



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

교육학석사 학위논문

그래프 이론을 이용한
서술형 문항 자동채점

2024년 7월

서울대학교 대학원
과학교육과 물리전공
김진우

그래프 이론을 이용한 서술형 문항 자동채점

지도교수 조 정 효

이 논문을 교육학석사 학위논문으로 제출함
2024년 7월

서울대학교 대학원
과학교육과 물리전공
김 진 우

김진우의 석사 학위논문을 인준함
2024년 7월

위 원 장 하 민 수 (인)

부위원장 송 진 응 (인)

위 원 조 정 효 (인)

국문초록

교육목표의 도달 여부를 확인하고 학습자의 인지 능력을 측정하며 학습과정의 개선 및 피드백을 위한 교육평가의 목표에 적합한 문항을 제작하는 것은 교육연구의 주요 관심분야이다. 그 중 응답 특성과 문제 해결과정을 더 자세히 볼 수 있어 학습자의 이해 수준을 효과적으로 점검하고, 이를 참고해 학습자에게 맞춤형 피드백을 제공할 수 있다는 점에서 서술형 문항(Descriptive Assessment)을 이용한 평가는 매우 중요하다. 그럼에도 불구하고 현장에서 선택형 문항이 주요하게 사용되는 이유는, 채점과정에 드는 노력과 비용의 문제와 평가 결과의 신뢰도 때문이다.

서술형 평가 채점에 대한 문제를 컴퓨터 기술로 해결하기 위한 자동채점 연구가 활발히 진행되고 있다. 정답 키워드의 출현 여부를 확인하는 통계기반 기술부터 시작된 자동채점 연구는 크게 자질(feature)을 추출하고 분류모델을 학습시키는 머신러닝(machine learning) 기술, 자질을 사용하지 않는 인공 신경망 기반의 모델을 학습시키는 딥러닝(deep learning) 기술을 활용하는 방법으로 나눌 수 있다. 최근에는 학습과정 없이 즉각적으로 사용할 수 있는 생성형 AI(Generative Artificial Intelligence)를 자동채점에 활용하려는 연구도 활발히 진행되고 있다. 하지만 선행 연구들은 모델의 정확도 확보를 위해 양질의 대량의 데이터를 필요로 하고, 타당한 자질 추출에 어려움을 겪는다. 또한 구축된 모델은 학습 문항에만 적용할 수 있어 일반화 능력이 떨어진다. 추가적으로 머신러닝 모델은 타당한 자질 추출에 어려움을 겪고, 딥러닝 모델은 채점 결과 해석이 어렵다. 생성형 AI를 활용한 연구도 응답의 수준과 편차가 커 신뢰도 확보가 어렵다는 한계가 존재한다. 따라서 신뢰도와 타당성이 높은 가운데 대량의 데이터를 필요로 하지 않고, 자

질 추출이 용이하여 채점 결과 해석이 가능하고, 결과와 피드백을 즉각적으로 반환가능한 모델에 대한 연구가 필요하다.

본 연구에서는 위의 목표에 알맞는 그래프 이론(Graph theory)을 이용한 자동채점 모델을 제안한다. 자연어처리능력이 뛰어난 생성형 AI를 활용하여 텍스트 답안의 삼중항을 추출한 뒤, 이를 기반으로 구축한 이분 그래프를 인접행렬로 변형하여 참조 답안과의 그래프 유사도를 정량화한다. 이 과정에서 텍스트의 특징인 유사어와 명사어등을 생성형 AI를 이용하여 식별하고, 정답 요소로 반영하여 모델의 정확도를 높였다. 또한 추출한 삼중항간의 관계를 인과 그래프로 표현하여, 절(Clause)간의 논리 관계를 그래프로 시각화할 수 있다. 두 그래프 유사도를 바탕으로 참조 답안과 학생 답안의 유사도를 연속형 숫자로 표현하고, 이를 바탕으로 채점 기준에 따른 자동 채점 점수를 반환한다. 이때 채점기준을 2가지로 구분하였다. 채점기준1은 본 연구에서 제안하는 모델의 채점 방식인 참조 답안과 일치여부를 통해 채점하는 방식이고, 채점기준2는 참조 답안을 포함한 과학적으로 타당한지 여부를 통해 채점하는 방식이다.

