https://dx.doi.org/10.20880/kler.2026.61.1.37

생성형 AI 기반 쓰기 피드백의 효과와 학습자 반응 분석

권태현 충북대학교 국어교육과 부교수

* 이 논문은 2025학년도 충북대학교 학술연구영역 사업의 연구비 지원에 의하여 연구되었음.

I. 머리말

쓰기 과정에서 학습자에게 제공되는 피드백은 글을 개선하고 필자의 쓰기 능력을 신장시키는 데 핵심적인 역할을 수행한다(Graham, Harris, & Hebert, 2011; 박영민·가은아·권태현 외, 2025). 그러나 실제 학교 현장에서는 교사 개인이 감당해야 하는 학생 수가 많고 수업 시간이 제한되어 있어, 학습자의 쓰기 수준과 오류 양상에 맞춘 작문 피드백 제공에 현실적 어려움이 뒤따른다(Applebee & Langer, 2011).

최근에는 이러한 문제의 대안으로 생성형 인공지능(Generative AI)의 활용 가능성에 주목하고 있다. GPT 계열의 대규모 언어모델(LLM)은 문맥 기반 분석과 문장 생성 능력을 갖춰 기존 자동 평가 시스템(AWE)이 제공하지 못했던 개별화, 즉시성, 반복 가능성을 갖춘 피드백을 제공할 수 있다는 점에서 쓰기 교육 지원 도구로서의 가능성을 보여주고 있다(Dai, Tsai, Lin, et al., 2024; Steiss, Tate, Graham, et al., 2024; Wan & Chen, 2024; 권태현, 2024). 이러한 AI 기반 피드백 도입은 교사의 업무 부담을 완화하면서도 학생에게 상시적 피드백 접근권을 보장할 수 있다는 장점이 있다(Mekheimer,

2025).

반면, AI 피드백의 질, 적절성, 신뢰성에 대한 검토는 여전히 필요한 상황이다. 선행 연구에서는 대형언어모델과 같은 생성형 AI가 응집성 및 문법, 표현 개선과 같은 표면적 측면에서는 효과적이지만(Yoon, Miszoglad, & Pierce, 2023; Guo, Pan, Li, et al., 2024), 학습자의 논리적 사고 유도나 비판적 관점 형성 등 고차적 쓰기 전략 지원에는 한계가 있다고 지적하였다(Dai et al., 2024). 더불어 학습자가 AI 피드백을 어떤 감정과 태도로 수용하는가, AI 도구에 대한 신뢰 수준은 어떠한가에 따라 실제 학습 효과가 달라질 수 있음이 보고된 바 있다(Zhan & Yan, 2025). 이뿐 아니라 교육적 피드백 맥락에서 AI를 적극적으로 활용할 경우, 데이터 프라이버시, 기술에 대한 과도한 의존 가능성 및 표절과 관련된 윤리적 고려 역시 필요하다.

이러한 논란에도 불구하고 국내에서는 아직까지 AI 피드백의 효과나 학습자 인식을 탐색하는 실증적 연구가 매우 부족한 상황이다. 특히 AI를 통해 생성된 피드백임을 인지했을 때 학생이 느끼는 신뢰도나 도구 수용 경험에 대한 질적 탐구가 중요함에도 거의 연구되지 않았다.

따라서 본 연구는 생성형 AI 기반 피드백이 고등학생의 논술형 답안에 대한 고쳐쓰기 성취에 미치는 효과를 분석하고, AI 피드백을 경험한 학생들의 피드백 수용 경험, 신뢰도, 정서적 반응 등을 심층적으로 탐색하고자 한다. 특히 이 연구에서는 문단 수준의 논술형 답안에 대한 피드백을 대상으로 연구를 진행하였다. 문단 수준의 논술형 답안은 긴 글에 비해 수정해야 하는 부분이 대체로 명확하여 피드백을 통한 글의 개선 정도를 판단하기 수월하다. 따라서 AI 피드백의 활용 가능성이 보다 분명하게 드러날 것으로 예상된다. 이를 통해 본 연구에서는 생성형 AI 피드백의 교육적 유용성과 한계를 함께 규명하고, 학교 현장의 실제적 활용 가능성에 대한 구체적 시사점을 제공하고자 한다. 이를 위해 본 연구에서 다룬 연구 문제를 정리하면 다음과 같다.

1. 피드백 유형(AI 생성 피드백과 인간 작성 피드백)이 고등학생의 논술 답안 고쳐쓰기에 어떠한 영향을 미치는가?
2. 학생들의 쓰기 수준은 고쳐쓰기 후 쓰기 결과에 어느 정도 영향을 미치는가?
3. AI 피드백을 받은 학생들은 해당 피드백의 유용성, 신뢰성, 수용 가능성을 어떻게 인식하는가?

II. 이론적 배경

최근 생성형 AI를 활용한 작문 피드백 연구가 다양한 교육적 맥락에서 활발하게 이루어지고 있다. 대규모 언어모델(LLM)을 기반으로 학습자의 글을 분석하고, 언어적 정확성뿐 아니라 내용 조직, 논리적 구성, 독자 고려와 같은 고차 담화 능력까지 지원할 수 있다는 점이 부각되면서, 생성형 AI 기반 피드백은 전통적 자동 쓰기 평가(AWE)의 한계를 극복할 수 있는 새로운 교육적 도구로 평가받고 있다. 이에 따라 대학 글쓰기, 제2언어 글쓰기, 중등교육 글쓰기 등 다양한 상황에서 생성형 AI 피드백의 효과성을 검증하고자 하는 경험·실험 연구가 빠르게 축적되고 있다(Dai et al., 2024; Escalante et al., 2023; Mizumoto & Eguchi, 2023; Stahl et al., 2024; Steiss et al., 2024; Yoon et al., 2023; Zhan & Yan, 2025; Zhang, Aubrey, Huang, et al., 2025).

이러한 연구들 중 상당수는 쓰기 평가 도구로서 생성형 AI가 제공하는 작문 피드백의 특성과 효과를 분석하였다. 대표적으로 Dai et al.,(2024)는 영어 학습자 작문에 대해 GPT-3가 제공한 피드백 유형과 효과를 분석하였다. 이 연구에 따르면 GPT 기반 피드백은 인간 피드백에 비해 구조적으로 정형화되어 있으며, 대체로 긍정적인 어조를 유지하면서 문법, 어휘, 내용의 명료성에 대한 제안을 포함하였다. 다만 내용과 같은 일부 영역에서는 교사의

피드백과 같은 높은 수준의 동의를 얻기는 어려웠다는 사실을 발견하였다. Yoon et al.(2023)는 생성형 AI인 ChatGPT가 일관성과 논리에 대한 타당한 피드백을 생성할 수 있지만, 일부 비판을 "환각"하거나, 일반적인 조언에 그치거나, 맥락적으로 미묘한 쟁점을 놓칠 수 있다고 경고하였다. Mizumoto & Eguchi(2023)는 일본 대학생을 대상으로 GPT 기반의 피드백과 인간 피드백이 학습자의 수정 행동에 미치는 영향을 비교 분석하였다. 연구 결과, GPT 피드백은 주로 문장 수준에서의 미세 수정(micro-level revisions)을 유도하는 반면, 인간 피드백은 더 높은 수준의 재구성(reorganization)이나 논리적 전개에 대한 재고를 유도하는 경향이 있었다. Escalante et al.(2023)도 GPT-4 기반의 작문 피드백을 제공하는 실험 연구를 수행하였다. 연구 결과, AI가 생성한 작문 피드백을 받은 학생들은 학기 말의 학습 성과에서 인간 피드백을 받은 학생들과 차이가 없었으며 학습자의 선호도 역시 고르게 나타났다. AI의 경우 피드백의 명확성과 구체성이, 인간 피드백의 경우 후속 질문을 할 수 있는 능력과 높은 수준의 상호작용이 장점으로 꼽혔다. 이러한 연구들은 공통적으로 AI 피드백이 구체적이지만 부분적이고 기술 중심적인 경향이 있는 반면, 인간 피드백은 보다 포괄적이고 의미 중심적인 방향으로 작문 개선을 유도한다는 점을 확인하였다.

