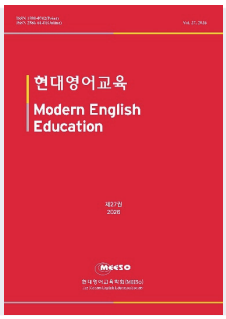
Modern English Education, Vol. 27, (2026) pp. 233-252
doi: https://doi.org/10.18095/meeso.2026.27.1.233

Modern English Education

ISSN: 1598-0782 / e-ISSN: 2586-6141
http://journal.meeso.or.kr



주의집중 기반 언어모델의 구조적 차이에 따른 자동화 영어 에세이 다면평가 양상의 비교 분석

Seongyeub Chu (Korea Advanced Institute of Science and Technology)

Received: 9 March 2026
Revised: 25 March 2026
Accepted: 17 April 2026

Chu, Seongyeub. (2026). A comparative study of automated multi-trait English essay scoring across attention-based language model architectures. *Modern English Education*, *27*, 233-252.

Keywords
Automated essay scoring, language model, attention, multi-trait essay scoring, artificial intelligence
자동화 에세이 평가, 언어모델, 주의집중, 에세이 다면평가, 인공지능

Seongyeub Chu
PhD Candidate
Graduate School of Data Science
Korea Advanced Institute of Science and Technology
chseye7@gmail.com

Abstract
Recent advances in attention-based language models have increased interest in automated essay scoring (AES) for English essays, particularly analytic scoring. However, prior work has primarily focused on improving score agreement with human raters, with limited attention to how different model architectures realize analytic assessment from an educational perspective. Accordingly, this study examines how architectural differences in attention-based language models influence analytic English essay scoring by comparing encoder-based (BERT), decoder-based (Qwen), and encoder–decoder-based (T5) models on the PERSUADE 1.0 dataset, consisting of essays written by U.S. students in grades 8–12 and rated by two trained raters, with final scores determined through adjudication by a third rater. The analysis evaluates agreement with human raters using Quadratic Weighted Kappa (QWK), attention patterns by classifying attention-weighted tokens into content and function words via part-of-speech tagging, and scoring efficiency by measuring average inference time per essay under both CPU and GPU environments. The results show that the encoder–decoder-based (T5) model achieves the highest agreement with human raters by focusing more on content words, reflecting a meaning-focused assessment strategy. In contrast, other models show lower agreement with less emphasis on content words. Despite moderate computational costs, the encoder–decoder-based model remains feasible for educational use. These findings highlight the importance of model architecture in analytic AES and offer guidance for selecting practical scoring systems.

서론

영어 에세이 평가는 학습자의 사고력, 영어 의사소통 능력, 담화 구성 능력을 종합적으로 반영할 수 있다는 점에서 영어 교육에서 중요한 평가 방식으로 자리 잡아왔다(Hamp-Lyons, 2003; Ibnian, 2011). 특히 분석적 에세이 평가(analytic scoring)는 하나의 점수로 수행을 판단하는 총체적 평가(holistic scoring)와 달리, 내용(content), 문법(grammar), 어휘(vocabulary), 응집성(cohesion) 등 복수의 평가 구인(construct)을 기준으로 학습자의 쓰기

능력을 세분화하여 진단할 수 있어 교육적 활용 가치가 크다. 이러한 분석적 평가는 학습자의 강점과 약점을 보다 구체적으로 파악할 수 있으며, 교수·학습 및 피드백 제공과의 직접적인 연계를 가능하게 한다는 점에서 그 중요성이 지속적으로 강조되어 왔다(Harsch & Martin, 2013; Winke & Lim, 2015). 그러나, 이러한 교육적 이점에도 불구하고, 하나의 에세이로부터 다양한 요소를 평가기준과 연계하여 정확하게 평가하는 것은 다수의 학생이 작성한 에세이를 평가해야 하는 교사에게 매우 큰 부담이며, 이는 교사의 피로도 누적과 함께 평가의 질을 저하시킬 수 있다. 이에 따라 교사의 부담을 경감하고 평가의 일관성을 확보하기 위한 방안으로 AI 기반 자동화 에세이 평가(automated essay scoring, AES) 기술 개발이 오랫동안 시도되어 왔다(Misgna et al., 2024). 국내에서도 서울시를 비롯하여 많은 시도교육청에서 교사를 지원하기 위한 AI 기반 서 논술 자동화 기술을 적극적으로 교육 현장에 도입할 방안을 모색하고, 관련 교사 연수를 실시하고 있는 점에서, 영어 글쓰기 평가 역시 가까운 시일 내에 AI 기술을 활용한 자동화 평가 체계가 본격적으로 도입될 가능성이 크다. 이러한 흐름에 발맞추어 국내 교육 현장의 맥락과 교육학적 관점에서 그동안 AI 분야에서 개발되어 온 자동화 에세이 평가 기술을 체계적으로 분석하고, 가용한 자원과 실제 교육 환경에 적합한 모델이 무엇인지 판단할 수 있는 근거 자료를 제공할 필요가 있다.

기존 자동화 에세이 평가 연구는 주로 모델의 평가 정확도를 향상시키는 데 초점을 두어 왔다. 특히 Vaswani 외 7인(2017)에 의해 ChatGPT의 근간이 된 트랜스포머(transformer) 모델 아키텍처가 제안된 이후, 사람과 유사하게 텍스트의 중요한 부분에 선택적으로 주의를 기울이는 주의집중 기법(attention mechanism)을 활용한 다양한 언어모델이 자동화 에세이 평가에 적용되고 있다(Do et al., 2024a; Kumar et al., 2022; Ramesh & Sanampudi, 2022). 이러한 언어모델은 대규모 텍스트 데이터로 학습된 사전학습 언어모델(pretrained language model, PLM)로, OpenAI의 ChatGPT, Google의 Gemini, Anthropic의 Claude 등 현재 상용화된 거대 언어모델(large language models, LLMs) 역시 PLM의 범주에 속한다.

이들 언어모델은 크게 텍스트 이해에 중점을 둔 인코더 기반 언어모델(encoder-based language model), 텍스트 생성을 중심으로 설계된 디코더 기반 언어모델(decoder-based language model), 그리고 텍스트 이해와 생성을 균형 있게 결합한 인코더–디코더 기반 언어모델(encoder–decoder-based language model)로 구분될 수 있으며, 자동화 에세이 평가 연구에서는 이러한 모델들이 다양한 형태로 변형 및 활용되어 왔다(Dong & Zhang, 2016; Lee et al., 2024; Ramesh & Sanampudi, 2022). 그러나 기존 연구들은 대체로 대규모 연산 자원을 활용하여 '사람 채점자의 점수와 얼마나 일치하게 에세이 점수를 예측하는가'에 초점을 둘 뿐, 각 유형의 언어모델이 점수 산출 과정에서 어떠한 내부적인 차이가 있는지 교육학적 설명을 충분히 제공하지 못한다는 한계를 지닌다.

또한, 사람 채점자와의 점수 일치도 향상에 과도하게 집중한 나머지, 언어모델을 인코더 기반, 디코더 기반, 인코더–디코더 기반이라는 구조적 특성을 고려하여 선택 및 설계하기보다는, 기존 모델에 다수의 새로운 모듈을 추가하는 방식이 주로 활용되어 왔다. 이로 인해 모델의 규모가 지나치게 커지고, 다량의 GPU와 같은 고가의 연산 자원을 요구하는 경우가 많아, 국내 교육 기관의 현실적 여건에서는 즉각적인 구축과 운영이 어려운 제약으로 작용한다. 그 결과, 국내 교육 현장에서 활용 가능한 언어모델이 무엇인지 판단할 수 있는 명확한 근거가 부족한 상황에서, 다양한 모델을 시행착오적으로 적용한 뒤 맥락에 맞는 모델을 선택할 수밖에 없으며, 이는 시간적·물적 측면에서 비효율성을 초래한다.

마지막으로, 분석적 에세이 평가 맥락에서 사람 평가자는 에세이의 전반적인 의미를 먼저 파악한 후, 각 평가 구인과 관련된 언어적 단서를 선택적으로 참조하며 판단을 내리는 복합적인 인지 과정을 거친다는 점이 선행연구를 통해 보고되어 왔다(Cumming et al., 2002; Lumley, 2005; Wolfe, 1997). 뿐만 아니라, 에세이를 평가할 때 내용어(content word)와 기능어(function word)를 중심으로 의미적인 요소(meaning-focused assessment)와 형식적인 요소(form-focused assessment)를 차등적으로 고려해서 평가한다(Winke & Lim, 2015; Wolfe, 1997). 그럼에도 불구하고, 기존 자동화 에세이 평가 연구에서는 서로 다른 구조를 지닌 언어모델이 이러한 평가 과정을 얼마나 유사하게 모사하는지, 그리고 모델의 주의집중 방식이 평가 양상에 어떠한 영향을 미치는지에 대한 체계적인 비교·분석이 충분히 이루어지지 않았다.

이에 본 연구는 주의집중 기반 언어모델의 구조적 차이에 주목하여, 인코더 기반, 디코더 기반, 인코더–디코더 기반 언어모델이 영어 에세이 다면평가에서 보이는 평가 양상을 (1) 사람 채점자와의 점수 일치도, (2) 에세이 내 내용어와 기능어에 대한 언어모델별 주의집중 양상, (3) 실제 교육 현장을 고려한 채점 속도라는 세 가지 측면에서 언어모델의 평가 특성을 종합적으로 검토하여 어떠한 시사점을 제공하는지 밝히는 것을 목표로 한다. 이를 통해 자동화 에세이 평가 모델이 단순한 '점수 예측기'를 넘어, 사람 채점자의 분석적 평가 과정을 얼마나 충실하게 반영하고 있는지 분석하고 교육 현장 수용 가능성을 논의함으로써, 향후 부분적으로나마 다양한 교육기관에서

현장의 상황과 자원을 고려하여 적절하고 현실적인 AI 언어모델을 활용한 평가 도구 설계와 적용에 이론적·실천적 근거를 제공하고자 한다. 이를 위해 본 연구는 다음과 같은 연구문제를 설정하였다.

연구문제 1) 분석적 영어 에세이 평가에서 인코더 기반, 디코더 기반, 인코더–디코더 기반 언어모델은 복수의 평가 구인별로 사람 평가자와 어느 정도의 채점 일치도를 보이는가?

연구문제 2) 분석적 영어 에세이 평가에서 인코더 기반, 디코더 기반, 인코더–디코더 기반 언어모델은 에세이의 내용어와 기능어를 참조하는 양상에서 어떠한 차이를 보이는가?

연구문제 3) 실제 교육 현장 및 대규모 평가 맥락을 고려할 때, 인코더 기반, 디코더 기반, 인코더–디코더 기반 언어모델은 에세이당 채점 속도에서 어떠한 차이를 보이며, 이는 영어 에세이 평가에서의 실제 활용 가능성에 어떠한 시사점을 제공하는가?

이론적 배경

자동화 에세이 평가(Automated Essay Scoring, AES)

자동화 에세이 평가는 오랜 기간 연구되어 온 분야로, 초기 연구들은 주로 에세이의 전반적인 완성도를 하나의 점수로 산출하는 총체적 평가 방식에 기반하였다. Page (1966)가 제안한 초기 자동 채점 시스템 이후, 2000년대에는 다양한 상용 및 연구용 자동 채점 시스템이 등장하였으며(Shermis & Burstein, 2013), 이들 시스템은 규칙 기반 접근이나 Random Forest와 같은 의사결정 트리(decision tree) 기반의 전통적인 머신러닝 기법을 중심으로 설계되었다(Ramesh & Sanampudi, 2022). 이후 2012년 Kaggle에서 공개된 Automatic Student Assessment Prize (ASAP) 에세이 데이터는 딥러닝 기반 접근을 에세이 평가 연구에 본격적으로 도입하는 계기가 되었고, AES 연구의 방법론적 전환을 촉진하였다(Hamner et al., 2012).

대표적으로, Taghipour와 Ng (2016)은 순환신경망(recurrent neural network, RNN)을 활용하여 에세이 평가 모델을 개발하였고, Dong과 Zhang (2016)은 합성곱신경망(convolutional neural network, CNN)을 기반으로 에세이 평가 모델을 개발하여 사람 채점자와 높은 수준의 일치도를 달성하였으며, 이후 두 신경망을 결합한 하이브리드 구조로까지 확장되었다(Zhang & Litman, 2020). 이후에는 데이터 수집 규모의 확대와 딥러닝 기반의 주의집중 기법의 발전에 따라(Vaswani et al., 2017), 대규모 말뭉치로 학습되어 다양한 과업에서 높은 수행 능력을 보이는 사전학습 언어모델을 활용함으로써, 실제 사람이 에세이를 이해하는 방식을 모사하도록 모델을 설계하고 사람 평가자와 유사한 수준의 평가 성능을 보이는 모델들이 점진적으로 개발되어 왔다(Misgna et al., 2024; Wang et al., 2022; Yang et al., 2020).