본 연구에서 제안하는 Graph 모델의 성능을 평가하기 위해 ChatGPT4o와 TF-IDF 방법을 대조군으로 설정하고, 서울시 소재 과학고등학교 1학년 학생들이 발전기의 역할과 원리에 대해서 응답한 137개의 데이터를 자동채점했다. 수동채점과의 정확도를 포함한 5가지 지표로 성능을 평가한 결과 채점기준1, 2에 대해 각각 .934, .763의 정확도를 나타낸 Graph 모델이 채점기준 1, 2에 대해 각각 .511, .431의 정확도를 나타낸 ChatGPT4o보다 우수한 성능을 보였다. 참조 답안과의 유사도로 채점하는 채점기준1에 대해서는 매우 우수한 성능을 보이나, 과학적으로 타당한 답안을 확인하는 채점기준2에 대해서는 그 성능이 저하되는 것을 확인할 수 있다.

이를 보완하기 위해 주의어를 사용자에게 표시하여 최종 점수 결정에 반영하도록 하였다.

Graph 모델은 동일한 점수의 학생들끼리도 그래프 유사도를 통해 문항과 관련된 개념에 대한 이해도를 연속적인 숫자로 정량화할 수 있어 개념 점검 및 맞춤형 피드백이 가능하고, 논리관계 존재 유무를 통해 과학적으로 타당한 논리표현으로 답안을 작성했는지 평가하고 피드백할 수 있다.

본 연구에서는 그래프 이론을 이용하여 내용 기반의 서술형 문항을 자동채점하는 모델을 제안하고, 교육적인 활용방법을 제시하였다. 본 모델은 첫째, 답안을 연속적인 숫자로 정량화하고, 채점요소가 명확하여 사용자는 채점결과를 해석하기 용이하다. 둘째, 보완해야 할 개념 또는 개념간의 관계를 교사가 인식하고 피드백에 즉각적으로 반영할 수 있다. 셋째, 학습과정이 필요하지 않아 양질의 대량의 데이터를 요구하지 않으므로 학급 수준부터 대규모 시험까지 활용 가능하다는 점에서 그 의의가 있다. 다만, 정답에서 포함되는 오답 요소를 식별하여 감점 요소로 채택하는 방법 개발과 다양한 참조 답안을 통해 유사 답안 채점 가능성의 폭을 넓히는 추가 연구를 통해 Graph 모델을 정교하게 발전시킬 필요가 있다.

주요어 : 그래프 이론, 자동채점, 서술형 문항, 인과 관계, 생성형 AI

학 번 : 2022-26856

목 차

1. 서론	1
1.1. 연구의 필요성	1
1.2. 연구 문제	6
1.3. 연구의 제한점	7
1.4. 용어의 정의	9
2. 이론적 배경	11
2.1. 서술형 문항(Descriptive Assessment)	11
2.1.1. 평가 문항의 유형	11
2.1.2. 서술형 평가의 의미와 도입 필요성	15
2.1.3. 서술형 평가 도입의 어려움	18
2.2. 자동채점(Automated Scoring)	19
2.2.1. 내용 기반 자동채점	19
2.2.2. 생성형 AI	20
2.2.3. 자동채점 연구 동향 및 한계점	22
2.3. 그래프 이론(Graph Theory)	32
2.3.1. 이분 그래프(bipartite graph)	33
2.3.2. 그래프를 이용한 자동채점	34
3. 연구 방법	36
3.1. 연구 절차	36
3.2. 분석 자료	39
3.3. 분석 방법	43
3.3.1. 프롬프트 엔지니어링을 통한 대조군 제작	44
3.3.2. TF-IDF를 통한 대조군 제작	47