AI 기반 피드백의 질과 관련하여 보다 주목할 만한 연구로는 Steiss et al.(2024)와 Stahl et al.(2024)을 들 수 있다. Steiss et al.(2024)에서는 중등학생이 작성한 역사 에세이를 대상으로, 인간 평가자와 ChatGPT가 제공하는 피드백의 품질을 다섯 가지 기준(기준 기반 피드백, 개선 방향의 명확성, 정확성, 우선순위 설정, 지원적인 어조)으로 비교하였다. 분석 결과, 인간 평가자는 기준 기반 피드백을 제외한 모든 측면에서 AI보다 우수한 평가를 받았으며, 특히 명확성과 정확성, 학습자 친화적인 어조에서 강점을 보였다. 반면 ChatGPT는 기준에 충실하고 일관된 피드백 제공 측면에서 높은 점수를 받았으며, 형식적 피드백을 빠르게 생성할 수 있다는 장점이 확인되었다.

반면 Stahl et al.(2024)는 오픈소스 LLM(Mistral 7B)을 활용하여 자동

채점과 피드백 생성을 통합하는 프롬프트 전략을 제안하였다. 이 연구에서는 제로샷(zero-shot)과 퓨샷(few-shot) 프롬프트 방식을 통해 채점과 피드백을 동시에 생성하는 과정을 실험하였으며, 채점의 일관성과 성능은 일정 수준 확보되었지만, 생성된 피드백은 구체성이 높음에도 불구하고 학습자의 인지적 부담을 높이는 과도한 정보 제공 또는 산만한 제안이 포함되는 경향이 있음을 지적하였다. 이는 LLM 기반 피드백이 정제 및 맥락화의 과정을 필요로 함을 보여준다.

그러나 피드백의 효과는 피드백 그 자체의 품질만으로 결정되지 않으며, 학습자가 해당 피드백을 어떻게 해석하고 수용하는지에 따라 실제 학습 결과가 크게 달라진다(Evans, 2013). 사회 인지 학습이론(Bandura, 1997)에 따르면, 학습자는 피드백을 적용하는 과정에서 자신의 능력에 대한 신념, 해당 도구가 실질적 향상으로 이어질 것이라는 기대, 피드백 수용 시 경험하는 감정적 반응과 같은 심리적 요인에 영향을 받는다. 이러한 요인들이 긍정적으로 작동할 때, 학습자는 피드백을 단순히 수정하라는 지시가 아니라 글의 향상을 위한 정보로 이해하고 능동적으로 활용하게 되며 결과적으로 더 큰 성취를 경험하게 된다(Zimmerman & Schunk, 2011). 특히 생성형 AI 기반 피드백에서는 피드백에 대한 신뢰 정도가 중요하다. 학습자가 AI가 제공하는 정보의 정확성과 유용성을 신뢰하지 못할 경우, 동일한 피드백을 받더라도 이를 무시하거나 피상적으로 적용하는 경향을 보인다(Ranalli, 2021). 반면 피드백의 타당성을 인정하고 자신의 수행에 실질적 도움을 준다고 판단할수록, 학습자는 더 깊이 있는 수정 행동을 수행하며, 이는 향상된 쓰기 능력으로 이어진다. 최근 연구에서는 이 과정에서 피드백 리터러시(feedback literacy)가 중요한 매개 변인으로 작동함을 확인하고 있다(Zhan & Yan, 2025). 즉, 학습자가 피드백을 인지적으로 해석하고 정서적으로 수용하며, 자신의 글에 적절히 적용할 수 있는 능력이 뒷받침되어야 생성형 AI 피드백이 실제 학습 효과로 전환될 수 있다.

종합하면, 생성형 AI 기반 피드백은 언어적 정확성 수준을 넘어 담화 조

직력, 비판적 사고, 독자 인식 등 고차 담화 능력의 향상을 지원할 수 있으나, 그러한 효과는 자동적이거나 보편적이지 않다. 동일한 피드백 조건이라도 학습자의 심리적 수용과 참여 방식에 따라 학습 성과가 달라지기 때문에, 향후 연구는 단순히 피드백을 통한 글의 향상 정도뿐만 아니라 학습자가 AI 피드백을 어떻게 받아들이고 있는지에 초점을 맞출 필요가 있다.

III. 연구방법

1. 연구 설계

본 연구에서는 준실험적, 사전-사후 테스트, 피험자 간 설계를 채택하여 AI를 통해 생성한 피드백과 교사가 생성한 피드백이 고등학생의 논술 답안 고쳐쓰기에 미치는 영향을 분석하였다. 이 연구에서는 초기 논술 답안의 점수를 기준으로 학생들의 쓰기 수준을 상 수준과 하 수준으로 구분하였다. 두 독립 변수는 피드백 유형(AI 대 교사)과 학생들의 쓰기 수준이었으며, 종속 변수는 논술형 답안에 대한 분석 점수 총점이다. 또한 본 연구에서는 양적 통계를 통해 확인하기 어려운 피드백에 관한 인식을 분석하기 위해 참가 학생들을 대상으로 자유 설문을 수행하였다. 학생들은 피드백의 주체가 누구인지 모르는 상태에서 고쳐쓰기를 수행하였으며 고쳐쓰기가 끝난 후 각자 받은 피드백의 유형을 확인한 후 모둠별로 모여 교사의 피드백과 AI 피드백을 비교하는 토의 활동을 수행하였다. 그리고 이를 바탕으로 Google Forms에 AI 기반 피드백의 특성에 대해 작성하도록 하였다.

2. 연구 대상

본 연구의 대상은 세종시 K고등학교와 서울 S고등학교에 재학 중인 2학년 학생 154명이다. K고등학교 학생이 77명, S고등학교 학생이 77명이었으며, 이 중 남학생이 78명, 여학생이 76명으로 성비가 균등했다. 학생들은 학교 수행 평가의 맥락에서 논술형 문항에 대해 답안을 작성하고 이에 대해 피드백을 받아 고쳐쓰기를 수행했다. 전체 154명의 학생들 중 교사의 피드백을 받은 통제집단 학생이 75명, AI 생성 피드백을 받은 실험집단 학생이 79명이었다. 이 연구에서는 모든 학급에 교사 피드백을 받은 학생과 AI 피드백을 받은 학생이 섞이도록 구성하였는데 이는 고쳐쓰기 이후 모둠별로 서로 다른 유형의 피드백을 비교해 보면서 AI 피드백의 특성에 대해 논의하도록 하기 위함이다. 전체 학생을 대상으로 초고 및 고쳐쓰기를 수행하였으나 AI의 피드백에 대한 자유 설문은 학교 사정으로 인하여 서울 S고등학교에서만 진행하였다.

3. 쓰기 과제 및 평가 기준

본 연구에서는 고등학교 2학년 '문학' 과목 수행 평가의 맥락에서 제시문을 읽고 [조건]에 따라 논술형 답안을 작성하도록 하였다. 쓰기 과제는 '능력주의의 이중성'에 관한 글 (가)와 소설의 일부인 (나)를 제시한 후 (가)의 관점을 분석 도구로 삼아 (나)에 나타난 인물의 태도를 비판하는 것이다. 학생들은 (가)에서 능력주의 비판의 논지를 추출하고 이를 기준으로 (나)에 등장하는 인물의 언어와 태도를 가치 판단적으로 해석해야 한다. 이는 단순히 (나)의 인물이 부당하다는 주장을 하는 것이 아니라 능력주의의 논리 속에서 그 부당함을 어떻게 정당화할 수 있는지를 묻는 것으로 분석과 적용이 포함된 논증 중심의 고차 사고 과제라고 할 수 있다.