초기 딥러닝 기반 자동화 에세이 평가 연구가 집중했던 총체적 평가 방식은 에세이의 전반적인 완성도를 효율적으로 평가할 수 있다는 장점을 제공하지만, 단일 점수로 평가 결과를 제시함으로써 쓰기 능력을 구성하는 다양한 하위 요소를 세밀하게 반영하지 못한다는 한계가 있다. 이러한 문제의식 속에서 최근 연구들은 분석적 평가에 기반한 다면평가(multi-trait scoring) 관점에서 에세이를 분석하려는 시도를 이어왔다. 초기 연구에서는 복수의 평가 구인별 예측 모델을 설계하여 여러 구인을 동시에 예측하는 방식이 제안되었고(Mathias & Bhattacharyya, 2020; Taghipour, 2017), 이후 하나의 모델이 여러 평가 구인의 특성을 공동으로 학습하는 멀티 태스크 학습(multi-task learning) 프레임워크로 발전하였다(Kumar et al., 2022). 이러한 접근은 평가 구인 간 정보를 공유함으로써 학습 효율성을 높이는 동시에, 다면평가를 보다 안정적으로 수행할 수 있음을 보여주었다.

최근 LLM의 급속한 발전에 따라, 다면적 에세이 평가에도 LLM을 직접 활용하려는 연구가 등장하고 있다. Lee 외 4인(2024)은 LLM이 평가 구인별 루브릭(rubric)을 자동으로 생성한 후 이를 바탕으로 점수를 산출하는 방식을 제안하였으며, Stahl 외 3인(2024)은 사전에 정의된 루브릭을 활용하여 점수 예측과 함께 학습자 피드백을 생성하는 접근을 시도하였다. 그럼에도 불구하고, 이러한 방법들은 모델이 산출한 점수가 사람 채점자의 평가 결과와 충분한 수준의 일치도를 보이지 못한다는 한계를 지니고 있다. 이에 따라 자동화 에세이 다면평가 분야에서는 여전히 언어모델을 대규모 데이터로 학습시키는 방식이 가장 효과적인 접근이라는 점에 대해 다수의 연구가 공감대를

형성하고 있다(Chu et al., 2025; Misgna et al., 2024).

이러한 딥러닝 기반 혹은 LLM 기반 에세이 평가 연구 흐름 속에서 자동화 평가 모델이 교육 현장에 수용되기 위해서는 에세이를 채점하는 과정에 대한 설명력이 있어야 한다는 문제의식이 공유되고 있고(Misgna et al., 2024), 에세이 평가의 근거를 함께 생성 및 분석하고자 하는 시도가 최근 이루어지고 있다(Chu et al., 2025; Do et al., 2026; Tang, 2026). 이처럼 설명 가능 에세이 평가는 근거를 포함한 평가(assessment with rationale)와 근거를 활용한 평가(assessment using rationale)로 크게 구분된다. 먼저, 근거를 포함한 평가에서는 Do 외 2인(2026)에서 제안된 바와 같이 에세이가 평가 구인별로 사람 채점자에게 채점된 점수와 함께 LLM에게 입력되어 구인별 점수에 대한 텍스트 형태의 근거를 생성한 후 이것을 사람 채점 점수와 함께 모델 학습 레이블(label)로 활용하여 작은 규모의 T5 언어모델(Raffel et al., 2020)을 학습하여 에세이가 모델에 입력되었을 때 점수와 함께 근거가 평가 결과에 포함되어 산출되도록 하였다. 다음으로, 근거를 활용한 평가와 관련해서 Chu 외 3인(2025)은 에세이와 구인별 평가 루브릭을 구인별로 작동하는 GPT-3.5 (OpenAI, 2022) 기반 에이전트(agent)에 입력하여 에세이가 어느 정도의 수준을 보이는지 각 구인별로 텍스트 형태의 근거를 생성한 후 이것을 에세이와 함께 T5 언어모델에 입력하는 방법을 제안하여 GPT-3.5가 생성한 평가 근거가 채점에 활용되는 구조를 제안하였다. 그리고, Tang (2026)은 BERT 언어모델(Devlin et al., 2019) 계열인 RoBERTa 언어모델(Liu et al., 2019)과 평가 구인별 주의집중 모듈을 결합하여 평가 구인별 모듈이 에세이의 각 단어에 주의집중한 정보를 구인별 채점 점수에 대한 근거로 활용할 수 있는 가능성을 제시하였다. 그러나, Do 외 2인(2026)과 Chu 외 3인(2025)의 방법은 구인별 점수의 근거를 LLM이 생성하도록 하고 이것을 다른 언어모델(예: T5 언어모델)이 학습 정보 혹은 추가 입력 정보로 활용하여 평가하도록 함으로써 실제로 평가를 수행하는 언어모델의 작동 과정은 여전히 블랙박스로 남게 된다. 또한, Tang (2026)이 제안한 방법은 에세이 채점 결과의 근거가 구인별 모듈의 주의집중에 의존하여 실제로 에세이의 내용을 파악하는 RoBERTa 언어모델과 구인별 주의집중 모듈 중 어떤 부분에 의해 채점 결과가 나왔는지 해석하기 어려운 문제가 존재한다. 추가적으로, 기존의 연구들은 각 모델들의 채점 과정을 단순히 에세이 내 단어 수준에서만 분석할 뿐, 실제로 사람 채점자가 평가하는 과정과 비교하여 영어평가적 관점에서의 고찰이 수행되지 않았다.

이처럼, 데이터 기반 지도학습(supervised-fine-tuning)을 통해 사람 채점자와 더욱 유사한 수준의 평가 수행 능력을 확보하거나 설명 가능한 에세이 평가를 위해 언어모델의 구조를 새롭게 설계하거나 학습 전략을 개선하려는 연구는 지속적으로 증가하고 있다 (Chu et al., 2025; Do et al., 2024b; Li & Ng, 2025, Tang, 2026). 그러나 이러한 연구들은 대체로 언어모델을 기저모델(backbone model)로 활용하여, 모델이 에세이를 이해한 결과로부터 후속 작업을 통해 추가적인 정보를 추출하는 방식에 초점을 두고 있으며, 데이터 마이닝(data mining) 관점에서의 접근에 머무르고 있다. 이에 따라 서로 다른 유형의 기저 언어모델이 지니는 주의집중 기법의 구조적 차이가 에세이 평가 과정에 어떠한 영향을 미치는지, 그리고 이러한 차이가 사람 채점자의 평가 양상과 어떻게 연결되는지에 대한 분석은 충분히 이루어지지 않았다. 실제 교육 현장에서 자동화 에세이 평가 모델이 활용되고 영어교육 전문가들로부터 수용되기 위해서는, 단순한 평가 결과 중심의 연구를 넘어 기저 언어모델 간의 본질적인 차이와 그 차이가 평가 과정 전반에 미치는 영향을 체계적으로 분석할 필요가 있다. 이에 본 연구는 이러한 문제의식을 바탕으로, 주의집중 기법을 활용하는 세 가지 유형의 언어모델(인코더 기반, 디코더 기반, 인코더–디코더 기반)이 영어 에세이 다면평가에서 보이는 평가 양상을 분석한다. 구체적으로, 사람 채점자와의 점수 일치도를 정량적으로 비교하는 한편, 모델이 에세이 내 단어에 주의집중하는 양상을 사람 채점자의 평가 과정과 비교·분석하고, 더 나아가 모델별 채점 속도를 측정함으로써 실제 교육 현장에서의 활용 가능성을 종합적으로 검토하고자 한다.

주의집중 기반 언어모델(Attention-based Language Model)

트랜스포머는 Vaswani 외 7인(2017)이 제안한 신경망 아키텍처로, 주어진 데이터의 모든 부분을 비슷하게 참고하던 기존 RNN이나 CNN과 달리 수행 과업의 성격에 따라 데이터에서 관련된 부분만 선택적으로 참고하는 주의집중 기법(Bahdanau et al., 2014)을 기반으로 문맥 정보를 모델링한다는 점에서 언어모델의 중요한 전환점을 마련하였다. 트랜스포머의 핵심 구성 요소인 자기 주의집중(self-attention)은 입력 텍스트를 토큰(token) 단위로 분해한 뒤, 각 토큰이 텍스트 내의 다른 모든 토큰과의 관계를 직접적으로 참조하도록 설계된 메커니즘이다. 여기서 토큰이란 문장을 구성하는 최소 의미 단위로, 하나의 단어가 될 수도 있고 단어의 일부(subword) 단위가 될 수도 있다. 이러한 자기 주의집중 구조를 통해 트랜스포머는 사람이 글을 읽을 때 문장이나 문단 내에서 위치상 멀리 떨어져 있으나

서로의 해석에 영향을 주는 단어들을 연결하여 이해하는 특성, 즉 장거리 의존성(long-range dependency)을 효율적으로 포착할 수 있다. 이 과정에서 모델은 각 토큰에 서로 다른 주의집중 가중치(attention weight)를 할당함으로써, 텍스트 내에서 의미적으로 중요한 정보와 상대적으로 덜 중요한 정보를 구분하여 처리한다. 이러한 주의집중 기법은 단순한 성능 향상을 넘어, 모델이 텍스트를 어떻게 해석하고 어떤 언어적 단서에 주목하는지를 분석할 수 있는 가능성을 제공한다. 특히 에세이 평가와 같이 복합적인 의미 해석과 판단이 요구되는 과업에서는, 주의집중 양상이 모델의 평가 전략을 간접적으로 드러내는 중요한 단서로 작용할 수 있다(Clark et al., 2019; Yang et al., 2020; Zhang & Litman, 2020).

트랜스포머의 등장 이후, 해당 구조를 기반으로 다양한 유형의 언어모델이 개발되었으며, 이들은 크게 인코더 기반, 디코더 기반, 인코더–디코더 기반 모델로 구분된다. 먼저, 인코더 기반 언어모델은 입력 텍스트를 양방향으로 인코딩하여 문맥적 표현을 학습하는 데 초점을 둔다. Devlin 외 3인(2019)에 의해 구축된 BERT는 이러한 인코더 기반 모델의 대표적인 예로, 문장 이해, 의미 비교, 분류 과업 등에서 뛰어난 성능을 보여 왔다. 인코더 기반 모델은 입력 텍스트 전체를 동시에 참조하는 자기 주의집중 구조를 통해 텍스트의 전반적인 의미를 효과적으로 파악하지만, 텍스트 생성보다는 텍스트 이해 중심의 과업에 최적화되어 있다. 반면, 디코더 기반 언어모델은 이전 토큰을 조건으로 다음 토큰을 순차적으로 생성하는 자기회귀적(autoregressive) 구조를 가지며, 주로 텍스트 생성 과업에 활용된다. 구체적으로, ‘What’s your name?’이라는 문장을 입력받아 ‘My name’까지 디코더 기반 모델이 생성하였다면 그 다음으로 올 수 있는 토큰인 ‘is’를 생성하기 위해 입력된 문장과 이전까지 생성된 토큰을 연결한 ‘What’s your name? My name’ 부분을 모두 참고하며 이 과정에서 중요한 토큰을 더 많이 참고하는 주의집중 기법이 활용된다. 이처럼 모델이 생성하고자 하는 텍스트의 모든 토큰에 대해 이러한 순차적인 방식을 거쳐야 하기 때문에 디코더를 활용하는 언어모델의 처리 시간은 인코더만 활용하는 모델에 비해 상대적으로 길다. OpenAI (2023)에 의해 개발된 GPT 계열 모델이나 Alibaba Qwen Team (2023)에 의해 개발된 Qwen 언어모델은 이러한 디코더 기반 구조를 채택하고 있다. 디코더 기반 모델은 텍스트 생성 능력에서는 강점을 보이지만, 입력 텍스트를 구조적으로 분해하여 분석하는 데에는 상대적으로 제한이 있을 수 있다. 마지막으로, 인코더–디코더 기반 언어모델은 입력 텍스트의 의미적 표현(representation) 정보를 인코더에서 먼저 파악한 뒤, 디코더가 이를 참조하여 과업에 맞는 출력을 생성하는 구조를 가지며, 텍스트 시퀀스를 입력 및 처리하여 새로운 텍스트 시퀀스를 생성하는 시퀀스-투-시퀀스(sequence-to-sequence, Seq2Seq) 방식의 모델이다. Raffel 외 8인(2020)에 의해 개발된 T5 언어모델은 모든 자연어 처리 과업을 시퀀스-투-시퀀스 형식으로 통합한 대표적인 인코더–디코더 기반 모델로, 입력 텍스트 이해와 출력 텍스트 생성을 균형 있게 결합하는 모델이다.