3.3.3. 자동채점 모델 구축	50
3.3.4. 자동채점 모델 평가	70
4. 연구 결과	74
4.1. 문항별 모델 자동채점 결과	74
4.1.1. 자동채점 모델 성능 평가 및 지표	74
4.1.2. 오채점 답안 분석	80
4.2. 자동채점 모델의 활용 방법	89
4.3. 자동채점 모델 프로그램	98
5. 결론 및 제언	101
5.1. 결론	101
5.2. 제언	105
참고문헌	107
Abstract	121

표 목 차

[표 2-1] 선택형 문항과 구성형 문항의 기준	12
[표 2-2] 평가 유형별 목표 적합도	13
[표 2-3] 서술형·논술형 문항 유형 비교	16
[표 2-4] 자동채점의 유형과 특징	19
[표 2-5] 자동채점 기술별 내용 기반 자동채점 선행 연구	28
[표 3-1] 데이터셋 개요	39
[표 3-2] 데이터셋의 문항, 채점기준1,2, 배점	41
[표 3-3] 데이터셋의 점수 척도 및 채점기준, 점수별 데이터 수	41
[표 3-4] 자동채점을 위한 프롬프트 디자인	45
[표 3-5] 17번 답안 ChatGPT4o 자동 채점결과(채점기준1, 2)	46
[표 3-6] 테스트 셋의 수동채점, 자동채점 점수	48
[표 3-7] TF-IDF를 이용한 중요 단어 상위 20개	49
[표 3-8] 삼중항 및 관계 추출 프롬프트 디자인	52
[표 3-9] 2번 답안의 삼중항 및 관계 추출 예시	53
[표 3-10] 추출 횟수간 삼중항 추출 일치율	53
[표 3-11] 12번 답안의 삼중항 추출 결과 및 누락요소	54
[표 3-12] 유사 동사 식별 프롬프트 디자인	58
[표 3-13] 명사어 식별 프롬프트 디자인	61
[표 3-14] 모델 성능 평가 지표(이진 분류)	71
[표 3-15] 모델 성능 평가 지표(다중 분류)	72
[표 3-16] QWK값에 따른 일치도 해석	73
[표 4-1] 자동채점 성능 평가 지표 (a) 채점기준1, (b) 채점기준2	74

[표 4-2] Graph 자동채점 오채점 유형 및 해결방안	80
[표 4-3] ChatGPT4o 자동채점 오채점 근거	85
[표 4-4] 텍스트, 채점 근거 및 피드백, 점수 (a) 5번 답안, (b) 32 번 답안	86
[표 4-5] 텍스트, 채점 근거 및 피드백, 점수 (a) 7번 답안, (b) 54 번 답안	87
[표 4-6] 70번 답안 텍스트, 채점기준1, 2 채점근거 및 피드백, 점 수	88

그림 목 차

[그림 2-1] Bloom의 신교육목표분류표(인지적 영역)	14
[그림 2-2] 기억, 이해 수준과 평가 유형의 관계	15
[그림 2-3] Chain-of-thought Prompt를 활용한 추론 과제 성능 향상의 예시	22
[그림 2-4] 노드(v1~v6)와 간선(e1~e5)으로 이루어진 그래프	32
[그림 2-5] 이분 그래프와 인접행렬의 예시	34
[그림 3-1] 연구 절차	38
[그림 3-2] 통합과학 교과서 내 발전기에 대한 설명	40
[그림 3-3] 그래프를 이용한 자동채점 연구 과정 개요 ...	42
[그림 3-4] 자연어 생성 품질 평가 프레임워크 G-Eval의 개요	44
[그림 3-5] 삼중항 및 인과관계 추출	51
[그림 3-6] 참조 답안 지식 그래프	56
[그림 3-7] 참조 답안 관계 그래프	56
[그림 3-8] 2번 답안의 핵심 개념 및 동사 노드 표시(유사 동사 식별)	59
[그림 3-9] 2번 답안의 핵심 인과 관계 표시	60
[그림 3-10] 핵심 개념 및 동사 노드 표시(유사 동사, 명사어 식별) (a) 41번 답안, (b) 39번 답안	62
[그림 3-11] 99번 답안의 핵심 개념 및 동사 노드 표시(유사 동사, 명사어, 주의어 식별)	63
[그림 3-12] 참조 답안의 이분 그래프	64
[그림 3-13] 22번 답안의 이분 그래프	64