다만 본 쓰기 과제는 분량을 300~400자 내외의 한 문단 쓰기 수준으

로 제한하였다. 이는 실제 고등학교 수업 맥락에서 빈번히 활용되는 단문 논증 과제일 뿐만 아니라 피드백 유형의 효과를 분석하기에도 적절한 과제라고 판단하였다. 장문의 쓰기 과제에 대한 피드백은 글의 구조나 조직 등 거시적 조언으로 흐를 가능성이 높아 피드백의 효과를 파악하는 데 다소 어려움이 있다. 이와 달리 한 문단 정도의 쓰기 과제는 주장의 명료성, 근거 적합성, 추론의 비약 등 미시적이고 기능적인 피드백에 초점이 맞춰짐으로써 피드백 및 그에 따른 고쳐쓰기 효과를 보다 세밀하게 분석할 수 있다는 장점이 있다. 이 쓰기 과제에 대한 평가 기준 및 예시 답안은 〈표 1〉과 같다.

〈표 1〉 채점 기준 및 예시 답안

채점 기준	• 내용(4점) – (가)의 핵심 내용을 정확히 이해하고 이를 자신의 비판 근거로 적절하게 활용하였는가? – (나)의 '배운 사람들'의 말과 행동에서 비판할 점을 정확히 파악하였는가? • 구성(3점) – 주장과 근거가 자연스럽게 연결되며 전체적으로 논증 구조가 잘 구성되어 있는가? • 표현(3점) – 답안의 표현이 논증적 글쓰기에 적합하며 어법에 문제가 없는가?
예시 답안	(나)의 '배운 사람들'은 가난한 사람들을 게으르고 낙오된 존재로 여기며 함부로 판단한다. 이 발언은 자신들의 특권은 외면한 채 약자의 실패를 개인의 문제로 돌리는 무책임한 태도다. (가) 제시문에 따르면 능력주의는 겉으로는 공정한 경쟁처럼 보이지만, 실제로는 구조적 불평등 속에서 일부 계층에게만 유리하게 작동한다. 특히 교육 기회와 경제력은 부모 세대에서 자식 세대로 세습되고 있으며, 노력만으로는 이를 극복하기 어렵다. 따라서 배운 사람들의 발언은 능력주의가 지닌 맹점을 그대로 드러낸 것으로, 사회적 약자를 더욱 소외시키고 차별을 정당화하는 위험한 인식이라고 할 수 있다.

채점 기준은 내용, 구성, 표현의 세 영역으로 구분되었으며, 특히 내용 영역에서는 제시문 (가)의 핵심 논지를 비판 근거로 적절히 활용하였는지와 제시문 (나)에 나타난 인물의 태도를 정확히 해석하였는지를 중심으로 평가하도록 설계하였다. 이는 제시문의 단순 요약이나 감정적 반응을 배제하고, 제시문 간 관계 설정과 개념을 기반으로 한 비판 능력을 중심으로 논증 수행 여부를 판단하기 위함이다.

예시 답안은 채점 기준에 부합하는 상위 수준의 응답을 제시하기 위한

참고 자료로 활용되었다. 즉, 피드백은 예시 답안의 재현 여부가 아니라, 학생 답안이 제시문을 개념적으로 연결하여 논증을 구성하였는지, 주장과 근거의 관계가 논리적으로 형성되었는지에 초점을 두도록 하였다. 이상의 쓰기 과제 및 평가 기준은 연구자 및 현직 고등학교 국어 교사 2인의 검토를 통해 내용타당도를 확보하였다.

본 연구에서는 논술형 답안에 대한 채점의 신뢰도를 확보하기 위하여 교사 3인이 독립적으로 채점에 참여하였다. 사전 검사와 사후 검사 모두 동일한 채점 기준을 적용하였으며, 채점자 간 일치도를 검증하기 위해 급내상관계수(intra-class correlation coefficients; ICCs)를 산출하였다. ICC 분석은 채점자를 무작위 효과로 간주한 이원무선효과 모형(two-way random effects)을 사용하였고, 점수의 절대적 일치를 기준으로 하여 평균 측정치 ICC를 산출하였다.

그 결과, 사전 검사 답안의 ICC는 .852로, 피드백 이후 고쳐 쓴 답안을 대상으로 한 사후 검사의 ICC는 .906으로 나타났다. 사전·사후 검사 모두에서 ICC 값이 .80 이상으로 나타났다는 점에서, 본 연구에서 사용한 채점 기준은 논술 수행을 신뢰롭게 평가하는 데 적절한 도구임을 확인할 수 있다. 특히 사후 검사에서 더 높은 ICC가 산출된 것은, 수정된 답안이 논증 구조와 표현 측면에서 보다 명확해짐에 따라 채점자 간 해석의 차이가 감소한 결과로 해석할 수 있다. 이에 따라 본 연구에서는 이후의 모든 분석에서 교사 3인의 채점 점수를 평균하여 학생의 대표 점수로 활용하였다.

4. 피드백 생성

1) 인간 피드백

인간 피드백은 경력 3년 이상의 현직 국어 교사 2인이 참여하여 생성하였다. 이들은 본 연구의 쓰기 채점에는 참여하지 않았는데 이는 피드백을 작성한 교사가 채점에도 참여할 경우, 무의식적으로 자신의 피드백이 반영된

글을 관대하게 채점하는 편향(bias)이 발생할 수 있을 것이라고 판단했기 때문이다. 먼저 두 명의 교사가 동일한 예비 답안에 대해 개별적으로 시범 피드백을 작성한 후, 상호 검토 및 협의를 통해 피드백의 방향성과 표현 수준, 내용 구조 등을 조정하였다. 이 과정에서 채점 루브릭에 따른 피드백 관점과 표현 지침을 공유하고 합의하였으며, 최종적으로 각 교사는 통제집단 75명의 학생 답안을 절반씩 나누어 독립적으로 피드백을 작성하였다. 이와 같은 절차는 교사 간 의견 일치를 도모함과 동시에, 각 교사의 개별성을 유지하는 데 목적이 있다. 한편 본 연구에서는 피드백 생성 조건의 공정성을 확보하기 위해, AI 피드백 프롬프트에 포함된 핵심 설계 원칙(지지적 어조, 핵심 중심 조언, 과도한 분량 제한 등)을 인간 피드백 작성 교사에게도 공유하였다. 다만 인간 피드백의 전문성과 자연스러운 교수적 판단을 유지하기 위하여, 구체적인 표현 방식이나 피드백 구조는 교사의 자율에 맡겼다.

2) AI 피드백

AI 피드백은 생성형 인공지능 모델을 활용하여 자동 생성하였다. 본 연구에서는 ChatGPT-4o를 사용하였으며, 인간 피드백과의 기능적 비교를 위해 프롬프트 디자인 프레임워크(prompt design framework)[1]를 적용하였다. 이는 대규모 언어모델이 교육적 피드백 과제를 충실히 달성할 수 있도록 프롬프트를 구조화하는 방안으로서 본 연구는 AI 피드백 관련 선행 연구를 참조하여 목표 명료화, 문맥 제공, 응답 제약 설정, 교육적 피드백 원칙 적용과 같은 설계 요소를 구조화하였다(Jacobsen et al., 2025). 다음은 프롬프트

1) 프롬프트 디자인 프레임워크는 생성형 AI(LLM)에게 원하는 응답을 얻기 위해 프롬프트를 구조화·설계하는 체계적 접근법을 의미한다. 단순히 질문하는 문장이 아니라, 목표·맥락·지침·제약 조건 등을 명확히 포함하여 AI가 일관성 있고 높은 품질의 출력을 생성하도록 유도하는 설계 원칙의 집합이다. 예를 들어 CLEAR 프레임워크는 정보 리터러시와 생성형 AI 시대의 소통 능력을 향상시키기 위해 개발된 프롬프트 설계 가이드로서, Concise(간결성), Logical(논리성), Explicit(명시성), Adaptive(적응성), Reflective(성찰성)의 다섯 요소로 구성된다(Lo, 2023).

설계 원리에 따라 본 연구에서 개발한 프롬프트 전문이다.

〔역할(Role)〕
당신은 고등학생의 논술형 글쓰기를 지도하는 국어 교사이다.