이처럼 서로 다른 구조를 지닌 언어모델들은 주의집중 기법을 활용하는 방식에서도 다음과 같이 본질적인 차이가 존재한다. 인코더 기반 모델은 자기 주의집중을 통해 입력 텍스트 내 단어 간 의미적 관계를 정밀하게 파악하는 데 집중하며, 이는 텍스트의 전반적인 의미 이해에 유리하다. 그러나 이러한 주의집중은 평가 구인과 같은 외부 기준과 직접적으로 연결되기보다는, 입력 표현을 추상화하는 데 주로 사용된다. 디코더 기반 모델의 경우, 주의집중은 출력 텍스트 생성을 위한 조건부 참조에 초점이 맞추어져 있다. 즉, 모델은 다음 토큰을 생성하기 위해 입력 텍스트와 이미 생성된 토큰을 참조하며, 이 과정에서 주의집중은 앞으로 생성할 토큰의 생성 확률을 조정하는 역할을 수행한다. 따라서 디코더 기반 모델의 주의집중은 출력 중심적(output-oriented) 성격을 지닌다. 이에 비해, 인코더–디코더 기반 모델에서는 인코더의 자기 주의집중과 디코더의 교차 주의집중(cross-attention)이 결합되어 작동한다. 교차 주의집중은 디코더가 출력 텍스트 생성 과정에서 인코더가 추출한 입력 텍스트 표현 정보의 특정 부분을 선택적으로 참조하도록 한다.

기존 자동화 에세이 평가 연구에서 주의집중 기법은 주로 성능 향상을 위한 내부 기제로 활용되어 왔으며, 모델 유형 간 주의집중 양상의 차이를 평가 과정 차원에서 분석하려는 시도는 제한적이었다. 특히, 서로 다른 구조를 지닌 언어모델이 에세이 평가에서 어떤 단어에 주목하고, 이러한 주의집중 양상이 사람 채점자의 평가 전략과 어떻게 연결되는지에 대한 체계적인 분석은 충분히 이루어지지 않았다. 이에 본 연구는 인코더 기반, 디코더 기반, 인코더–디코더 기반 언어모델을 대상으로, 사람 채점자와의 점수 일치도, 내용어와 기능어에 대한 주의집중 양상, 그리고 실제 교육 현장을 고려한 채점 실용성(practicality)을 종합적으로 분석한다. 이를 통해 단순한 성능 비교를 넘어, 세 가지 유형의 주의집중 기반 언어모델이 영어 에세이를 다면적으로 평가하는 방식 자체를 설명 가능한 관점에서 재조명하고, 어떤 구조의 언어모델이 영어 쓰기 평가 맥락에서 사람 채점자의 판단을 가장 충실하게 모사하는지 분석하고자 한다.

분석적 에세이 평가(Analytic Scoring of Essay)

분석적 에세이 평가는 학습자의 쓰기 능력을 하나의 점수로 판단하는 총체적 평가(holistic scoring)와 달리, 쓰기 능력을 구성하는 여러 평가 구인에 대해 각각 평가하는 다면평가 방식을 의미한다. 일반적으로 분석적 평가는 내용(content), 조직(organization), 문법(grammar), 어휘(vocabulary), 구문(syntax), 응집성(cohesion) 등과 같이 사전에 정의된 평가 구인에 근거하여 점수를 산출하며, 각 구인은 명시적인 채점 기준(rubric descriptor)에 따라 평가된다(Jacobs et al., 1981; Setyowati et al., 2020; Weigle, 2002). 분석적 에세이 평가는 학습자의 쓰기 수행을 보다 세밀하게 진단할 수 있다는 점에서 교육적 의의가 크다. 총체적 평가가 에세이의 전반적인 완성도나 인상을 빠르게 판단하는 데 유리한 반면, 분석적 평가는 학습자의 강점과 약점을 구체적으로 드러내어 피드백과 교수·학습 개선에 직접적으로 활용할 수 있는 정보를 제공한다는 장점을 지닌다(Barkaoui, 2010; Weigle, 2002).

사람 채점자가 분석적 에세이 평가를 수행하는 과정은 단순히 점수를 나열하는 기계적 절차가 아니라, 에세이의 전체적인 의미를 먼저 파악한 뒤, 각 평가 구인을 기준으로 텍스트의 특정 부분을 선택적으로 재읽기(selective rereading)하는 인지적 과정을 포함한다(Freedman & Calfee, 1983; Lumley, 2002). 선행연구에 따르면, 숙련된 채점자일수록 에세이를 처음부터 끝까지 동일한 비중으로 읽기보다는, 평가 구인과 직접적으로 관련된 언어적 단서(linguistic cues)에 주의를 집중하여 판단을 내리는 경향을 보인다(Cumming et al., 2002; Wolfe, 1997). 또한, 높은 채점자 간 신뢰도(inter-rater reliability)를 보이는 채점자들은 공통적으로 루브릭에 제시된 구인별 기준을 일관되게 적용하며, 형식적 오류의 존재 여부 자체보다는 의미 전달에 미치는 영향을 중심으로 평가하는 것으로 보고된다(Lumley, 2005; Wolfe, 1997). 이러한 점에서 분석적 평가는 단순한 분해 평가가 아니라, 의미 중심 판단과 기준 참조 판단이 결합된 복합적 평가 방식으로 이해할 수 있다.

자동화 에세이 평가 연구 초기에는 총체적 점수 예측이 주된 목표였으나, 최근에는 분석적 평가의 교육적 가치를 반영하여 다면평가적 접근이 점차 확대되고 있다. 다면평가 기반 자동화 에세이 평가는 여러 평가 구인을 동시에 예측함으로써, 총체적 점수 예측에 비해 해석 가능성과 교육적 활용 가능성을 높일 수 있다(Do et al., 2024a; Kumar et al., 2022; Mathias & Bhattacharyya, 2020). 본 연구는 이러한 분석적 에세이 평가의 특성에 주목하여, 서로 다른 구조를 지닌 주의집중 기반 언어모델이 영어 에세이 다면평가에서 어떠한 평가 양상을 보이는지를 비교·분석한다. 특히 사람 채점자의 분석적 평가 과정에서 나타나는 선택적 재읽기와 기준 참조 판단이, 인코더 기반, 디코더 기반, 인코더–디코더 기반 언어모델의 주의집중 기법과 어떻게 대응되는지를 탐구함으로써, 언어모델의 평가 과정과 사람 채점자의 평가 과정의 유사성을 분석한다.

연구 방법

영어 에세이 평가 언어모델

본 연구에서는 영어 에세이 평가에서 언어모델의 구조적 차이가 사람 채점자와의 일치도, 단어 참조 양상, 그리고 채점 속도에 어떠한 영향을 미치는지를 분석하기 위해 서로 다른 구조를 지닌 세 가지 언어모델을 비교 대상으로 선정하였다. 구체적으로, 인코더 기반 언어모델인 BERT (Devlin et al., 2019), 디코더 기반 언어모델인 Qwen (Qwen Team, 2023), 그리고 인코더–디코더 기반 언어모델인 T5 (Raffel et al., 2020)를 활용하였으며, Figure 1은 각 모델의 에세이 평가 과정을 보여준다. BERT는 입력 텍스트를 인코더를 통해 의미적 표현으로 변환하는 데 특화된 구조를 지니며, 문장 이해 과업에서 강점을 보이는 대표적인 인코더 기반 모델이다. 반면, Qwen은 디코더 중심 구조를 기반으로 텍스트 생성을 주된 목표로 설계된 언어모델로, 출력 텍스트 생성 품질에 초점을 두는 특성을 지닌다. 마지막으로 T5는 인코더–디코더 구조를 통해 입력 텍스트의 의미적 정보를 인코더에서 종합적으로 처리한 후, 디코더가 해당 정보를 바탕으로 과업에 특화된 출력 텍스트를 생성하는 모델이다.

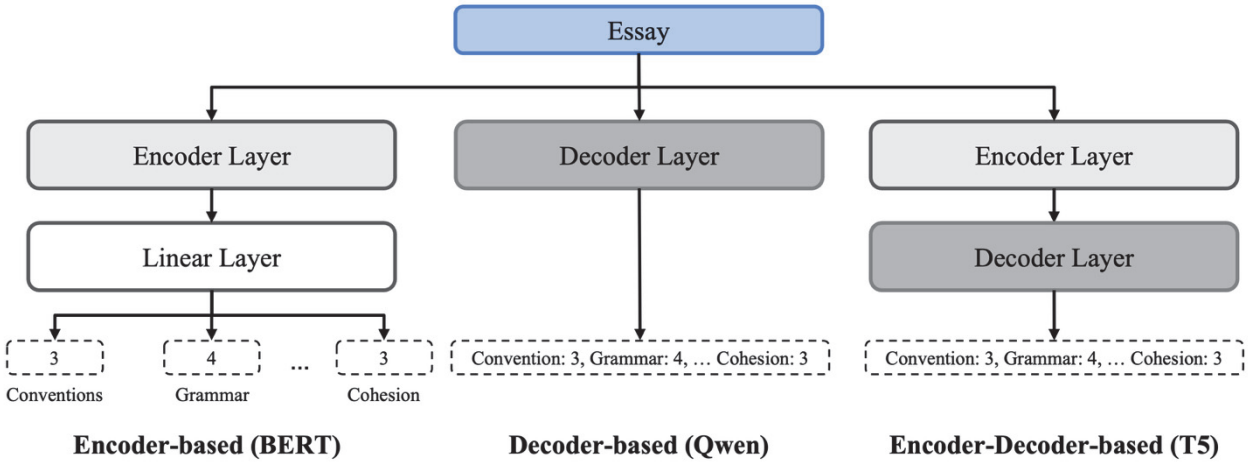


FIGURE 1
Architecture Comparison of Language Models for Essay Scoring

일반적으로 인코더 기반 모델의 인코더는 입력된 텍스트의 정보를 다차원으로 수치화된 임베딩 벡터(embedding vector)로 변환하며, 이 임베딩은 평가 점수 산출과 같은 다양한 후속 과업(downstream task)에 활용된다. 본 연구에서는 인코더가 변환한 에세이 임베딩 벡터를 각 평가 구인별 점수로 매핑하기 위해 단일 선형 레이어(linear layer)를 적용하여, 구인의 개수에 해당하는 점수를 직접 산출하도록 설계하였다. 본 연구에서는 총 여섯 가지 평가 구인을 사용하므로, 인코더 기반 모델은 최종적으로 여섯 개의 점수를 출력한다. 반면, 디코더 기반 모델과 인코더–디코더 기반 모델은 디코더를 통해 텍스트 생성을 수행하는 구조를 지닌다. 이에 따라 이들 모델은 인코더 기반 모델과 달리, 평가 구인별 점수를 개별 수치로 출력하지 않고 '평가 구인: 사람 채점자가 부여한 점수' 형태로 모든 구인의 채점 결과를 하나의 문장으로 나열(예: 'Conventions: 3, Grammar: 4, Phraseology: 5, Vocabulary: 3, Syntax: 3, Cohesion: 3')하여 생성한다. 디코더 기반 모델은 에세이를 입력받아 디코더를 통해 곧바로 구인별 점수로 구성된 텍스트를 생성하는 반면, 인코더–디코더 기반 모델은 먼저 인코더가 에세이 정보를 함축한 임베딩 벡터를 생성한 후 이를 디코더에 전달하며, 디코더는 해당 정보를 바탕으로 구인별 점수로 구성된 텍스트를 산출한다.