[그림 3-14] 참조 답안의 인접 행렬	65
[그림 3-15] 34번 답안의 인접 행렬	66
[그림 3-16] 핵심인접행렬 (a) 참조 답안 (b) 2번 답안 ...	67
[그림 3-17] 핵심인접행렬 (a) 34번 답안 (b) 77번 답안 ·	68
[그림 3-18] 혼동 행렬 예시 (a) 이진 분류 혼동 행렬 (b) 다중 분 류 혼동 행렬	71
[그림 4-1] Graph 자동채점 점수 분포 (a) 채점기준1, (b) 채점기 준2	75
[그림 4-2] 채점기준별 수동채점 점수에 따른 Graph 자동채점 점 수	76
[그림 4-3] 반올림된 Graph 자동채점 점수 분포 (a) 채점기준1, (b) 채점기준2	77
[그림 4-4] 반올림된 Graph 자동채점 점수와 수동채점 점수의 혼 동행렬 (a) 채점기준1, (b) 채점기준2	78
[그림 4-5] ChatGPT4o 자동채점 점수 분포 (a) 채점기준1, (b) 채 점기준2	79
[그림 4-6] ChatGPT4o 자동채점 점수와 수동채점 점수의 혼동행 렬 (a) 채점기준1, (b) 채점기준2	79
[그림 4-7] 7번 답안 텍스트, 지식 그래프, 인접 행렬	81
[그림 4-8] 80번 답안 텍스트, 지식 그래프, 인접 행렬 ...	82
[그림 4-9] 45번 답안 텍스트, 지식 그래프, 인접 행렬 ...	83
[그림 4-10] 80번 답안 텍스트, 지식 그래프, 인접 행렬 ...	90
[그림 4-11] 23번, 114번 답안의 텍스트, 인접행렬, Automated score, 수동채점 점수	91
[그림 4-12] 54번, 134번 답안의 텍스트, 인접행렬, Automated score, 수동채점 점수	92
[그림 4-13] 21번, 76번 답안의 텍스트, 지식 그래프, 인접행렬,	

Automated score, 수동채점 점수	94
[그림 4-14] 73번, 123번 답안의 텍스트, 지식 그래프, 인과 그래프, Automated score, 수동채점 점수	96
[그림 4-15] 자동채점 모델 프로그램의 전체 학생 분석 화면 예시	98
[그림 4-16] 자동채점 모델 프로그램의 학생별 분석 화면 예시	99

1. 서론

1.1. 연구의 필요성

교육평가의 타당성을 확보하고 공정성과 신뢰성을 보장하기 위해 평가 문항 제작은 매우 중요하다. 평가 문항의 유형으로는 지필식 시험인 선다형, 진위형, 연결형 문항을 포함하는 선택형 문항과 서답형 시험인 단답형, 완성형, 논술형 문항을 포함하는 구성형 문항이 있다. 한국교육과정평가원에서 주관하는 우리나라의 국가수준 대규모 평가인 대학수학능력시험에서는 수학 영역에서 30%의 문항이 단답형인 점을 제외하고는 모두 선다형 문항이다. 국가수준 학업성취도 평가는 문항 비율 또는 기출문제가 공개되어 있지 않아 정확히 파악할 수는 없지만 2009년부터 서답형 문항을 포함하고 있고 2022년부터는 구성형 문항을 포함하고 있다(박인용, 2023). 시·도 교육청 수준에서는 학업성적관리지침에 따라 서·논술형 평가 권장 반영 비율을 설정하고, 이는 단위학교에서 평가에 대한 지침이 되고 있다. 하지만 현장의 교사들은 많은 유사 답안, 엄청난 채점 물량, 성적 처리에 턱없이 부족한 물리적 시간, 문항 및 채점기준표 작성의 어려움과 같은 이유로 정기고사에서 서·논술형 평가 실시를 기피하고 수행평가에서 다루는 경향이 높다(박혜영, 2019). 스위스의 비영리 교육재단 IBO가 운영하는 IB(International Baccalaureate) 교육과정도 도입하여 토론수업의 논, 서술형 평가를 도입하려는 움직임이 점차 나타나고 있으나, 정기고사를 포함해 채점 근거의 엄밀함이 요구되는 큰 시험에서는 선다형 위주의 평가가 이루어지고 있는 것은 사실이다.