〔목표(Goal)〕
아래에 제시된 학생 글을 분석하여, 학생이 스스로 글을 고쳐 쓸 수 있도록 돕는 교육적이고 실행가능한 종합 피드백을 작성하라.

〔맥락(Context)〕
본 피드백은 사후 수정 활동을 위한 것이다. 다음에 제시되는 쓰기 과제, 루브릭, 모범 답안은 학생 글의 수준을 판단하고 개선 방향을 설정하기 위한 참고 자료이다. 모범 답안을 정답으로 삼아 학생 글을 비교·판단하지 말고, 학생 글 자체의 논증 수행을 중심으로 평가하라.

〔입력 자료(Input)〕

〔쓰기 과제〕, 〔루브릭〕, 〔모범 답안〕, 〔학생 답안〕

〔피드백 작성 지침(Instructions)〕
- 핵심 중심 지적: 학생 답안에서 발견되는 모든 문제를 나열하지 말고, 논증의 타당성과 완성도에 실질적인 영향을 미치는 핵심 문제점을 중심으로 제시하라.
- 지지적·정중한 톤: 교사의 피드백으로서 지지적이고 정중한 표현을 사용하며, 학생의 노력을 인정하는 격려를 반드시 포함하라.
- 긍정적 시작 후 개선 제안: 먼저 학생 글의 강점이나 적절한 시도를 구체적으로 언급하라. 이후 내용, 구성, 표현 중에서 루브릭에 비추어 개선이 필요한 핵심 요소만을 구체적으로 제안하라.
- 실행 가능성: 학생이 바로 고쳐 쓸 수 있도록, 추상적 평가가 아닌 수정 방향 중심의 조언을 제시하라.

〔출력 형식(Output)〕
- 하나의 연속된 단락으로 작성하되, 교사 피드백 문체를 유지하라.
- 전체 분량은 500자 이내로 제한한다.

피드백 사전 지침은 일관된 평가 관점과 학생 중심의 피드백 질을 동시에 확보하기 위함이다. 특히 긍정적 시작, 핵심 개선 사항 중심 제시, 500자 제한 등의 요소는 과잉 피드을 방지하고, 학습자가 실제 고쳐쓰기 과제에 적용 가능한 구체적이고 실행 가능한 조언을 제공하는 데 초점을 두었다. 이러한 설계는 단순 오류 지적을 넘어 학습자의 자기 수정(self-revision) 능력 증진을 목표로 하며, 교사와 AI 간 피드백 차이를 비교 평가하기 위한 공통의 기준점(anchor)으로 기능할 수 있도록 하였다.

실제 피드백 생성 시 예시 답안은 인간 교사와 AI 모두에게 참고용 자료

로 제공되었다. 이는 피드백의 톤과 방향, 논증 평가 기준이 일정 수준 이상 유지되도록 유도하기 위한 것으로, 예시 답안이 정답 텍스트로 기능하지 않도록 유의하였다. AI 프롬프트는 동일한 지침을 유지한 상태에서 각 학생 답안을 입력하여 개별 피드백이 생성되었으며, 생성된 텍스트는 연구자가 사전 정의된 지침과 일치하는지를 추가 검토하였다. 필요 시 문법적 오류나 불필요한 표현만을 수정하였으나, 피드백의 내용적 의미나 조언의 방향은 원문을 유지하였다.

5. 분석 방법

본 연구에서는 인간 평가자와 ChatGPT에 의해 산출된 쓰기 피드백의 효과를 분석하기 위해 집단별 답안 점수의 평균, 표준편차 등 기술적 통계를 산출하였으며 정규성 및 등분산 검정, 대응표본 t-검정을 사용하여 두 피드백 유형에서 사후 검사와 사전 검사 점수의 차이를 확인하였다. 또한 쓰기 수준의 주요 효과와 피드백 개입 후 쓰기 성과에 대한 잠재적 상호 작용을 동시에 검토하기 위해 이원분산분석(ANOVA)을 수행하였다. 본 연구에서는 단순한 통계적 유의성 검증을 넘어, 피드백이 실제로 학습자에게 어느 정도 의미 있는 변화를 가져왔는지 판단하기 위해 모든 추론 통계 분석에 효과크기(effect size)를 함께 산출하였다. 효과크기 척도로는 독립표본 t-검증에서는 cohen's d, ANOVA에서는 부분 에타제곱(η^2)을 활용하였다. 이상의 통계 분석에는 Python 3.14.1의 scikit-learn 라이브러리를 활용하였다.

또한 본 연구에서는 AI 기반 피드백에 대한 참여자의 자유 응답을 분석하기 위해 주제 분석(Thematic Analysis)과 내용 분석(Content Analysis)을 통한 질적 분석을 수행하였다. 이는 참여자의 피드백 인식을 보다 심층적으로 이해하고, 정량적 결과를 보완하기 위한 목적에서 수행되었다(Braun & Clarke, 2006). 질적 분석은 자유 응답에 참여한 서울 S고등학교 학생 97명의 응답을 활용하였으며 총 637개의 문장 단위의 응답이 수집되었다. 본 연

구에서는 자료 익숙화, 초기 코딩, 주제 생성, 내용 분석의 과정을 거쳐 질적 분석을 수행하였다. 모든 질적 분석은 연구자 외의 작문 교육 박사 이상의 전문가 2인과의 교차 검토를 통해 신뢰도를 확보하였으며, 데이터 정리 및 주제별 코딩, 빈도 분석에는 Python 3.14.1을 활용하였다.

IV. 연구 결과

1. 생성형 AI 피드백을 활용한 고쳐쓰기 결과

1) 기술 통계

〈표 2〉는 피드백 유형(AI 피드백, 인간 피드백)과 학습자의 초기 쓰기 수준(상, 하)에 따른 사전·사후 쓰기 점수 및 변화량에 대한 기술 통계 결과를 제시한 것이다.

〈표 2〉 집단(피드백 유형 및 쓰기 수준)별 기술 통계

피드백 유형	쓰기 수준	N	사전 M(SD)	사후 M(SD)	변화량 M(SD)
AI	상	44	8.54(0.65)	9.00(1.03)	0.46(1.05)
	하	33	7.07(0.61)	8.44(0.98)	1.37(1.16)
인간	상	40	8.90(0.80)	9.10(1.04)	0.20(1.13)
	하	37	6.45(1.14)	7.53(1.83)	1.08(1.51)

우선 전체 집단 차원에서 볼 때, AI 피드백 집단과 인간 피드백 집단 모두에서 사후 점수가 사전 점수에 비해 상승하는 경향이 관찰되었다. AI 피드백 집단의 사전 점수 평균은 7.91(SD=0.97), 사후 점수 평균은 8.76(SD=1.04)로 나타났으며, 평균 변화량은 0.85(SD=1.18)였다. 인

간 피드백 집단 역시 사전 점수 평균 7.72(SD=1.57)에서 사후 점수 평균 8.34(SD=1.66)로 증가하였고, 평균 변화량은 0.62(SD=1.39)로 나타났다. 이러한 기술 통계 결과는 두 피드백 유형 모두에서 쓰기 성과 향상이 관찰되었음을 시사한다.

쓰기 수준에 따른 분포를 살펴보면, 하 수준 집단의 평균 변화량이 AI 피드백 집단에서 1.37(SD=1.16), 인간 피드백 집단에서 1.08(SD=1.51)로 나타나, 상 수준 집단의 변화량(AI: 0.46, 인간: 0.20)에 비해 상대적으로 큰 값을 보였다. 이는 초기 쓰기 수준이 낮은 학습자일수록 피드백 개입 이후 점수 향상이 더 크게 나타나는 경향이 있음을 시사한다.