분석 대상 에세이

본 연구는 미국의 8~12학년 학습자가 작성한 영어 에세이를 활용하여 언어모델 기반 자동 에세이 평가를 수행하였다(Crossley et al., 2022). 언어모델 학습과 분석에 활용된 에세이는 Vanderbilt 대학교에서 수집되어 총 3,911편으로 구성되어 있으며, 다양한 사회경제적 배경과 인종을 지닌 학습자들이 작성한 에세이를 분석 대상으로 하였다. 학습자들은 컴퓨터 기반 환경에서 제시된 8가지 주제 중 하나를 선택하여 최소 150 단어 이상으로 구성된 에세이를 작성하였으며, 해당 주제의 구체적인 내용은 데이터 수집 관련 논문에서 공개되지 않았다(Crossley et al., 2022). 에세이 채점을 위한 평가 루브릭은 언어학, 언어교육학, 기계학습, 담화분석 분야의 전문가들이 참여한 복수의 협의와 개선 과정을 통해 개발되었다. 이후 2년 이상의 에세이 채점 경력을 가진 두 명의 사람 채점자가 해당 루브릭을 기반으로 평가 훈련을 받은 뒤, 이중 맹검(double-blind) 방식으로 각 에세이에 대해 여섯 가지 평가 구인별 분석적 평가를 수행하였다. 최종 점수는 두 평가자의 채점 점수 차이가 1점 이내이면 두 점수의 평균을 최종 점수로 하고, 점수 차이가 2점 이상이면 제3 평가자의 중재(adjudication)를 통해 합의하여 산출하였다. 본 연구에서는 이 점수를 언어모델 평가 결과와 비교하기 위한 준거 점수로 활용하였고, 8:2 비율로 나누어 언어모델의 학습(training)과 검증(validation)에 활용하였다. Table 1은 에세이 평가에 활용된 여섯 가지 평가 구인에 대한 세부 정보를 제시한다(Crossley et al., 2022, Crossley et al., 2023). 본 연구에서 해당 에세이 자료를 선택한 이유는 영어 에세이 평가를 위한 언어모델 학습을 위해서는 대규모 데이터가 요구되지만, 국내에는 1,000 건 이상의 영어 에세이를 대상으로 분석적 평가 방식에 따라 복수의 평가 구인에 대해 전문가 채점 결과를 공개한 자료가 부재하기 때문에 이에 대한 대안으로 해당 자료를 분석 대상 에세이로 선택하였다. 따라서, 본 연구의 결과를 국내 EFL 학습자 맥락으로 일반화하고 전이 효과를 검증하기 위해서는, 국내 EFL 환경에서 분석적 채점이 포함된 영어 에세이

데이터를 구축하는 것이 향후 매우 중요한 연구 과제로 남아 있다.

TABLE 1
Details of Essays and Constructs

Number of Essays	Average Number of Words (Standard Deviation)	Average Number of Sentences (Standard Deviation)	Construct	Description	Score Range
3,911 (Training: 3,129 / Validation: 782)	402.31 (188.38)	20.56 (9.76)	Conventions	The correct use of writing mechanics, including spelling, capitalization, and punctuation, to ensure clarity of meaning.	1 - 5
			Grammar	The accurate and consistent use of grammatical rules with few or no errors.	
			Phraseology	The effective use of natural and varied expressions, such as idioms and collocations, to convey precise and nuanced meanings.	
			Vocabulary	The use of a wide and appropriate range of words to convey precise meanings with minimal inaccuracies.	
			Syntax	The effective and flexible use of a range of sentence structures, including simple, compound, and complex forms, with minimal errors.	
			Cohesion	The ability to logically connect ideas across sentences and paragraphs using appropriate linguistic devices such as references and transitions.	

분석 대상 언어모델

본 연구에서 분석된 언어모델에 대한 구체적인 정보는 Table 2에 제시하였다. 언어모델이 데이터를 통해 특정 과업을 수행하도록 학습되는 과정에서, 사람의 두뇌에서 뉴런(neuron)이 수행하는 과정을 모사하는 요소를 파라미터(parameter)라 하며, 일반적으로 파라미터의 수가 증가할수록 모델의 표현력과 과업 수행 능력은 향상되는 경향을 보인다(Kaplan et al., 2020; Wei et al., 2022). 이에 따라 본 연구에서는 모델 간 비교의 공정성을 확보하기 위해, 파라미터 규모가 크게 상이하지 않으며 선행 연구에서 유사한 수준의 언어모델로 비교되어 온 모델들을 대표로 선정하였다. 이후 연구 결과 분석에서 자세히 논의하겠지만, 에세이 평가 과업에서는 파라미터 수의 증가가 평가 성능의 뚜렷한 향상으로 반드시 이어지지는 않는 것으로 나타났다.

TABLE 2
Details of Language Models

Model	Parameter Size	Developer	URL
BERT-Large-Uncased (BERT)	0.3B	Google	https://huggingface.co/google-bert/bert-large-uncased
Qwen3-0.6B (Qwen)	0.6B	Alibaba	https://huggingface.co/Qwen/Qwen3-0.6B
T5-Base (T5)	0.2B	Google	https://huggingface.co/google-t5/t5-base

Note. Parameter size is reported in billions (B) of parameters.

언어모델 학습 환경 및 구현 정보

본 연구에서 활용된 모든 언어모델은 일관된 실험 조건을 유지하기 위해 동일한 학습 및 추론 환경에서 실행되었다. 모델 구현 및 실험은 다양한 언어모델을 오픈소스로 제공하는 허깅페이스(Hugging Face, https://huggingface.co/)

플랫폼에서 공개된 사전학습 모델 가중치를 활용하여 수행하였다. 모델 학습을 위해 Python 3.12.3 프로그래밍 언어를 사용하였으며, 딥러닝 프레임워크인 PyTorch 2.8과 허깅페이스에서 제공하는 언어모델 학습 관련 라이브러리를 활용하였다. 모든 실험은 Intel i7-12700 CPU, NVIDIA A100 (80GB) GPU, 32GB RAM이 장착된 Linux (Ubuntu 22.04) 환경에서 수행되었다.

인코더 기반 언어모델인 BERT는 안정적인 학습을 위해, 구인별 점수를 산출하는 단일 선형 레이어에 시그모이드(sigmoid) 함수를 활성화 함수로 적용하여 출력 점수가 0.0과 1.0 사이로 정규화되도록 설계하였다. 이에 따라 사람 채점 점수 또한 최솟값(1점)과 최댓값(5점)을 기준으로 0.0–1.0 범위로 정규화하였으며, 손실 함수로는 평균 제곱 오차(mean squared error)를 사용하였다. 이후, 검증 과정에서는 모델이 산출한 점수를 다시 1점과 5점 사이의 범위로 재정규화하여 사람 채점 점수와의 일치도를 측정하였다. 반면, Qwen과 T5 언어모델은 평가 결과를 텍스트 형태로 생성하는 구조를 가지므로, 별도의 점수 정규화 과정 없이 5점 척도의 원점수를 그대로 학습에 활용하였다. 이들 모델은 텍스트 생성 기반 학습 방식을 사용하므로, 각 토큰의 생성 확률을 기반으로 하는 크로스엔트로피(cross-entropy)를 손실 함수로 적용하였다. 세 가지 언어모델 모두 허깅페이스 플랫폼에서 제공하는 SFTTrainer 모듈을 사용하여 학습되었으며, 학습률(learning rate)과 최적화 알고리즘(optimizer)은 SFTTrainer의 기본 설정(학습률: 0.00005, 최적화 알고리즘: AdamW)을 그대로 적용하였다. 언어모델 학습에 활용된 에세이의 수가 비교적 적은 것을 고려하여, 모델 학습은 5폴드 교차 검증(five-fold cross-validation) 방식으로 수행하였다. 각 폴드에 대해 모델은 최대 15에폭(epoch)까지 학습되었으며, 학습 데이터의 내용만 지나치게 학습하여 실제 새로운 에세이를 정확하게 평가하지 못하는 과적합(overfitting)을 방지하기 위해 조기 종료(early stopping) 기법을 적용하였다. 구체적으로, 검증 손실(validation loss)이 2에폭 연속으로 감소하지 않을 경우 학습을 종료하도록 설정하였다. 이러한 설정은 제한된 데이터 환경에서 모델의 일반화 성능을 안정적으로 확보하기 위한 목적에서 적용되었다.

마지막으로, 언어모델의 채점 실용성을 분석하기 위해 실제 교육 현장에서의 활용 가능성을 고려하여 CPU 환경과 GPU 환경에서의 추론 시간을 모두 측정하였다. GPU 환경에서는 병렬 연산이 가능한 설정을 활용하였으며, CPU 환경에서는 별도의 가속 장치 없이 순수 CPU 연산만을 사용하여 추론 시간을 산출하였다.

분석 도구

본 연구에서는 여섯 가지 평가 구인별로 언어모델의 평가 결과와 에세이 참조 양상을 다각도로 분석하기 위해 다음과 같은 분석 도구를 활용하였다. 먼저, 언어모델과 사람 채점자 간의 점수 일치도를 평가하기 위해 Quadratic Weighted Kappa (QWK) 지표를 사용하였다. QWK는 순서형 척도로 채점된 평가 결과 간의 일치도를 측정하는 지표로, 모델이 예측한 점수와 사람 채점 점수 간의 불일치 정도와 그 크기를 함께 반영한다. 이러한 특성으로 인해 QWK는 점수 간 선형적 관계만을 측정하는 Pearson 상관계수에 비해 평가 결과의 실질적인 일치도를 보다 엄격하게 반영하며, 에세이 평가 연구에서 사람 채점자 간 신뢰도 산출과 자동 채점 시스템의 성능 평가에 널리 활용된다.

다음으로, 언어모델의 단어 참조 양상을 분석하기 위해 각 언어모델의 주의집중 기법을 활용하였다. 모델별로 토큰 단위로 분할된 단어는 단어 단위로 재통합하여 주의집중 점수가 높은 상위 100개 단어를 추출한 후 분석하였다. 또한, 단어 유형 분석을 위해 spaCy 텍스트 마이닝 오픈소스 라이브러리(https://spacy.io/)를 활용하여 에세이 내 모든 단어를 문맥 기반으로 품사 태깅(POS tagging)하였다. 이후 상위 100개 단어에 부여된 품사 태그(POS tag)를 기준으로, British Council의 분류 체계와 Nation (2001)의 분류 체계를 참고하여 단어를 Table 3과 같이 내용어와 기능어로 구분하였다. spaCy는 기본적인 8품사뿐만 아니라 보조 동사(auxiliary verb)와 같이 각 단어의 의미적 쓰임새에 따른 태그도 제공하며, 유형별로 spaCy가 제공하는 태그는 Table 3에 제시되어 있다. 이때, spaCy는 Yes/No answers를 감탄사 태그(INTJ)로 분류하고, 소유대명사(possessive pronouns)와 관계대명사(relative pronouns)를 모두 대명사 태그(PRON)로 분류하여 더욱 세분화된 태깅이 제한되지만 내용어와 기능어의 구분 과정에서 겹치는 부분이 없기 때문에 spaCy의 태깅 방식을 그대로 유지하였다. 이후 각 모델이 참조한 내용어 및 기능어의 비율을 산출하고, 모델 간 차이를 검증하기 위해 일원배치 분산 분석(ANOVA) 및 사후 분석(post-hoc analysis)을 실시하였다. 본 연구에서는 ANOVA와 사후 분석을 수행하기에 앞서 분산 동질성 검정(Levene's test)을 실시하여 분산의 동질성을 확인하였고, 사후 분석으로는 Tukey의 HSD 검정을 적용하였다. 또한 동일한 에세이를 세 가지 언어모델로 평가한 값을 비교하기 때문에 평가 결과값 간 독립성 가정이 완전히 충족되지 않을 가능성이 있어 일원배치 ANOVA만으로는 모델 간 차이가 과대 추정될 위험을 보완하고자 에세이를 랜덤 효과(random effect)로 포함한 선형혼합모형(linear mixed-

effects model)을 추가적으로 적용하여 분석하였다. 본 연구의 모든 통계적 분석은 SPSS 29.0.2.0을 사용하여 수행하였다.

마지막으로, 언어모델 간의 에세이 평가 속도의 차이를 비교하기 위해 모델별로 에세이를 평가하는 데 소요된 평균 시간을 비교하였다. 딥러닝 기반 언어모델은 일반적으로 병렬 연산이 가능한 GPU 환경에서 높은 처리 효율을 보이지만, 실제 교육 현장에서는 GPU 연산 자원의 구축 및 유지가 현실적으로 어려운 경우가 많다. 이에 본 연구는 일선 교육 현장에서의 활용 가능성을 고려하여 CPU 연산 자원을 활용한 경우와 GPU 연산 자원을 활용한 경우의 각 모델의 채점 속도를 모두 산출하고, 모델 간 차이를 검증하기 위해 ANOVA 및 사후 분석과 선형혼합모형 기반 분석을 수행하였다.

TABLE 3

Part-of-speech (POS) Categories Used to Define Content and Function Words

Word Type	Part-of-speech (POS)
Content Words (POS Tags)	Nouns (NOUN), Verbs (VERB), Adjectives (ADJ), Adverbs (ADV), Negatives (PART), Exclamation (INTJ), Numbers (NUM), Yes/No answers (INTJ), Question words (QUE)
Function Words (POS Tags)	Auxiliary verbs (AUX), Prepositions (ADP), Conjunctions (CC), Determiners (DET), Pronouns (PRON), Possessive pronouns (PRON), Relative Pronouns (PRON)

연구 결과 및 논의

영어 에세이 평가에서 언어모델과 사람 채점자와의 일치도(연구문제 1)

연구문제 1을 분석하기 위해, 본 연구는 서로 다른 언어모델이 산출한 에세이 점수와 사람 채점자 점수 간의 일치도를 분석하였다. Table 4는 Conventions, Grammar, Phraseology, Vocabulary, Syntax, Cohesion의 여섯 가지 평가 구인에 대해, 각 언어모델과 사람 채점자 간의 QWK 점수를 제시한다. 분석 결과, 언어모델에 따라 사람 채점자와의 일치도에서 뚜렷한 차이가 나타났다. 전반적으로 인코더-디코더 기반 언어모델(T5)이 모든 평가 구인에서 가장 높은 일치도를 보였으며, 그 다음으로 디코더 기반 언어모델(Qwen), 마지막으로 인코더 기반 언어모델(BERT) 순으로 높은 일치도를 보였다.