답안을 만들어 내야 하는 구성형 문항은 보기가 주어진 문항과 비교하여 응시자들의 인지적 과정을 직접적으로 측정하고, 선다형 문항에 비하여 추측의 요인이 배제되어 변별도와 검사의 신뢰도가 높다(Bennett, 1993). 선택형 문항은 답안 중 최선의 답을 찾으므로 창의력이 발휘되지

않고 어휘능력도 쓰이지 않으나, 구성형 문항은 새로운 문제를 해결하는 능력을 기르기에 유리하다(lee, 2011). 또한 서술형 평가는 선택형 평가와 비교하여 진단 기능적 측면, 사고의 촉진 측면, 개념 변화의 측면에서 효과적이다(하민수, 2019). 4차 산업혁명의 미래 사회에서는 풍부한 지식을 갖춘 인재를 넘어 창의성, 문제해결력, 협력과 소통, 포용력과 인성을 갖춘 인재 양성을 요구한다(송진웅, 2019). 이에 필요한 역량 함양에도 현재의 지필식 시험 위주의 지식 중심 평가가 아닌, 구성형 문항으로 이루어진 과정 중심 평가가 강조되어야 한다. 따라서 학습자의 고차원적인 사고 능력을 신장하고 평가하기 위해서 내용 타당도가 높고, 질적 평가가 가능한 구성형 문항에 대한 응답을 평가하는 것이 학생들의 역량과 사고력을 측정할 수 있다.

구성형 문항의 중요성은 정부 차원에서도 인지하고 장려하고 있는데, 교과 영역 간 통합형 문항으로 서술형이나 논술형 문항의 출제를 제안하거나(이용상, 2015), 2023년 12월 교육부가 발표한 ‘2028 대학입시제도 개편 확정안’에서 지식암기 위주의 5지선다형과 같은 문항을 지양하고 창의력·사고력·문제해결력 평가가 가능한 서·논술형 평가를 확대하고, 고교 교사의 평가 전문성 향상연수 집중 실시 계획에서 확인할 수 있다. 이는 단위 시·도 교육청에서도 활발히 장려되고 있으며, 서울시 교육청에서는 ‘2023학년도 중등 학생평가 내실화 계획’에서 서·논술형 평가를 학기 단위 성적의 20% 이상 권장하고 있다.

하지만 서·논술형 평가는 즉각적인 진단과 피드백이 어렵고, 다수 응답을 채점하는 과정에서 채점자의 일관성을 확보하는 것이 어렵다. 또한 대단위 평가에서는 일반적으로 2인 이상의 복수 채점을 적용하는데, 이는 인력과 시간이 훨씬 많이 요구 된다. 이처럼 채점의 일관성, 객관성, 공정성의 문제나 채점에 소요되는 시간과 비용의 문제는 서·논술형 평가의 현실적인 제한점이라 할 수 있다(박종임, 2023). 위와 같은 많은 어려움 중 채점자가 가장 부담스러워하는 부분은 단연 채점이다. 문항의 채점은 매번 채점자의 판단에 따라 상당한 시간과 노력을 필요로 하는데, 채점자 특성, 채점 피로도, 채점 시기 등의 기타 제반 요인에 영향을 받

을 가능성 또한 잠재되어 있다(노은희, 2016).

