot of information. Many articles and ads go up every day. If we use internet info well, we can get useful and fast information, and our life can be better.	5	Clearly presents a stance and develops the argument logically throughout.
	In modern society, the internet is like an ocean of information. Using information is very useful in many ways. First, the internet is a convenient tool. If there was no internet, we would have to go to libraries or research centers to get information. Second, the internet has a lot of information. Books have a limited amount of information, but the internet has tons of information. Let's use information efficiently and beneficially to improve our quality of life.	4	Presents a clear stance, but the logical flow between arguments is somewhat uneven. Consider improving the connections between paragraphs to enhance coherence.
	Internet is needed in today's society. Because it already became our tool, and people say it's dangerous but they still use it. They even made block functions for the dangers, so I don't get why they still talk about bad things now.	3	Attempts an argumentative essay, but the reasoning is difficult to follow due to poor logic and weak coherence. Focus on clarifying how each point supports the main stance.
	Internet is useful. I like to use internet for many things. It helps me find information. But sometimes it's bad too. Internet has good things and bad things. I think we should use internet.	2	Shows an attempt at an argumentative essay but does not fully develop the argument. Need to provide clear supporting reasons and maintain consistency throughout.
	I think we should use the internet a little more carefully.	1	Does not present a clear stance or coherent argument. Need to state a position and organize supporting points logically.