구체적으로, T5 모델의 구인별 일치도는 0.588(최솟값, Vocabulary)에서 0.658(최댓값, Grammar) 범위로 나타났으며, 0.41과 0.60 사이의 QWK 점수로 'Moderate'한 수준의 일치도를 보이는 Vocabulary를 제외한 모든 구인에서 0.61과 0.80 사이의 QWK 점수로 'Substantial' 한 수준의 일치도를 보였다(Landis & Koch, 1977). 디코더 기반 언어모델인 Qwen은 0.543(최솟값, Vocabulary)에서 0.651(최댓값, Conventions) 범위에서 일치도를 보이며, 모든 평가 구인에서 BERT보다는 높은 일치도를, T5보다는 낮은 일치도를 나타냈다. 또한, 'Moderate'한 수준의 일치도를 보이는 Phraseology(0.600), Vocabulary(0.543), 그리고 Cohesion(0.557)을 제외한 모든 구인에서 'Substantial'한 수준의 일치도를 보였다. 이는 디코더를 중심으로 설계된 언어모델이 인코더를 중심으로 설계된 언어모델보다 사람 채점자의 평가 경향을 더 잘 반영하는 것을 보여줌과 동시에, 인코더의 부재가 모델의 평가 정확도를 일부 하락시킬 수 있음을 시사한다. 반면, 인코더 기반 언어모델인 BERT는 0.514(최솟값, Cohesion)에서 0.587(최댓값, Conventions) 범위에서 일치도를 보이며 전반적으로 'Moderate'한 수준에서 사람 채점자와의 일치도를 보였고, 다른 언어모델과 비교하였을 때 그 일치도는 가장 낮았다. 이는, 인코더만 활용하는 언어모델은 인코더와 디코더를 모두 활용하거나, 디코더만 활용하는 모델과 비교하였을 때, 전반적으로 사람 채점자의 판단을 충분히 반영하는 데에는 한계가 있는 것을 시사한다.

이러한 모델별 채점 양상을 사람 채점자가 영어 에세이를 분석적 평가 관점에서 수행하는 평가 과정과 비교하면 다음과 같다. 선행 연구에 따르면 사람 채점자의 영어 에세이 평가 과정에는 개인차가 존재하지만, 일반적으로 에세이의 전체 내용을 1차적으로 파악한 후 평가 구인별 채점 기준과 연결하여 선택적 재읽기를 수행하며 구인별 점수를 산출한다(Cumming et al., 2002; Freedman & Calfee, 1983; Lumley, 2002). 특히 하나의 에세이로부터 복수의 평가 구인을 판단해야 하는 분석적 평가 상황에서, 높은 채점자 간 신뢰도를 보이는 채점자들은 공통적으로 구인별 채점 기준에 더욱 집중하여 평가하는 경향을 보인다(Barkaoui, 2010; Lumley, 2005).

이러한 평가 과정은 T5 모델에서 인코더와 디코더 간에 이루어지는 교차 주의집중 메커니즘과 유사하다(Raffel et al., 2020). T5 모델의 인코더는 먼저 자기 주의집중을 통해 에세이 내 단어들 간의 의미적 관계를 파악함으로써 텍스트의 전반적인 내용을 이해한다. 이후 디코더는 인코더가 추출한 에세이 표현을 바탕으로, 모델 학습 과정에서 내재화된 구인별 정보를 교차 참조하며 관련된 텍스트 단서를 선택적으로 활용하여 최종적으로 구인별 점수를 텍스트 형태로 생성한다. 이와 같은 구조적 특성은 사람 채점자의 평가 과정과 가장 높은 유사성을 보이며, 인코더–디코더 기반 모델이 사람 채점자와 가장 높은 일치도를 보인 이유로 해석할 수 있다.

다음으로, 디코더 기반의 Qwen 모델과 사람 채점자의 평가 과정을 비교하면 다음과 같다. 디코더 기반 모델은 구조적으로 입력 텍스트를 출력 텍스트 생성 관점에서 처리하며, 출력 결과의 생성 품질에 상대적으로 더 큰 비중을 두는 특성을 지닌다(Fu et al., 2023; Qorib et al., 2024). 이는 인코더–디코더 기반 모델이 입력과 출력을 보다 균형적으로 고려하는 것과 대비된다. 이러한 특성은 분석적 평가 과정에서 사람 채점자가 구인별 채점 기준을 중심으로 판단을 수행하는 양상과 유사하며, 그 결과 Qwen 모델은 사람 채점자와 두 번째로 높은 일치도를 보인 것으로 해석할 수 있다.

마지막으로, 인코더 기반 BERT 모델은 자기 주의집중을 통해 에세이 내 단어들 간 의미적 관계를 파악하는 데에 초점을 둔 구조를 지닌다(Devlin et al., 2019). 즉, 에세이의 전반적인 내용 이해에는 강점을 보이나, 이를 구인별 채점 기준과 명시적으로 연결하여 평가 점수로 전환하는 과정은 구조적으로 제한적이다. 이러한 특성은 사람 채점자의 평가 과정 중 일부 측면만을 반영하는 것으로 볼 수 있으며, 이로 인해 인코더 기반 모델이 인코더–디코더 기반 또는 디코더 기반 모델에 비해 사람 채점자와의 일치도가 상대적으로 낮게 나타난 이유를 설명할 수 있다.

종합하면, 본 연구의 결과는 에세이 평가에서 인코더–디코더 기반 언어모델이 사람 채점자의 평가 과정을 가장 충실하게 모사하며, 이에 따라 사람 채점자의 판단과 가장 높은 일치도를 보인다는 점을 보여준다. 동시에 이는 모델이 일관된 채점 기준을 안정적으로 적용하고 있음을 의미하며, 본 연구에서 활용된 에세이가 3인 전문가의 협업을 통해 채점된 것을 고려하였을 때(Crossley et al., 2022) , 인코더-디코더 기반 언어모델이 훈련된 사람 채점자 수준의 에세이 평가 능력을 달성하고 있음을 보여준다. 한편, 사람 채점자와의 일치도는 디코더 기반 언어모델, 인코더 기반 언어모델 순으로 감소하는 경향을 보였는데, 이는 각 언어모델의 텍스트 이해 방식과 구인별 점수 산출 과정이 사람 채점자의 평가 과정과 얼마나 유사한지에 따라 평가 일치도가 영향을 받음을 시사한다.

TABLE 4

Comparison of Essay Scoring Performance

Model	Con	Gra	Phr	Voc	Syn	Coh	Avg.
BERT (Encoder-based)	0.587 (0.018)	0.576 (0.029)	0.568 (0.032)	0.530 (0.035)	0.563 (0.018)	0.514 (0.010)	0.556 (0.021)
Qwen (Decoder-based)	0.651 (0.028)	0.635 (0.039)	0.600 (0.055)	0.543 (0.056)	0.613 (0.036)	0.557 (0.023)	0.600 (0.032)
T5 (Encoder-Decoder-based)	0.652 (0.020)	0.658 (0.026)	0.632 (0.016)	0.588 (0.035)	0.646 (0.038)	0.610 (0.023)	0.631 (0.028)

Note. Scoring traits include: Con: Conventions, Gra: Grammar, Phr: Phraseology, Voc: Vocabulary, Syn: Syntax, Coh: Cohesion, Avg.: Average. Values in parentheses denote the standard deviation across 5 folds.

영어 에세이 평가에서 언어모델의 내용어 및 기능어 참조 양상(연구문제 2)

연구문제 2를 분석하기 위해, 본 연구는 각 언어모델이 영어 에세이 평가 과정에서 에세이 내 개별 단어를 어떠한 양상으로 참조하는지 분석하였다. 트랜스포머 구조로 설계된 언어모델은 입력된 텍스트의 각 토큰에 주의집중을 할당하는 방식으로 텍스트를 참조한다. 그러나 모델이 입력된 모든 토큰을 동일한 수준으로 참조하지는 않기 때문에, 상대적으로 중요도가 낮은 토큰으로 인한 노이즈를 제거하고자 본 연구에서는 모델별 주의집중 점수(0.0–1.0)를 기준으로 상위 100개 단어를 추출하여 분석하였다. 분석의 일관성을 확보하기 위해, 모델별 토크나이저(tokenizer)가 하나의 단어를 여러 토큰으로 분할한 경우(예: playing → play, ing), 이를 하나의 단어로 통합하고, 해당 단어를 구성하는 토큰들의 주의집중 점수를 합산하여 단어 단위의 주의집중 점수로 산출하였다. Figure 2는 각 모델이 주의집중한 상위 100개 단어에 대한 주의집중 점수를 검증에 사용된 전체 에세이(782편)에 대해 평균한 분포를

나타낸다. Figure 2에서 확인할 수 있듯이, 상위 약 50개 단어 이후부터는 주의집중 점수가 점진적으로 감소하는 양상이 나타나며, 상위 약 70개 단어 이후부터는 주의집중 점수가 급격하게 감소하는 것을 확인할 수 있다. 이는 각 모델이 에세이 내에서 약 50개 내외의 단어를 특히 집중적으로 참조하고 있음을 시사한다.

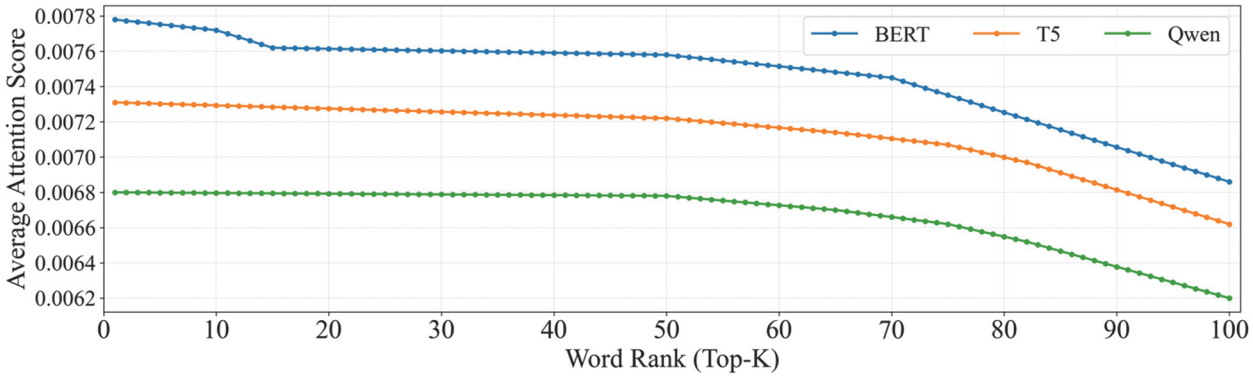


FIGURE 2
Average Attention Scores across Word Ranks for Different Language Models

선행연구에 따르면, 사람 채점자는 영어 에세이를 평가할 때 의미적 요소와 형식적 요소를 모두 고려하여 수행한다(Hamp-Lyons, 2003; Lumley, 2005; Winke & Lim, 2015). 이에 따라 본 연구에서는 모델이 참조한 상위 100개 단어가 의미적 요소와 형식적 요소의 관점에서 어떠한 특성을 보이는지를 분석하고자 하였다. 이를 위해 앞서 설명한 바와 같이 에세이 내에서 문맥 기반으로 모든 단어의 품사를 태깅한 후, 의미적 쓰임새에 따라 상위 100개 단어에 부여된 품사 태그를 활용하여 Table 3을 기준으로 내용어와 기능어로 분류하였다. 이후 에세이별로 각 모델이 주의집중한 내용어와 기능어의 비율을 산출하고, 모델 간 차이를 검증하기 위해 ANOVA와 선형혼합모델을 적용한 결과를 Table 5, 6, 7에 제시하였다.

TABLE 5
Results of Analysis of Variance on the Proportion of Content and Function Words by Model

Word Type	BERT (N = 782)	Qwen (N = 782)	T5 (N = 782)	Mean Square (Between Groups)	F	η^2
Content Words	.628	.653	.689	.741	271.670***	.188
(Function Words)	(.372)	(.347)	(.311)			

Note. *p < .05, **p < .01, ***p < .001.