2) 집단 간 평균 비교

분석에 앞서 각 집단 점수의 분포가 정규성을 충족하는지 확인하였다. Shapiro－Wilk 검정 결과 정규성 가정을 크게 위반하지 않는 것으로 나타났다($p>.05$). 또한 Levene 검정을 실시한 결과 집단 간 분산의 동질성도 충족되는 것으로 확인되었다($p>.05$). 이에 따라 집단 간 평균 검정을 수행하였다. 피드백 유형별 사전－사후 쓰기 점수의 변화를 검증하기 위해 대응표본 t－검증을 실시한 결과는 〈표 3〉과 같다.

〈표 3〉 집단(피드백 유형) 간 점수 차이

피드백 유형	N	사전 M(SD)	사후 M(SD)	t(df)	p	Cohen's d
AI	77	7.91(0.97)	8.76(1.04)	6.32(76)	<.001	0.72
인간	77	7.72(1.57)	8.34(1.66)	3.93(76)	<.001	0.45

AI 피드백 집단과 인간 피드백 집단 모두에서 사전 점수 대비 사후 점수가 유의하게 향상된 것으로 나타났다. AI 피드백 집단의 경우 사후 점수(M=8.76, SD=1.04)가 사전 점수(M=7.91, SD=0.97)에 비해 유의하게 높았으며($t(76)=6.32$, $p<.001$), 효과크기 또한 중간에서 큰 수준($d=0.72$)이었다.

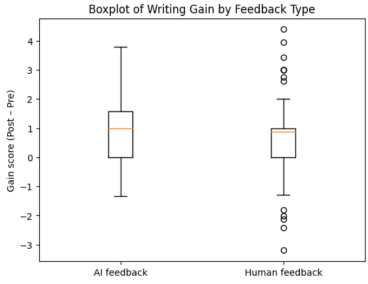
인간 피드백 집단에서도 사후 점수(M=8.34, SD=1.66)가 사전 점수(M=7.72, SD=1.57)에 비해 유의한 향상을 보였고($t(76)=3.93$, $p<.001$), 효과크기는 중간 수준(d=0.45)이었다. 이러한 결과는 두 피드백 유형 모두 학습자의 쓰기 성과 향상에 통계적으로 유의미한 기여를 했으나 그 효과는 AI 피드백이 인간 피드백에 비해 미세하게 높게 나타났다. 그러나 이것이 곧바로 AI 피드백의 우월성을 의미하지는 않는다. AI 피드백은 즉각성, 분량, 반복 가능성 측면에서 피드백 노출량이 많았을 가능성이 있으며, 이러한 구조적 특성이 효과크기에 반영되었을 가능성이 있다. 두 집단 모두에서 사전–사후 유의한 향상이 나타났다는 점은, 피드백의 효과가 제공 주체의 '인간성' 자체보다는 피드백이 제공하는 정보의 구조화, 시의성, 수정 가능성에 의해 설명될 수 있음을 시사한다. 다만 AI 피드백 집단에서 중간 이상의 효과크기(d=0.72)가 관찰되었다는 점은, 생성형 AI가 단순한 정보 제공 도구를 넘어 학습자의 실제 쓰기 수행을 개선하는 교육적 개입으로 기능할 수 있음을 시사한다. 아래 〈그림 1〉은 피드백 유형(AI 피드백, 인간 피드백)에 따른 총점 변화량(Post – Pre)의 분포를 박스플롯으로 제시한 것이다. 인간 피드백 집단은 AI 피드백 집단에 비해 사분위 범위가 넓고 이상치의 분포가 더 크게 나타나, 학습자 간 점수 변화의 변동성이 상대적으로 큰 양상을 보였다. 반면 AI 피드백 집단은 변화량이 비교적 좁은 범위에 분포하여, 학습자 반응이 보다 균질하게 나타나는 경향을 보였다. 이러한 분포상의 차이는 AI 피드백과 인간 피드백의 작동 방식 차이를 반영하는 결과로 해석될 수 있다. AI 피드백은 동일한 기준과 구조에 따라 피드백을 제공함으로써 학습자 전반에 비교적 일관된 개선 효과를 유도한 반면, 인간 피드백은 개별 학습자의 특성이나 답안의 질에 따라 효과가 크게 달라질 수 있음을 시사한다. 이는 선행연구에서 지적된 생성형 AI 피드백의 체계성과 일관성이라는 특질과 연결되며, 본 연구의 질적 분석에서 확인된 '전형적이지만 안정적인 피드백'이라는 학습자 인식과도 맥을 같이한다.

추가로 본 연구에서는 내용, 구성, 표현 세부 영역의 점수 차이를 함께

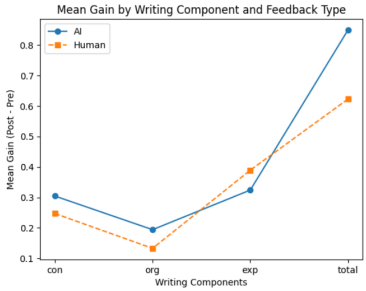
분석하였다. 〈표 4〉와 〈그림 2〉는 세부 영역별 점수 차이를 나타낸 것이다.

〈표 4〉 집단(피드백 유형 및 쓰기 수준)간 점수 차이

영역	사전M(SD)	사후M(SD)	t	p
내용	3.05 (0.71)	3.33 (0.61)	5.19	<.001
구성	2.40 (0.55)	2.57 (0.50)	3.44	.001
표현	2.29 (0.52)	2.65 (0.48)	8.18	<.001
총점	7.82 (1.30)	8.55 (1.40)	7.08	<.001



〈그림 1〉 피드백 유형별 점수 변화



〈그림 2〉 쓰기 영역별 점수 변화

〈표 3〉과 〈그림 2〉는 피드백 유형에 따라 내용, 구성, 표현의 세부 영역과 및 총점에서 나타난 사전–사후 변화량의 평균을 제시한 것이다. 분석 결과, 두 피드백 유형 모두 모든 하위 영역에서 점수 향상이 관찰되었으나, 영역별 변화 양상에는 차이가 나타났다. AI 피드백 집단은 총점 변화량이 가장 크게 나타났으며, 특히 내용과 표현 영역에서 비교적 고른 향상을 보였다. 반면 인간 피드백 집단은 표현 영역에서 가장 큰 변화량을 보여, 문체나 표현의 세밀한 조정과 관련된 영역에서 상대적으로 높은 효과를 나타냈다. 구성 영역은 두 집단 모두에서 변화량이 가장 작게 나타났는데, 이는 텍스트 구조나 논증 조직과 같은 고차원적 쓰기 기능이 단기적 피드백만으로는 개선되는 데 한계가 있음을 시사한다. 전반적으로 두 피드백 유형은 하위 영역 전

반에서 유사한 변화 패턴을 보이면서도, 강조되는 쓰기 요소에는 차이를 보임을 확인할 수 있다.

3) 이원분산분석

〈표 5〉는 피드백 유형(AI 피드백, 인간 피드백)과 학습자의 초기 쓰기 수준(상, 하)이 쓰기 점수의 변화량에 미치는 영향을 검토하기 위해 실시한 이원분산분석(ANOVA) 결과를 제시한 것이다.

〈표 5〉 이원분산분석 결과

요인	SS	df	F	p	partial η^2
피드백 유형	2.87	1	1.94	.166	.013
쓰기 수준	30.70	1	20.73	<.001	.121
상호작용	0.00	1	0.00	.959	.000

분석 결과, 학습자의 초기 쓰기 수준에 따른 주효과가 통계적으로 유의하게 나타났다($F(1,150)=20.73$, $p<.001$, partial $\eta^2=.121$). 이는 사전 점수 기준 하 수준 학습자가 상 수준 학습자에 비해 피드백 개입 이후 더 큰 쓰기 성과 향상을 보였음을 의미한다. 반면 피드백 유형의 주효과는 통계적으로 유의하지 않았다($F(1,150)=1.94$, $p=.166$, partial $\eta^2=.013$). 이는 본 연구에서 사용된 두 피드백 유형 간에 쓰기 성과 변화량의 차이가 통계적으로 확인되지 않았음을 나타낸다. 또한 피드백 유형과 쓰기 수준 간의 상호작용 효과 역시 유의하지 않은 것으로 나타났다($F(1,150)=0.00$, $p=.959$, partial $\eta^2 \approx .000$). 이는 학습자의 초기 쓰기 수준에 따라 피드백 유형의 효과가 다르게 나타났다고 볼 근거가 없음을 시사한다.