이러한 한계를 기술을 통해 보완하기 위하여 ‘자동채점(Automated Scoring)’에 대한 연구가 활발히 수행되고 있다. 컴퓨터 알고리즘과 소프트웨어를 사용하여 학습자들의 답안을 채점하는 자동채점은 사람이 채점하는 것과 비교하여 동일한 평가 기준을 일관성있게 적용하고, 효율적이며 즉각적인 맞춤형 피드백이 가능하다는 장점이 있다(Zhang, 2013). 영어권에서 자동채점은 1960년대 컴퓨터를 이용한 방식으로 시작되었다(Page, 1966). 이후 교육연구기업들이 주도하여 에세이(essay)로 통칭하는 긴 글 수준의 자동채점에 대한 연구(Automated Essay Scoring)가 활발히 진행되었는데, E-rater, Intelligent Essay Assessor, IntelliMetric 등이 있다. 국내 기관 수준에서 한국교육과정평가원에서는 서답형 문항 자동채점 프로그램 방안 연구(진경애, 2006; 2007; 2008)를 시작으로 문장 수준의 서답형 문항을 채점할 수 있는 프로그램을 개발했다(노은희, 2012; 2013; 2014; 2015; 2016). 이후 한국교육과정평가원에서는 컴퓨터 기반 서·논술형 평가를 위한 자동채점 방안에 대해 2차 연도 연구까지 진행되었다(박종임, 2022; 2023). 국내 개인 수준에서도 인공지능을 활용하여 서술형 평가를 자동채점 하거나(하민수, 2016; 2019), 영어 에세이(쓰기) 평가(박강운, 2021) 또는 한글 에세이(쓰기) 평가(조희련, 2021; 이용상, 2022)에 관한 연구가 진행되었다.

선행 연구에서 사용된 기술은 크게 머신러닝에 기반한 분류모델을 활용한 연구와 인공지능 기반의 딥러닝을 활용한 연구로 나눌 수 있다. 머신러닝을 활용한 연구는 자질(feature) 설정이 잘 이루어졌을 때 채점 결과의 해석이 용이하나, 자질을 탐색하고 설정하는 것에 많은 시간과 노력이 든다(장지현, 2021). 또한 모델에 설정하는 자질은 교과, 문항 유형, 응답자에 따라 달라지므로 문항마다 새롭게 설정하거나, 일반화할 수 없는 단점이 있다. 한편 딥러닝을 활용한 연구는 많은 데이터를 학습할 수록 인간 채점 결과와의 일치도가 높아지는 장점이 있으나, 블랙박스과 같은 은닉층으로 인하여 채점 결과를 해석하기 어렵다는 단점이 뚜렷하다. 이는 교육평가의 목적인 학습 과정 전반에 대한 지원과 개선에 적합

하지 않다. 무엇보다 기존의 연구방법은 인간이 채점한 결과와의 일치도를 향상시키기 위해서 모델을 학습시킬 많은 데이터가 필요하다. 2023년 서울시 학교 1개당 평균 학생 수¹⁾가 중학교는 509명, 고등학교는 643명 이므로 서울시 학교 1개당 평균 1개 학년의 학생 수는 약 250명 이내라고 볼 수 있다. 2023년 전국 중학교, 고등학교 개수로 학년별 학생수를 나누면 전국 학교 1개당 평균 학생 수²⁾가 중학교는 약 405명, 고등학교는 약 531명이므로 전국 학교 1개당 평균 1개 학년 학생 수는 약 200명 이내라고 볼 수 있다. 모든 학생이 충실하게 답안을 작성하여 모두 데이터로 활용 가능하다고 하더라도 200~250개 데이터는 분명 모델이 학습하기에 충분한 데이터라고 하기엔 무리가 있고, 이는 곧 모델의 신뢰도가 떨어짐을 의미한다.