Table 5에 제시된 바와 같이, ANOVA 분석 결과 각 모델이 참조한 내용어와 기능어의 비율에는 통계적으로 유의한 차이가 나타났다(p < .001, η^2 = .188). 선형혼합모형 분석 결과, Table 6에서 보고되는 바와 같이 에세이를 임의 효과로 포함하였을 때 에세이 수준의 분산은 0.002, 잔차 분산은 0.001로 나타났으며, 이에 따라 산출된 층내상관계수(intraclass correlation coefficient, *ICC*)는 약 0.667이었다. 이는 각 모델이 참조한 내용어와 기능어 비율 변동의 상당 부분이 에세이 간 차이에 기인함을 시사하며, 에세이를 임의 효과로 포함하는 것이 타당함을 뒷받침한다. 또한, 모델별 고정 효과를 고려한 후에도 Table 6과 같이 각 모델 간 내용어와 기능어의 참조 비율에서 통계적으로 유의한 차이가 확인되어 모델 특성이 평가 결과에 미치는 영향이 존재함을 확인할 수 있다. 그러나, 효과크기(η^2)는 작은 수준(small effect, η^2 < 0.2)에 해당하여, 각 모델이 내용어와 기능어를 참조하는 정도에 차이는 있으나 그 폭은 작아 실제적인 영향력은 제한적일 가능성을 시사한다. 이러한 결과를 언어모델과 사람 채점자의 평가 과정을 비교하며 분석하면 다음과 같다.

먼저, 사람 채점자와 모든 평가 구인에서 가장 높은 일치도를 보인 T5 모델은 내용어를 가장 높은 비율로 참조하고(68.9%), 기능어를 가장 낮은 비율로 참조하는 양상(31.1%)을 보였다. 그 다음으로 Qwen 모델(내용어 65.3%, 기능어 34.7%), BERT 모델(내용어 62.8%, 기능어 37.2%) 순으로 내용어 참조 비율이 감소하고 기능어 참조

비율이 증가하는 경향이 확인되었다. 또한, 모델 간 내용어 및 기능어 참조 비율의 차이는 Table 7의 사후 분석 결과에서 모두 통계적으로 유의한 것으로 나타났다($p < .001$). 이러한 결과는 사람 채점자와 보다 높은 수준의 일치도를 보이는 언어모델일수록 기능어에 비해 내용어에 더욱 집중하는 경향이 있음을 의미한다. 이는 선행 연구에서도 실제로 훈련된 사람 채점자가 영어 에세이를 분석적으로 평가할 때, 형식적 오류의 존재 여부 자체보다는 에세이가 의미를 효과적으로 전달하고 있는지(Lumley, 2005; Winke & Lim, 2015), 사고력과 조직력이 충분히 드러나는지(Setyowati et al., 2020)와 같은 의미 중심의 평가 요소에 더 큰 비중을 둔다고 보고된 바 있다. Weigle (2002)에 따르면 형식적 요소의 경우에도 사람 채점자는 고려하는 비중이 크지 않지만 에세이의 전반적인 의미 전달을 현저히 저해하는지 여부를 중심으로 판단하는 경향이 있다는 점에서, 본 연구의 결과는 사람 채점자와 유사한 양상으로 영어 에세이를 평가하는 모델일수록 그 채점 과정이 실제 사람 채점자의 채점 과정과 더욱 유사하다는 것을 시사한다.

TABLE 6
Results of Linear Mixed-Effects Model on the Proportion of Content and Function Words (parentheses) by Model

Fixed Effects	Estimate	*SE*	*t*	Random Effect	Variance	Residual Variance	*ICC*
Qwen-BERT	Content: .025 Functional: -.025	.002	15.685***	Essay	0.002	0.001	0.667
T5-BERT	Content: .062 Functional: -.062	.002	23.119***				
T5-Qwen	Content: .037 Functional: -.037	.002	23.225***				

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

TABLE 7
Results of Post-hoc Analysis on the Proportion of Content and Function Words by Model

	Mean Difference (row – column)		
Model	1	2	3
1. BERT	-		
2. Qwen	.025***	-	
3. T5	.061***	.037***	-

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

종합하면, 내용어를 상대적으로 더 많이 참조하는 언어모델일수록 사람 채점자와 전반적으로 높은 일관성을 보이며, 이는 언어모델 기반 영어 에세이 평가에서 내용어 중심의 평가 전략이 보다 정확한 평가를 가능하게 함을 시사한다. 나아가 이러한 결과는 인코더–디코더 기반 아키텍처가 영어 글쓰기 다면평가에서 의미적 요소와 형식적 요소를 차등적으로 고려하는 사람 채점자의 판단 과정을 어느 정도 충실하게 모사하고 있음을 보여준다. 그러나, 모델 간 차이의 효과크기가 작은 수준인 점을 고려하면, 이러한 차이는 통계적으로 유의하더라도 채점 양상의 실질적 차이를 크게 설명한다고 보기는 어렵다. 따라서 내용어 중심 참조 경향은 트랜스포머 기반 언어모델의 핵심 작동 원리인 주의집중 기법을 기반으로 사람 채점자와의 유사성을 설명하는 하나의 보조적 신호로 해석될 수 있으며, 추가적인 요인에 대한 후속 연구가 필요하다.

언어모델별 영어 에세이 평가 사례 분석: 내용어 및 기능어 주의집중

Table 8은 본 연구에서 제안한 방법을 적용하여, 하나의 학습자 에세이에 대해 세 개의 언어모델이 참조한 단어별 주의집중 양상의 시각화 자료를 제시하고 있다. 또한, 표의 우측에는 여섯 가지 구인에 대해 각 모델이 예측한 점수가 제시되어 있으며, 표의 제목에는 사람 채점자가 평가한 점수가 제시되어 있다. 이는 앞서 설명한 바와 같이 두 명의 훈련된 사람 채점자가 평가 루브릭을 활용하여 채점한 후 점수 차가 1점 이내면 평균하고, 2점 이상이면

제3 평가자의 중재에 의해 최종 부여된 점수이다. 구인별 점수를 텍스트 형태로 생성하는 T5 언어모델과 Qwen 언어모델은 모델 학습에 레이블로 활용된 형태와 동일하게 평가 점수를 출력하여 사람 채점자와 유사하게 점수를 0.5 단위로 산출하였으나(예: Cohesion: 2.5), BERT 언어모델은 최종적인 점수 산출이 텍스트 형태가 아니라 에세이 임베딩 벡터가 단일 선형 레이어에 의해 변환된 실수인 점을 고려하여 소수점 둘째 자리에서 반올림하여 나타내었다(예: Cohesion: 1.2). Table 8에서 에세이의 모든 단어 중 각 모델이 부여한 주의집중 점수의 상대적 크기에 따라 상위 30%(57개) 단어는 굵은 글씨와 밑줄로 표현하였고(예: **believe**), 하위 30% 단어는 굵은 글씨와 밑줄로 표현하되, 회색으로 구분하였다(예: **we**).

시각화 결과를 분석하면, T5 언어모델은 'believe', 'inactivity', 'important', 'refresh', 'agree' 등의 에세이의 핵심 의미를 구성하는 명사, 동사, 형용사와 같은 내용어에 상대적으로 더 높은 주의집중을 할당하는 경향을 보여 앞서 표5, 6, 7에서 분석된 결과와 일관되는 것을 확인할 수 있다. 이러한 단어들은 에세이의 주제 진술과 논지 전개에 직접적으로 기여하는 요소로, 사람 채점자가 에세이 분석적 평가에서 고려하는 의미적 단서와 일치한다. 반면, 'we', 'are', 'can', 'to', 'the', 'and', 'that' 등 대명사, 보조 동사, 부정사와 같은 기능어는 문법적 연결에는 필요하지만 의미 판단에 미치는 영향이 상대적으로 낮아 전반적으로 약한 주의집중을 보인다. 또한, 'possiblilities', 'exshuased', 'ours bodys', 'mind set'과 같이 철자나 복수형 오류가 포함된 내용어에도 높은 주의집중이 할당되어, 의미 전달에 중요한 단어의 오류를 모델이 포착하여 평가에 반영하고 있음을 확인할 수 있다. 이러한 양상은 일부 구인에서 사람 채점 점수와 완전히 일치하지는 않지만, 대부분 0.5점 이내의 오차 범위에서 비교적 정확한 구인별 채점을 가능하게 하였으며, 이는 채점 과정에서 사람 채점자와의 유사성에 기인한 것으로 해석할 수 있다. 한편 일부 기능어가 중간 수준 이상의 주의집중을 유지하는 점은 모델이 의미 중심 평가를 수행하면서도 문장 연결성과 담화 흐름에 필요한 최소한의 형식적 단서를 함께 고려하고 있음을 시사한다. 그러나, 'arent'와 같은 축약형 표현의 오류는 포착하지 못해 T5 언어모델의 주의집중 양상에도 여전히 한계가 존재하는 것을 확인할 수 있으며 사람 채점자와의 채점 일치도 향상을 위해서는 주의집중이 더욱 정확하게 이루어질 수 있는 방법론에 대한 후속 연구가 필요하다.

Qwen 언어모델 역시 내용어 중심의 주의집중 경향을 보이지만, T5 언어모델에 비해 'believe', 'things', 'possiblilities', 'think' 등의 일부 내용어에 대한 주의집중 점수가 상대적으로 낮고, 'is', 'we', 'the', 'a'와 같은 기능어에 대한 주의집중의 가중치가 증가하는 양상이 나타났다. 또한, T5 언어모델이 주의집중한 오류 표현(예: 'possiblilities', 'exshuased', 'bodys')에 대한 주의집중도 일부 감소하였다. 이러한 차이는 Qwen 언어모델의 채점 결과가 T5 언어모델에 비해 Syntax, Grammar, 그리고 Conventions와 같이 문장이나 어구 간 연결 또는 철자나 문법 오류와 같은 부분을 채점하는 평가 구인에서 사람 채점 점수와 더 큰 오차를 보이게 한 원인으로 해석될 수 있다.

마지막으로, BERT 언어모델은 다른 두 언어모델에 비해 기능어에 주의집중하는 경우가 비교적 많으며, 오류 표현이나 의미적으로 중요한 내용어에 대한 주의집중은 상대적으로 부족한 경향을 보인다. 특히, 에세이의 주제 전달과 관련이 없고, 평가자의 의미 이해를 보조하는 정도가 낮은 보조 동사, 대명사, 관사 등에 대한 주의집중이 더욱 증가하는 것을 확인할 수 있다. 이러한 패턴은 BERT 언어모델의 구인별 평가 점수가 사람 채점 점수와 크게 차이나는 현상과 연관될 가능성을 시사한다.

종합하면, 본 사례 분석은 인코더-디코더 기반의 T5 언어모델이 영어 에세이를 평가하는 과정에서 인코더 혹은 디코더만으로 구성된 모델에 비해 인코더가 에세이 내용을 충분히 이해하고, 디코더가 각 구인별 특성과 에세이 내용을 효과적으로 연결시키는 작용을 균형있게 활용함으로써 내용어를 중심으로 의미적 정보를 우선적으로 처리하고, 기능어는 보조적인 담화 단서로 활용하는 전략을 취하고 있음을 정성적으로 보여준다. 이는 본 연구의 정량 분석 결과를 보완하는 사례로서, 인코더-디코더 기반의 언어모델이 사람 채점자의 의미 중심적이고 분석적인 평가 과정을 가장 충실하게 모사하고 있음을 뒷받침한다. 나아가, T5 언어모델이 추출한 내용어 중심의 주의집중 정보는 단순한 해석을 넘어 실제 교육 현장에서 활용 가능한 시각적, 형성평가적 피드백(formative visual feedback)으로 확장될 수 있다. 예를 들어, 에세이 내 단어별 주의집중 점수를 기반으로 학습자에게는 자신의 글에서 의미 전달에 핵심적인 내용어가 어떻게 사용되었고 평가되었는지 직관적으로 보여줄 수 있으며, 이는 학습자가 자신의 글이 얼마나 주제 중심적으로 구성되었는지 여부와 단어 및 문장 구성을 스스로 점검하도록 돕는다. 또한, 철자 오류나 부적절한 표현이 포함된 내용어에 높은 주의집중이 할당된 경우, 해당 단어를 강조 표시하여 교사가 별도의 설명 없이도 학습자가 스스로 오류를 인식하고 수정하도록 유도할 수 있다. 교사의 입장에서는 이러한 시각화 정보를 활용하여 학생별로 진단하고, 개별화된 피드백을 구성하기 위한 보조자료로 활용할 수 있다. 이는 기존의 점수 중심 피드백을 넘어, 학습자의 사고 과정과 글의 의미 구조를 가시화한다는 점에서 설명가능한 AI 기반 형성평가 도구로서의 잠재력을 지닌다.