종합하면, 이원분산분석 결과는 피드백 개입 이후의 쓰기 성과 변화가 피드백 제공 주체의 유형보다는 학습자의 초기 쓰기 수준에 의해 더 크게 설명됨을 보여준다. 이러한 결과는 이후 논의에서 피드백 효과의 해석을 학습

자 특성 요인과 연계하여 검토할 필요성을 시사한다.

2. 학습자 반응 분석

AI 피드백에 대한 학생들의 경험과 인식에 대한 더 깊은 통찰을 얻기 위해, S고등학교 학생 97명을 대상으로 반구조화 인터뷰를 수행한 후 관련 데이터를 주제별 분석으로 수행하였다. 분석 결과, AI 쓰기 피드백에 대한 특질과 관련하여 5가지 주요 주제가 드러났다. 이는 (1) 구체성과 실현 가능성, (2) 체계성과 일관성, (3) 표현 개선 중심, (4) 전형성, (5) 메타인지 유도의 한계이다. 〈표 6〉은 이러한 주제 및 하위 내용을 요약한 것이다.

〈표 6〉 생성형 AI 피드백에 대한 인식

주제 (Theme)	하위 코드 (Sub-code)	주요 내용 요약	대표 예시 인용문
구체성과 실행 가능성	구체적 설명	AI 피드백은 무엇이 좋은지, 어떻게 고쳐야 할지를 구체적으로 설명함.	"내용의 긍정적인 부분을 인간 피드백보다 더 자세하게 설명하며, 표현의 매끄러움에 대해 많은 예시를 들고 있다."
	수정 방향 제시	단순 평가가 아니라 실제 수정 방향을 알려준다는 점을 장점으로 인식	"AI 피드백은 학생이 글을 실질적으로 개선할 수 있도록 구체적이고 체계적인 도움을 준다."
체계성과 일관성	체계적 구조	피드백이 일정한 틀과 순서로 제시되어 이해하기 쉬움.	"피드백이 체계적으로 정리되어 있어서 읽기 쉽고 뭐부터 고쳐야 할지 알 수 있다."
	반복 기준 적용	답안이 달라도 유사한 기준으로 평가한다고 느낌	"AI 피드백은 전형적인 틀로 항상 비슷한 기준에서 말하는 것 같다."
표현 개선 중심	문장 다듬기	표현의 매끄러움과 문장 수정에 특히 도움이 된다고 인식	"표현의 매끄러움 같은 부분에서 AI 피드백이 더 도움이 됐다."
	가독성 향상	글이 더 읽기 쉬워졌다고 느낌	"문장 형식이나 표현에 대한 AI 피드백을 반영하니까 글이 정돈된 느낌이 들었다."

전형성	전형성 인식	피드백이 다소 틀에 박혀 있다고 인식	“인간 피드백은 글에 따라 다른 피드백을 제시하는 반면, AI 피드백은 대부분 유사하고 일정한 틀에 맞춰 전형적이라는 느낌이 든다.”
메타인지 유도의 한계	깊이의 한계	내용의 깊이나 독창성까지는 다루지 못한다고 느낌	“표현은 잘 봐주는데, 생각 자체를 더 깊게 만들어주진 않는 것 같다.”

우선 구체성과 실현 가능성 측면에서 AI 피드백은 학습자가 무엇을 어떻게 수정해야 하는지를 명확히 제시하는 경향을 보였는데, 이는 생성형 AI 피드백이 즉시 적용 가능한 수정 제안을 풍부하게 제공한다는 기존 연구 결과와 일치한다(Lin & Crosthwaite, 2024; Rashkin et al., 2025; Steiss et al., 2024). 분석된 피드백 전문에서는 “한 문장으로 요약한 뒤 근거를 연결하라”, “문단의 첫 문장에서 이전 내용을 정리하라”와 같이 수정의 대상과 방식이 구체적으로 제시되는 사례가 다수 확인되었다. 이러한 피드백은 학습자가 추가적인 해석이나 판단을 거치지 않고도 즉각적인 수정 행동으로 옮길 수 있다는 점에서 높은 실현 가능성을 지닌다. 이러한 특질은 AI 피드백이 글 고쳐쓰기 과정에서 행동 지향적(action-oriented) 피드백으로 기능하고 있음을 시사한다. 즉, AI 피드백은 학습자의 사고를 평가하는 데 초점을 두기보다는, 현재 작성된 텍스트를 기준으로 수정 가능한 지점을 명확히 드러내는 역할을 수행한다. 학생 인식 분석에서도 AI 피드백이 “무엇을 고쳐야 하는지 바로 알 수 있다”, “수정할 때 참고하기 쉽다”는 평가가 반복적으로 나타났으며, 이는 구체성과 실현 가능성이 학습자의 피드백 수용도를 높이는 핵심 요인으로 작용하고 있음을 보여준다.

두 번째 특질인 체계성과 일관성은 AI 피드백이 비교적 안정된 구조와 동일한 기준에 따라 생성된다는 점에 기인한다. 본 연구에서 분석한 AI 피드백은 ‘전반적인 평가–개선이 필요한 지점–수정 제안’이라는 담화 구조를 반복적으로 유지하고 있었으며, 피드백의 어조와 구성에서도 큰 변동이 관찰되지 않았다. 이러한 체계성과 일관성은 학습자에게 예측 가능하고 안정

적인 피드백 환경을 제공한다는 점에서 긍정적으로 작용한다. 그러나 동시에 이러한 특질은 피드백이 답안의 맥락이나 학습자의 개별적 특성을 세밀하게 반영하기보다는, 표준화된 기준을 우선 적용하는 방향으로 수렴될 가능성을 내포한다(Lin & Crosthwaite, 2024). 이는 AI 피드백의 강점이자 구조적 한계로 함께 이해될 수 있다.

세 번째 특질은 AI 피드백이 내용의 깊이나 논증 전략보다는 표현의 명료성, 문장 구성, 가독성 등 언어적 표층 요소의 개선에 상대적으로 집중하는 경향을 보인다는 점이다. 분석된 피드백에서는 문장의 중복, 문단 간 연결의 자연스러움 등 표현 수준의 문제를 지적하는 사례가 빈번하게 나타났다. 반면, 주장 간 논리적 관계나 대안적 논증 전략에 대한 심층적 논의는 제한적으로 제시되었다. 선행 연구에서도 생성형 AI 피드백이 언어 요소의 개선에는 효과적인 반면, 내용의 깊이나 논증 구조와 같은 고차원적 쓰기 요소에는 상대적으로 제한적 효과를 보인다고 보고해 왔다(Dai et al., 2024; Guo et al., 2024; Steiss et al., 2024; Yoon et al., 2023; Zhang et al., 2025). 본 연구의 AI 피드백 전문에서도 표현의 중복, 문단 연결, 문장의 간결성 등 표현 수준의 지적이 두드러지게 나타났으며, 학생 인식 분석에서도 표현 개선이 가장 체감되는 효과로 반복 언급되었다. 이러한 결과는 양적 분석에서 표현 영역의 변화량이 상대적으로 크게 나타난 결과를 설명하는 기제로도 해석될 수 있다.

한편, 전형성과 메타인지 유도의 한계는 선행 연구의 논의를 부분적으로 확장하는 지점이다. 우선 전형성은 AI 피드백이 학습자 개인의 고유한 사고 과정이나 글쓰기 맥락을 세밀하게 반영하기보다는, 비교적 정형화된 구조와 표현을 반복적으로 사용하는 경향을 지님을 의미한다. 이러한 전형성은 한편으로는 피드백의 이해 가능성과 적용 가능성을 높이는 요인으로 작용한다. 정형화된 구조와 반복되는 표현은 학습자가 피드백의 핵심을 빠르게 파악하고, 수정 방향을 예측 가능하게 인식하도록 돕는다. 이는 전형성이 단순한 한계라기보다는, AI 피드백의 체계성과 일관성을 뒷받침하는 구조적

특질로 기능할 수 있음을 보여준다.