TABLE 8
Visualization of Model Attention to Content and Function Words

Model	Attention	Model-Predicted Scores
T5	I **believe** Thomas Jefferson is right. If we are always doing something we **get things done faster** and there is **possiblilities** of **starting** new things. Why should we **wait** to **finish** something we can **finish** the same **day,** we never know what can **happen.** I **also believe** the **inactivity** is **good** not only does it give a chance to **rest** but we **also** can **think** through our **ideas.** We can **perfect things** we **didnt think** we could. It can **give** a **chance** to **see** different views of are **own ideas.** Maybe **taking** a **break** can **refresh** are **memories** and remind us how **important** things are to us. Woring all the time can **exshuased ours bodys** and make us think a **little** less. With a fresh **mind set** it can **make things easier** than we **thought. Sometimes** we over thinks that arent that big of a deal and a big part of it **comes** from not **taking** the **time** to **rest** and **refresh** our **memories. Many** people would agree with Thomas Jefferson **just** like I do but would also **agree** that taking **time** to **rest** is also getting things **done.**	Cohesion 2.5/5, Syntax 2.5/5, Vocabulary 2.5/5, Phraseology 2.5/5, Grammar 2.5/5, Conventions 3.0/5
Qwen	I believe Thomas Jefferson **is** right. If **we** are always **doing** something we **get** things **done faster** and there is possiblilities of starting **new** things. **Why** should **we** **wait** to **finish** something we can **finish** the **same day,** we never **know what** can **happen.** I **also believe the inactivity** is **good** not only does it give a **chance** to **rest** but we also **can** think through **our ideas.** We can **perfect** things we didnt think we could. It can **give** a **chance** to see different views of are **own ideas. Maybe taking** a **break can refresh** are memories and remind us how **important** things are to us. Woring all the **time** can exshuased ours bodys and make **us** think a **little** less. With a **fresh mind** set it can **make** things **easier** than we **thought. Sometimes** we over thinks that arent that big of **a** deal and a big part of it **comes** from not **taking** the **time** to **rest** and **refresh** our **memories.** Many people would agree with Thomas Jefferson **just** like I do but would also agree that taking **time** to **rest** is also getting things done.	Cohesion 2.5/5, Syntax 2.0/5, Vocabulary 2.5/5, Phraseology 2.5/5, Grammar 3.0/5, Conventions 2.0/5
BERT	**I** believe Thomas Jefferson is right. If **we** are always doing something **we** get things done **faster** and there is possiblilities **of** starting new **things.** Why should **we** wait to finish something we can **finish** the **same day,** we never **know** what can **happen.** I also believe the **inactivity** is **good** not only **does** it give **a chance** to rest **but** we also **can think** through **our ideas.** We can perfect **things** we didnt think we could. **It** can **give** a chance to **see different views of** are own **ideas.** Maybe taking a **break can refresh** are memories **and** remind us how **important** things are to us. Woring all **the time** can exshuased ours bodys and make us think **a little less.** With **a** fresh **mind** set it can **make things** easier than we **thought. Sometimes** we over thinks **that** arent that big of **a** deal and a big part of it **comes** from not **taking** the **time** to **rest** and **refresh** our **memories.** Many people would agree with Thomas Jefferson just like I do but would also agree that taking **time** to **rest** is also **getting** things done.	Cohesion 1.2/5, Syntax 1.4/5, Vocabulary 4.3/5, Phraseology 3.6/5, Grammar 4.0/5, Conventions 1.1/5

Note. Words in the top 30% of attention weights are shown in bold and underlined, whereas words in the bottom 30% are also shown in bold and underlined but in gray. (Real Scores: Cohesion 2.5/5, Syntax 2.5/5, Vocabulary 3.0/5, Phraseology 2.5/5, Grammar 2.5/5, Conventions 3.0/5)

영어 에세이 평가에서 언어모델의 에세이 채점 속도(연구문제 3)

연구문제 3과 관련하여 Table 9는 각 언어모델의 CPU 환경과 GPU 환경에서의 에세이별 채점 속도의 평균을 제시한다. Table 9에 제시된 결과에 따르면, 인코더 기반의 BERT 모델은 CPU 환경에서 평균 0.337초, GPU 환경에서 평균 0.014초로 가장 짧은 처리 시간을 보였다. 다음으로 인코더–디코더 기반의 T5 모델은 CPU 환경에서 0.792초, GPU 환경에서 0.181초가 소요되었으며, 디코더 기반의 Qwen 모델은 CPU 환경에서 1.506초, GPU 환경에서 0.772초로 가장 긴 처리 시간을 보였다. ANOVA 결과, CPU 및 GPU 환경 모두에서 모델 간 채점 속도의 차이는 통계적으로 유의하게 나타났다($p < .001$, CPU $\eta^2 = .038$, GPU $\eta^2 = .131$). 선형혼합모형 분석 결과, 에세이를 임의 효과로 포함하여 CPU와 GPU 환경 모두에서 에세이 수준의 분산과 잔차 분산을 고려하였을 때 층내 상관 계수가 각각 0.590과 0.631로 산출되어 모델 간 채점 속도 변동의 상당 부분이 에세이 간 차이에 기인함을 시사하며, 에세이를

임의 효과로 포함하는 것이 타당함을 뒷받침한다. 또한, 모델별 고정 효과를 고려한 후에도 Table 10과 같이 각 모델 간 채점 속도에서 통계적으로 유의한 차이가 확인되었다. 추가적으로, Table 11의 사후 분석 결과 또한 모든 모델 쌍 간 차이가 통계적으로 유의함을 보여준다($p < .001$). 그러나, CPU와 GPU 환경 모두의 효과 크기는 모두 작은 수준($\eta^2 < 0.2$)으로 나타나, 모델 간 채점 속도 차이의 실제적 영향력은 제한적일 가능성을 시사한다. 따라서 이러한 결과는 모델 간 채점 속도 차이가 통계적으로는 유의하더라도 그 실질적인 차이는 비교적 크지 않을 수 있음을 의미하며, 해석 시 주의가 필요하다. 이에 따라 언어모델 간 에세이 채점 속도 차이를 논의할 때에는 단순한 속도 비교에 그치지 않고, 각 모델의 채점 정확도와의 관계를 함께 고려하여 보다 종합적으로 해석할 필요가 있다.

일반적으로 딥러닝 모델은 파라미터 규모가 증가할수록 성능 향상과 함께 처리 속도의 저하가 동반되지만(Kaplan et al., 2020; Wei et al., 2022), 언어모델의 에세이 채점 속도는 파라미터 수 자체보다도 아키텍처 구조와 연산 흐름의 복잡성에 더 크게 영향을 받는다. 인코더 기반의 BERT 모델은 입력 텍스트를 한 번의 전방 계산으로 처리하고 추가적인 생성 단계를 요구하지 않기 때문에 가장 빠른 처리 속도를 보인다. 반면, 디코더를 포함하는 T5와 Qwen 모델은 입력된 에세이 텍스트로부터 디코더가 구인별 점수를 순차적으로 생성하는 구조적 특성으로 인해 상대적으로 속도가 느려지지만, 에세이 평가 정확도는 오히려 향상되는 경향을 보였다. 특히 Qwen 모델이 T5 모델보다 약 세 배 많은 파라미터를 보유함에도 불구하고, T5 모델이 사람 채점자와 더 높은 일치도를 보였다는 점은 주목할 만하다. 이는 에세이 평가와 같이 에세이 내용 이해와 구인별 판단을 동시에 요구하는 과업에서는, 입력 자료와 출력 자료를 균형 있게 처리할 수 있는 인코더–디코더 결합 구조가 인코더 혹은 디코더만 활용하는 모델보다 사람 채점자의 평가 과정을 더 효과적으로 모사함을 시사한다.

TABLE 9

Results of Analysis of Variance on Essay Scoring Duration by Model (sec)

Computing Resource	BERT (0.3B)	Qwen (0.6B)	T5 (0.2B)	Mean Square (Between Groups)	*F*	η^2
CPU	.337	1.506	.792	272.483	46.422***	.038
GPU	.014	.772	. 181	123.807	176.757***	.131

Note. *$p < .05$, **$p < .01$, ***$p < .001$. Parameter size is reported in billions (B), as indicated in parentheses.

TABLE 10

Results of Linear Mixed-Effects Model on Essay Scoring Duration by Model (sec)

Computing Resource	Fixed Effects	Estimate	*SE*	*t*	Random Effect	Variance	Residual Variance	*ICC*
CPU	Qwen-BERT	1.171	.118	9.918***	Essay	0.036	0.025	0.590
	Qwen-T5	.716	.118	6.060***				
	T5-BERT	.455	.118	4.217***				
GPU	Qwen-BERT	.757	.042	17.916***	Essay	0.041	0.024	0.631
	Qwen-T5	.590	.042	13.961***				
	T5-BERT	.167	.042	77.027***				

Note. *$p < .05$, **$p < .01$, ***$p < .001$.

TABLE 11

Results of Post-hoc Analysis on Essay Scoring Duration by Model (sec)

		Mean Difference (row – column)		
Computing Resource	Model	1	2	3
CPU	BERT (0.3B)	-		
	Qwen (0.6B)	1.171***	-	
	T5 (0.2B)	.455***	.716***	-
GPU	BERT (0.3B)	-		
	Qwen (0.6B)	.757***	-	
	T5 (0.2B)	.167***	.590***	-

Note. *$p < .05$, **$p < .01$, ***$p < .001$. Parameter size is reported in billions (B), as indicated in parentheses.

더 나아가, 채점 속도는 자동화 평가 시스템의 실용성(Bachman & Palmer, 1996; Fulcher, 2024)을 구성하는 핵심 요소 중 하나이다. 학교 현장에서 다수의 학생 산출물을 제한된 기간 내에 평가해야 하는 교사의 업무 특성을 고려할 때, 신속하고 일관된 평가 도구에 대한 요구는 매우 크다(Yip & Cheung, 2005). 이에 따라 자동 채점 시스템이 실제 평가 상황에서 활용되기 위해서는 사람 채점자와의 높은 일치도뿐만 아니라, 대규모 응답을 효율적으로 처리할 수 있는 처리 속도 또한 중요한 모델 선택 기준으로 작용한다. 이러한 맥락에서 본 연구의 결과는 언어모델의 구조적 복잡도가 증가할수록 채점 속도에 일정 수준의 계산 비용이 수반됨을 보여준다. 그러나, 인코더–디코더 기반의 T5 모델은 BERT 모델에 비해 다소 긴 처리 시간을 요구함에도 불구하고, 앞선 연구문제에서 확인된 바와 같이 사람 채점자와의 높은 일치도를 보였다. 이는 평가의 정확성과 해석 가능성을 확보하기 위해 일정 수준의 계산 비용을 감수하는 것이 교육 현장을 고려한 평가 관점에서 합리적인 선택임을 보여준다.

결론

본 연구는 주의집중 기반 언어모델의 구조적 차이가 영어 에세이 다면평가에서 어떠한 평가 양상을 보이는지를 체계적으로 분석하였다. 구체적으로, 인코더 기반(BERT), 디코더 기반(Qwen), 인코더–디코더 기반(T5) 언어모델을 대상으로 (1) 사람 채점자와의 점수 일치도, (2) 에세이 내 내용어와 기능어에 대한 주의집중 양상, (3) 실제 교육 현장을 고려한 채점 속도 측면에서 비교 분석을 수행하였다. 이를 통해 기존 자동화 에세이 평가 연구가 주로 성능 향상에 집중해 온 것과 달리, 언어모델이 어떠한 근거와 과정으로 점수를 산출하며 그것과 사람 채점자의 유사도는 어떠한지 교육학적 관점에서 비교하며 분석하였다.

연구 결과, 인코더–디코더 기반 언어모델인 T5는 모든 평가 구인에서 사람 채점자와 가장 높은 점수 일치도를 보였으며, 내용어에 대한 주의집중 비율 또한 가장 높게 나타났다. 이는 T5가 에세이의 전반적인 의미를 먼저 파악한 뒤, 평가 구인과 관련된 언어적 단서를 선택적으로 참조하는 사람 채점자의 분석적 평가 과정을 비교적 충실하게 모사하고 있음을 시사한다. 반면, 인코더 기반 모델인 BERT는 텍스트 이해에는 강점을 보였으나, 구인별 판단으로 연결되는 과정이 구조적으로 제한적이었으며, 디코더 기반 모델인 Qwen은 생성 중심의 특성으로 인해 중간 수준의 일치도를 보였다. 또한, 내용어와 기능어에 대한 주의집중 분석 결과, 사람 채점자와 높은 일치도를 보이는 모델일수록 의미적 요소를 대표하는 내용어에 상대적으로 더 많은 주의를 기울이는 경향이 확인되었다. 이는 사람 채점자가 영어 에세이를 평가할 때, 의미 중심 판단이 형식 중심 판단보다 우선적으로 작동한다는 선행연구의 논의와 일치하는 결과이다(Lumley, 2005; Setyowati et al., 2020; Winke & Lim, 2015). 그러나, 내용어 및 기능어 집중 비율의 모델 간 통계적 차이에 대한 효과 크기가 작은 수준임을 고려하였을 때, 본 연구의 결과는 모델의 채점 과정을 부분적으로 설명하는 것이며, 다른 요소들에 대한 추가적인 연구가 필요하다. 채점 실용성 측면에서는, 인코더 기반 모델이 가장 빠른 채점 속도를 보였으나, 인코더–디코더 기반 모델 역시 교육 현장에서 허용 가능한 수준의 처리 속도(에세이당 CPU 기준 0.792초)를 유지하면서도 가장 높은 평가 정확도(평균 0.631)를 달성하였다. 이는 분석적 에세이 평가와 같이 복합적인 판단이 요구되는 과업에서는, 일정 수준의 계산 비용을 감수하더라도 입력 텍스트 이해와 출력 텍스트 생성을 균형 있게 처리하는 인코더-디코더 구조가 교육적으로 더 합리적인 선택이 될 수 있음을 시사한다.