그러나 동시에 이러한 전형성은 피드백의 개별화 수준을 제한하는 요인으로 작용할 수 있다. AI 피드백은 학습자의 독창적인 논증 전략이나 사고 전개의 특수성을 심층적으로 반영하기보다는, 일반적인 글쓰기 기준에 비추어 수정 방향을 제시하는 경향을 보였다. 이로 인해 일부 학습자는 AI 피드백이 자신의 글을 "개인적으로 평가받는 느낌보다는, 모범 답안에 맞추어 점검받는 느낌"에 가깝다고 인식하기도 하였다. 이는 AI 피드백을 글쓰기 지도에서 보편적 수정과 1차 점검을 위한 도구로 활용하되, 보다 개별화된 사고 확장과 전략적 지도는 교사 피드백과의 결합을 통해 보완할 필요가 있음을 시사한다.

생성형 AI 피드백의 마지막 특질은 '메타인지 유도의 한계'로 나타났다. 기존 연구에서는 AI 피드백이 질문이나 점검 문장을 포함하더라도 학습자의 사고를 장기적으로 확장하거나 논증 전략을 재구성하도록 유도하는 데에는 한계가 있다고 지적해 왔다(Yoon et al., 2023). 본 연구에서도 AI 피드백은 "어떻게 고칠 것인가"에 대한 직접적인 지시를 중심으로 구성되어 있었으며, 학습자의 선택이나 판단을 요구하는 질문형 피드백은 제한적으로 나타났다. 이에 대해 학생들 역시 AI 피드백이 글을 '더 잘 쓰게' 도와주기는 하지만, 자신의 생각을 근본적으로 더 깊게 만들어주지는 않는다고 인식하고 있었다. 이는 AI 피드백이 메타인지적 성찰보다는 즉각적 수정에 초점을 둔 도구로 기능하고 있음을 학습자 관점에서 재확인한 결과라 할 수 있다.

이상의 분석 결과, AI 피드백은 구체성과 실현 가능성, 체계성과 일관성, 표현 개선 중심성이라는 강점을 지니는 반면, 전형성과 메타인지 유도의 한계를 함께 내포하는 피드백으로 나타났다. 이는 선행연구에서 제시된 생성형 AI 피드백의 장점과 한계를 대부분 재현하면서도, 이러한 특질이 실제 학습자 경험 속에서 어떻게 인식되고 활용되는지를 보다 구체적으로 보여준다는 점에서 의의를 지닌다. 상기 분석 결과는 AI 피드백이 글쓰기 지도에서 효과적인 1차 수정 도구로 활용될 수 있음을 시사하는 동시에, 심층적 사고 확

장과 개별화된 지도는 인간 교사의 개입을 통해 보완될 필요가 있음을 보여준다.

V. 맺음말

본 연구는 생성형 AI가 학교 현장에서 쓰기 피드백 도구로 활용될 수 있는 가능성과 한계를 실증적으로 검토하고자, 인간 평가자와 ChatGPT에 의해 산출된 피드백이 고등학생의 논술 답안 고쳐쓰기 성과에 미치는 영향을 비교·분석하였다. 특히 본 연구는 한 문단 분량의 논술형 답안을 연구 대상으로 설정함으로써, 피드백 개입의 효과를 보다 명확하게 포착하고자 하였다.

양적 연구 결과, 인간 피드백과 AI 피드백을 받은 집단 모두에서 사전–사후 쓰기 점수가 유의미하게 향상된 것으로 나타나, 두 피드백 유형 모두 학습자의 글쓰기 성과 향상에 기여함을 확인할 수 있었다. 대응표본 t-검증 결과에서는 AI 피드백 집단의 효과 크기가 인간 피드백 집단에 비해 미세하게 높게 나타났으나, 그 차이는 제한적인 수준이었다. 이는 짧은 논술형 답안을 대상으로 한 본 연구의 맥락에서, 인간 피드백과 AI 피드백 간의 성과 차이가 크게 두드러지지 않았음을 의미한다. 즉, 제한된 분량의 논술 답안에서는 AI 피드백이 제공하는 구체적이고 즉각적인 수정 제안만으로도 인간 피드백과 유사한 수준의 학습 성과를 도출할 수 있음을 시사한다.

이원분산분석 결과에서는 쓰기 수준의 주효과만이 통계적으로 유의하게 나타났으며, 피드백 유형의 주효과와 피드백 유형과 쓰기 수준 간의 상호작용 효과는 유의하지 않은 것으로 확인되었다. 이는 학습자의 초기 쓰기 수준이 사후 성과를 설명하는 핵심 변인으로 작용한 반면, 피드백의 주체가 인간인지 AI인지는 성과 차이를 결정짓는 주요 요인으로 작동하지 않았음을 의미한다. 이는 AI 피드백의 교육적 효과를 학습자의 특성, 과제 성격 등을

함께 고려하여 해석할 필요가 있음을 시사한다.

학생 응답에 대한 질적 분석 결과는 이러한 양적 분석 결과를 보다 심층적으로 이해할 수 있는 근거를 제공한다. 본 연구에서는 학습자 인식 자료와 AI 피드백 텍스트 분석을 통해 AI 피드백의 특질을 '구체성과 실현 가능성', '체계성과 일관성', '표현 개선 중심성', '전형성', '메타인지 유도의 한계'라는 다섯 가지 범주로 도출하였다. AI 피드백은 무엇을 어떻게 수정해야 하는지를 명확히 제시함으로써 학습자의 즉각적인 수정 행동을 촉진하였고, 비교적 안정적이고 일관된 구조를 통해 이해 가능성과 적용 가능성이 높게 인식되었다. 특히 표현의 명료성, 문장 구성, 가독성 개선과 같은 영역에서 강점을 보였는데, 이는 짧은 논술 답안이라는 과제 특성과 결합되어 AI 피드백의 효과가 더욱 분명하게 나타난 결과로 해석할 수 있다.

반면, AI 피드백은 개별 학습자의 사고 맥락이나 논증 전략을 심층적으로 반영하기보다는 정형화된 서술 틀을 반복하는 전형성을 지니고 있었으며, 학습자의 사고를 확장하거나 자기조절적 성찰을 유도하는 메타인지적 기능에는 일정한 한계를 보였다. 이는 AI 피드백이 고차적 논증 요소나 사고의 질적 전환을 직접적으로 촉진하기보다는, 이미 작성된 텍스트를 보다 명료하고 완결된 형태로 다듬는 데 주로 기여하고 있음을 시사한다.

이상의 결과를 종합할 때, AI 피드백은 교사의 피드백을 대체하는 도구라기보다는, 글쓰기 과정에서 효율적인 1차 수정 도구로 활용될 가능성이 크다. 특히 혼합형 피드백의 관점에서 볼 때, 1차적으로 AI 피드백을 활용하여 표현과 구조 수준의 수정과 점검을 수행하고, 이후 교사가 고차적 논증 요소에 대한 피드백이나 학생의 메타인지를 유도하는 질문 중심의 피드백을 보완하는 방식이 보다 적절한 활용 방안이 될 수 있다.

본 연구는 생성형 AI 기반 피드백이 짧은 논술형 답안 맥락에서 인간 피드백과 비교할 때 유사한 수준의 학습 성과를 도출할 수 있으며, 특히 표현 중심의 글 고쳐쓰기 단계에서는 안정적이고 실질적인 학습 지원 도구로 기능할 수 있음을 실증적으로 보여준다. 동시에 AI 피드백의 전형성과 메타인

지 유도의 한계를 고려할 때, 이를 교사의 피드백을 대체하는 수단으로 이해하기보다는, 전체적인 교수·학습 설계 안에서 전략적으로 결합·보완해야 할 도구로 활용할 필요가 있음을 시사한다. 향후 연구에서는 과제의 분량과 유형을 확장하여 AI 피드백의 효과가 긴 논술문이나 복합적 논증 과제에서도 동일하게 나타나는지를 검증하고, 혼합형 피드백 모델의 실제 수업 적용 효과를 보다 정교하게 탐색할 필요가 있다.