그럼에도 불구하고, 본 연구는 몇 가지 한계를 지닌다. 첫째, 연구에서 언어모델 학습과 성능 평가에 활용된 에세이는 미국 학생이 작성한 에세이로, 국내 EFL 학습자의 쓰기 특성을 직접적으로 반영하지는 못한다. 따라서 본 연구 결과를 국내 교육 맥락으로 일반화하기 위해서는, 향후 국내 EFL 환경에서 분석적 채점이 포함된 대규모 에세이 데이터를 구축하고 이를 활용한 후속 연구가 필요하다. 둘째, 본 연구는 세 가지 대표적인 언어모델 구조를 비교 대상으로 선정하였으나, 동일한 구조 내에서도 모델 크기나 사전학습 데이터의 차이에 따라 평가 양상이 달라질 가능성을 충분히 탐색하지는 못하였다. 셋째, 주의집중 점수는 모델의 내부 판단을 간접적으로 해석할 수 있는 단서이지만, 이것이 곧바로 사람의 인지 과정을 그대로 반영한다고 결론짓기에는 해석상의 주의가 필요하다.

이러한 한계에도 불구하고, 본 연구는 몇 가지 중요한 시사점을 제공한다. 첫째, 이론적 측면에서 자동화 에세이 평가 연구에서 언어모델의 성능 비교를 넘어, 각 모델 자체의 구조와 주의집중 방식이 평가 과정에 어떠한 영향을 미치는지를 선행 연구에서 확인된 사람 채점자의 평가 양상과 비교하여 분석함으로써 언어모델 기반 평가에 대한 교육학적 설명 가능성을 확장하였다. 이는 에세이 평가 모델의 설명력 관련 선행연구(Chu et al., 2025; Do et al., 2026; Tang, 2026)가 에세이 내 단어와 구인별 점수의 연결성 분석에 주로 집중되었고, 평가 모델 자체의 채점 과정보다

복수의 LLM 혹은 내부 모듈간 결합을 통한 채점 과정에 대한 설명력에 초점을 두어 모델 아키텍처에 여전히 존재하는 블랙박스의 한계를 일부 보완하는 사례가 될 수 있다. 둘째, 방법론적 측면에서 내용어와 기능어에 대한 주의집중 분석을 통해 의미적 요소와 형식적 요소를 구분하여 모델의 평가 전략을 해석할 수 있는 분석 틀을 제안함으로써 향후 영어교육에 언어모델을 활용하는 연구에서 모델의 내부 작동 과정을 영어교육학적 시각으로 분석할 수 있도록 하였다. 마지막으로, 실천적 측면에서 국내 교육 현장에서 활용 가능한 자동화 에세이 평가 도구를 설계할 때, 무조건적으로 대규모 모델을 도입하기보다 평가 목적과 가용 자원을 고려하고, 에세이 평가 과정에서 나타나는 언어모델별 특성을 고려하여 적절한 구조의 모델을 선택하는 것이 중요함을 시사한다.

종합하면, 본 연구는 주의집중 기반 언어모델이 영어 에세이를 다면적으로 평가하는 방식을 사람 채점자의 분석적 평가 과정과 연결하여 재조명함으로써, 자동화 에세이 평가가 단순한 점수 예측을 넘어 교육적으로 해석 가능하고 현장 친화적인 평가 도구로 발전하기 위한 기초적 근거를 제시한다. 향후 연구에서는 국내 EFL 학습자 데이터를 기반으로 한 검증, 평가 구인별 피드백 생성과의 연계, 그리고 교사와 학습자의 실제 활용 경험을 반영한 연구가 이어진다면, AI 기반 영어 쓰기 평가의 교육적 활용 가능성은 더욱 확장될 수 있을 것이다.

References

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests (Vol. 1)*. Oxford University Press.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *arXiv*. https://arxiv.org/abs/1409.0473

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

British Council. (n.d.) Content words. Retrieved December 5th, 2025, from https://www.teachingenglish.org.uk/professional-development/teachers/teaching-knowledge-database/c/content-words

Chu, Seongyeub, Kim, Jongwoo, Wong, B., & Yi, Munyong (2025). Rationale behind essay scores: Enhancing S-LLM's multi-trait essay scoring with rationale generated by LLMs. *Proceedings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, NM*, 5796–5814. https://doi.org/10.18653/v1/2025.findings-naacl.322

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT Look at? An analysis of BERT's attention. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Florence*, 276–286. https://doi.org/10.18653/v1/W19-4828

Crossley, S. A, Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser., U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing*, *54*. https://doi.org/10.1016/j.asw.2022.100667

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, B., Picou, A., & Boser, U. (2023). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research*, *9*(2), 248–269. https://doi.org/10.1075/ijlcr.22026.cro

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Do, Heejin, Kim, Yunsu, & Lee, Gary. (2024a). Autoregressive score generation for multi-trait essay scoring. *Proceedings of the Association for Computational Linguistics: EACL 2024, St. Julian's*, 1659–1666. https://doi.org/10.18653/v1/2024.findings-eacl.115

Do, Heejin, Ryu, Sangwon, & Lee, Gary. (2024b). Autoregressive multi-trait essay scoring via reinforcement learning with scoring-aware multiple rewards. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL*, 16427–16438. https://doi.org/10.18653/v1/2024.emnlp-main.917

Do, Heejin, Ryu, Sangwon, & Lee, Gary. (2026). Teach-to-reason with scoring: Self-explainable rationale-driven multi-trait essay scoring. *Expert Systems with Applications*, 132119. https://doi.org/10.1016/j.eswa.2026.132119

Dong, F., & Zhang, Y. (2016). Automatic features for essay scoring– An empirical study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX*, 1072–1077. https://doi.org/10.18653/v1/D16-1115

Freedman, S. W., & Calfee, R. C. (1983). Holistic assessment of writing: Experimental design and cognitive theory. In L. T. P. Mosenthal (Ed.), *Research on writing: Principles and methods* (pp. 75–98). Longman.

Fu, Z., Lam, W., Yu, Q., So, A. M. C., Hu, S., Liu, Z., & Collier, N. (2023). Decoder-only or encoder-decoder? interpreting language model as a regularized encoder-decoder. *arXiv.* https://arxiv.org/abs/2304.04052

Fulcher, G. (2024). *Practical language testing*. Routledge. https://doi.org/10.4324/9781003373629

Hamner, B., Morgan, J., lynnvandev, Shermis, M., & Vander Ark, T. (2012). The Hewlett Foundation: Automated essay scoring. https://kaggle.com/competitions/asap-aes

Hamp-Lyons, L. (2003). Writing teachers as assessors. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 162–189). Cambridge University Press. https://doi.org/10.1017/CBO9781139524810.012

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: Issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice*, *20*(3), 281–307. https://doi.org/10.1080/0969594X.2012.742422

Ibnian, S. S. (2011). Brainstorming and essay writing in EFL class. *Theory and Practice in Language Studies*, *1*(3), 263–272. http://doi.org/10.4304/tpls.1.3.263-272

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv.* https://arxiv.org/abs/2001.08361

Kumar, R., Mathias, S., Saha, S., & Bhattacharyya, P. (2022). Many hands make light work: Using essay traits to automatically score essays. *Proceedings of the 2022 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA*, 1485–1495. https://doi.org/10.18653/v1/2022.naacl-main.106

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

Lee, S., Cai, Y., Meng, D., Wang, Z., & Wu, Y. (2024). Unleashing large language models' proficiency in zero-shot essay scoring. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL*, 181–198. https://doi.org/10.18653/v1/2024.findings-emnlp.10

Li, S., & Ng, V. (2025). Graph-based multi-trait essay scoring. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou*, 33325–33351. https://doi.org/10.18653/v1/2025.emnlp-main.1691

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *arXiv.* https://arxiv.org/abs/1907.11692

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing*, *19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang. https://doi.org/10.1016/J.ASW.2008.02.005

Mathias, S., & Bhattacharyya, P. (2020). Can neural networks automatically score essay traits? *Proceedings of the 15th workshop on innovative use of NLP for Building Educational Applications, Seattle, WA*, 85–91. https://doi.org/10.18653/v1/2020.bea-1.8

Misgna, H., On, Byung-Won, Lee, Ingyu, & Choi, Gyu Sang. (2024). A survey on deep learning-based automated essay scoring and feedback generation. *Artificial Intelligence Review*, *58*(2), 36. https://doi.org/10.1007/s10462-024-11017-5

Nation, I. S. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

OpenAI. (2023). *OpenAI: Gpt-4 technical report*. *arXiv*. https://arxiv.org/abs/2303.08774

OpenAI. (2022). *OpenAI: Introducing ChatGPT*. https://openai.com/index/chatgpt/

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, *47*(5), 238–243.

Qorib, M. R., Moon, G., & Ng, H. T. (2024). Are decoder-only language models better than encoder-only language models in understanding word meaning? *Proceedings of the Association for Computational Linguistics: ACL 2024, Bangkok*, 16339–16347. https://doi.org/10.18653/v1/2024.findings-acl.967

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P.J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*(140), 1–67. https://dl.acm.org/doi/10.5555/3455716.3455856

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Setyowati, L., Sukmawan, S., & El-Sulukiyyah, A. A. (2020). Exploring the use of ESL composition profile for college writing in the Indonesian context. *International Journal of Language Education*, *4*(2), 171–182. https://doi.org/10.26858/ijole.v4i2.13662

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Routledge. https://doi.org/10.4324/9780203122761

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024). Exploring LLM prompting strategies for joint essay scoring and feedback generation. *Proceedings of the 2024 workshop on innovative use of NLP for Building Educational Applications, Mexico City*, 283–298. *arXiv*. https://arxiv.org/abs/2404.15845

Taghipour, K. (2017). *Robust trait-specific essay scoring using neural networks and density estimators* [Unpublished doctoral dissertation]. National University of Singapore.

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX*, 1882–1891. https://doi.org/10.18653/v1/D16-1193

Tang, X. (2026). Beyond the black box: Interpretable Multi-Trait Essay Scoring with Trait-Aware Transformer. *Electronics*, *15*(5), 1066 https://doi.org/10.3390/electronics15051066

Team, Q. (2023). *Qwen technical report*. Alibaba. *arXiv*. https://arxiv.org/abs/2309.16609

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. & Polosukhin, I. (2017). Attention is all you need. *arXiv*. https://arxiv.org/abs/1706.03762

Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the use of BERT for automated essay scoring: Joint learning of multi-scale essay representation. *Proceedings of the 2022 Conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA*, 3416–3425. https://doi.org/10.18653/v1/2022.naacl-main.249

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent abilities of large language models. *arXiv*. https://arxiv.org/abs/2206.07682

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Winke, P., & Lim, Hyojung. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, *25*, 38–54. https://doi.org/10.1016/j.asw.2015.05.002

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing*, *4*(1), 83–106. https://doi.org/10.1016/S1075-2935(97)80006-2

Yang, R., Cao, J., Wen, Z., Wu, Y., & He, X. (2020). Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. *Proceedings of the 58th annual meeting of the Association for Computational Linguistics, Seattle, WA*, 1560–1569. https://doi.org/10.18653/v1/2020.findings-emnlp.141

Yip, D. Y., & Cheung, D. (2005). Teachers' concerns on school-based assessment of practical work. *Journal of Biological Education*, *39*(4), 156–162. https://doi.org/10.1080/00219266.2005.9655989

Zhang, H., & Litman, D. (2020). Automated topical component extraction using neural network attention scores from source-based essay scoring. *Proceedings of the 58th annual meeting of the Association for Computational Linguistics, Seattle, WA*, 8569–8584. https://doi.org/10.18653/v1/2020.acl-main.759