* 본 논문은 2026.01.26. 투고되었으며, 2026.02.08. 심사가 시작되어 2026.03.07. 심사가 종료되었음.

참고문헌

권태현(2024),「ChatGPT를 활용한 쓰기 채점 및 피드백 방안 – 프롬프트 전략을 중심으로 -」,『새국어교육』141, 7-42.

박영민 · 가은아 · 권태현 · 박종임 · 박찬홍 · 이수진 · 이재기(2025),『작문교육론』. 서울: 역락.

심지연(2024),「생성형 AI 를 활용한 한국어 쓰기 피드백 방안 -ChatGPT 와 한국어 교사의 쓰기 피드백에 대한 학습자 인식 및 선호도 차이를 중심으로」,『Journal of Korean Culture (JKC)』67, 45-79.

최숙기(2024),「국어 교사의 챗지피티 (ChatGPT)를 활용한 작문 피드백 경험에 대한 사례 연구」,『교원교육』40(1), 379-404.

최진영 · 김형성 · 송보라 · 김지수(2023),「생성형 인공지능 ChatGPT를 활용한 고등학생의 작문 과제 고쳐쓰기 양상」,『리터러시 연구』14(6), 79-119.

Applebee, A. N. & Langer, J. A. (2011), "EJ Extra: A snapshot of writing instruction in middle schools and high schools [free access]". English Journal 100(6), 14-27.

Bandura, A. (1997), *Self-efficacy: The exercise of control* (Vol. 11). New York: Freeman.

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., ... & Chen, G. (2024), "Assessing the proficiency of large language models in automatic feedback generation: An evaluation study", *Computers and Education: Artificial Intelligence* 7, 100299.

Escalante, J., Pack, A., & Barrett, A. (2023), "AI-generated feedback on writing: Insights into efficacy and ENL student preference", *International Journal of Educational Technology in Higher Education 20*(1), 57.

Evans, C. (2013), "Making sense of assessment feedback in higher education". *Review of educational research, 83*(1), 70-120.

Graham, S., Harris, K., & Hebert, M. (2011), *Informing Writing: The Benefits of Formative Assessment. A Report from Carnegie Corporation of New York*, New York: Carnegie Corporation of New York.

Guo, K., Pan, M., Li, Y., & Lai, C. (2024), "Effects of an AI-supported approach to peer feedback on university EFL students' feedback quality and writing ability", *The Internet and Higher Education 63*, 100962.

Jacobsen, L. J. & Weber, K. E. (2025), "The promises and pitfalls of large language models as feedback providers: A study of prompt engineering and the quality of AI-driven feedback", *AI 6*(2), 35.

Jacobsen, L. J., Mertens, U., Jansen, T., & Weber, K. E. (2025). AI, Expert or Peer? – –Examining the Impact of Perceived Feedback Source on Pre-Service Teachers Feedback Perception and Uptake. arXiv preprint arXiv:2507.16013.

Lin, S. & Crosthwaite, P. (2024), "The grass is not always greener: Teacher vs. GPT-assisted written corrective feedback", *System 127*, 103529.

Lo, L. S. (2023), "The CLEAR path: A framework for enhancing information literacy through prompt engineering", *The Journal of Academic Librarianship 49*(4), 102720.

Mekheimer, M. (2025), "Generative AI-assisted feedback and EFL writing: a study on proficiency, revision frequency and writing quality", *Discover Education 4*(1), 170.

Mizumoto, A. & Eguchi, M. (2023), "Exploring the potential of using an AI language model for automated essay scoring", *Research Methods in Applied Linguistics 2*(2), 100050.

Ranalli, J. (2021), "L2 student engagement with automated feedback on writing: Potential for learning and issues of trust", *Journal of Second Language Writing 52*, 100816.

Rashkin, H., Clark, E., Huot, F., & Lapata, M. (2025), "Help Me Write a Story: Evaluating LLMs' Ability to Generate Writing Feedback", In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 25827-25847.

Stahl, M., Biermann, L., Nehring, A., & Wachsmuth, H. (2024), "Exploring LLM prompting strategies for joint essay scoring and feedback generation", arXiv preprint arXiv:2404.15845.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., ... & Olson, C. B. (2024), "Comparing the quality of human and ChatGPT feedback of students' writing", *Learning and Instruction 91*, 101894.

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*(1), 010152.

Yoon, S. Y., Miszoglad, E., & Pierce, L. R. (2023), "Evaluation of ChatGPT feedback on ELL writers' coherence and cohesion", arXiv preprint arXiv:2310.06505.

Zhan, Y. & Yan, Z. (2025), "Students' engagement with ChatGPT feedback: Implications for student feedback literacy in the context of generative artificial intelligence", *Assessment & Evaluation in Higher Education*, 1-14.

Zhang, Z., Aubrey, S., Huang, X., & Chiu, T. K. (2025), "The role of generative AI and hybrid feedback in improving L2 writing skills: a comparative study", *Innovation in Language Learning and Teaching*, 1-19.

Zimmerman, B. J. &, Schunk, D. H. (2011), "Self-regulated learning and performance: An introduction and an overview" In B. Zimmerman & H. Dale(Eds.), *Handbook of self-regulation of learning and performance*, Florence, KY: Routledge.

초록

생성형 AI 기반 쓰기 피드백의 효과와 학습자 반응 분석

권태현

본 연구에서는 생성형 AI 기반 쓰기 피드백의 교육적 효과와 특성을 검증하기 위해 인간 평가자와 ChatGPT에 의해 산출된 피드백이 고등학생의 논술 답안 고쳐쓰기 성과에 미치는 영향을 비교·분석하였다. 이를 위해 154명의 고등학교 2학년 학생들을 대상으로 준실험적 사전–사후 설계를 적용하여 양적 효과를 분석하고, 학생 인식 설문을 통해 질적 특질을 도출하였다. 분석 결과, AI 피드백의 효과 크기가 인간 피드백에 비해 미세하게 높게 나타났으며, 이원분산분석에서는 쓰기 수준의 주효과만이 유의하게 확인되었다. 질적 분석 결과, AI 피드백은 구체성과 실현 가능성, 체계성과 일관성, 표현 개선 중심이라는 강점을 지니는 반면, 전형성과 메타인지 유도의 한계를 함께 내포하는 것으로 나타났다. 이러한 결과는 생성형 AI 피드백이 1차 수정 도구로 활용될 수 있음을 드러내며, 고차적 논증 요소와 메타인지 유도를 위해서는 교사 피드백과의 혼합 활용이 필요함을 시사한다.

핵심어 생성형 AI, 쓰기 피드백, 고쳐쓰기, 인간 피드백, 챗지피티

ABSTRACT

Effects of Generative AI–Based Writing Feedback and Learner Responses

Kwon Taehyun

This study compared the effects of feedback generated by human evaluators and ChatGPT on high-school students' revision performance in argumentative writing to investigate the educational effectiveness and characteristics of generative AI-based writing feedback. A quasi-experimental pretest-posttest design was used with 154 second-year high school students to examine quantitative effects and identify qualitative features through a learner perception survey. The results showed that the effect size of AI feedback was slightly larger than that of human feedback, and the two-way ANOVA revealed a significant main effect only for writing proficiency level. According to qualitative analysis, AI feedback demonstrated strengths in specificity, feasibility, systematicity, consistency, and a focus on improving linguistic expression, but it also had limitations related to stereotypical patterns and insufficient support for metacognitive engagement. These results suggest that generative AI feedback can be used effectively as a tool for initial revision, whereas a hybrid approach that combines it with teacher feedback is required to address higher-order argumentative elements and promote metacognitive processes.

KEYWORDS Generative AI, writing feedback, revision, human feedback, ChatGPT