한편 사전 학습된 대형 언어 모델(Pre-trained Large Language Model)인 생성형 AI(generative Artificial Intelligence)의 급격한 발전으로 인하여 교육 분야의 평가 도구로 사용하기 위한 관심이 커지고 있다 즉, 교육 분야를 위해 특화시킨 AI가 아닌 상용화된 인공지능 기술의 활용 기술에 대한 탐색이 필요하다(백중호, 2023). 생성형 AI는 별도의 학습의 과정을 거치지 않거나 미세 조정(Fine tuning)을 통해 채점 도구로 활용될 수 있어 이에 관한 연구도 진행되고 있다(Jessica, 2023; Lartif, 2024; 박소영, 2023). 학습에 필요한 데이터가 작거나 없어도 된다는 점에서 기존의 모델을 학습시키는 연구에 비해 큰 장점이 있다. 하지만 생성형 AI를 이용한 자동채점은 응답이 가능한 범위나 수준의 측면에서 폭이 너무 넓어, 신뢰도와 타당성을 확보해야 하는 채점의 과정에 적용하기에는 한계가 있다는 의견도 있다(백중호, 2023).

종합하면 자동채점 기술로 교과, 문항 유형, 응답자, 데이터 수에 구애받지 않고 다양한 분야에 적용할 수 있는 범용적이고 일반화할 수 있으며 신뢰도와 타당성을 확보할 수 있는 기술이 필요하다. 위와 같은 기술로 본 연구에서는 그래프 이론(Graph theory)을 기반으로 한 자동채점

1) 서울특별시 (2024). 서울통계통합플랫폼. <https://stat.eseoul.go.kr/>

2) 한국교육개발원 (2024). KESS 교육통계서비스. <https://kess.kedi.re.kr/index>

방법을 제안한다. 그래프 이론은 수학자 오일러(Leonhard Euler)가 1736년에 발표한 이론으로, 개별 객체를 뜻하는 정점 또는 노드(node)와 정점 또는 객체 사이의 관계를 설명하는 간선(edge)로 구성된다. 학습자의 응답을 조사를 제외한 기본형의 형태소로 분리하면, 노드에 해당하는 개별 객체를 추출할 수 있고 문장을 분석하여 객체 사이의 관계를 이어 간선을 그릴 수 있다. 이를 통해 모범 답안으로 구축한 그래프와 다른 응답으로 구축한 그래프를 분석하여 그 유사도를 통해 예상 점수를 반환한다. 이 과정에서 생성형 AI를 통해 응답을 형태소로 분리, 유사어 추출, 문장간 인과관계 추출, 형태가 다른 품사 인식 등의 기능을 수행한다. 이와 같은 그래프 이론을 이용한 자동채점 방식은 첫째. 학습이 필요하지 않으므로 데이터 수에 의존하지 않으므로 단위 학교, 학급 수준에서도 활용 가능하다. 둘째. 그래프의 유사도를 통해 채점하기 때문에 채점 결과를 이해하고 응답자의 피드백에 활용할 수 있다. 셋째. 그래프 구축을 통한 채점 방식이 일관적이므로 교과, 문항에 제한이 없어 활용 가능성이 높다. 넷째. 모범 답안의 수준을 설정하여 그래프를 구축할 수 있으므로 응답자의 연령, 학력 수준을 고려한 채점이 가능하다. 위와 같은 장점을 갖는 그래프를 이용한 자동채점 방법을 제안하고자 한다.

1.2. 연구 문제

본 연구의 목적은 서술형 문항에 대한 응답을 그래프 이론을 활용하여 자동으로 채점하는 방법을 제안하고 평가 보조 도구로써의 타당성을 평가하는 데 있다. 나아가 현장에서의 활용 가치를. 이러한 목적을 달성하기 위한 구체적인 연구 문제는 다음과 같다.

- (1) 그래프 이론을 이용한 자동채점은 평가 보조 도구로써 타당한가?
- (2) 그래프 이론을 이용한 자동채점은 실제 교육현장에서 어떻게 활용될 수 있는가?

1.3. 연구의 제한점

연구과정에서 생성형 AI인 ChatGPT를 활용하고 있다. 이 과정에서 생성형 ChatGPT의 한계점이 반영된다.

첫째. 다양한 분야의 일반적인 데이터로 학습한 대화형 인공지능 챗봇이므로, 훈련 데이터가 포함된 시점까지의 지식만으로 결과를 반환한다. 따라서 최신 정보나 전문적인 분야에 대한 지식이 부족할 수 있다.

둘째. 영어보다 한글을 분석하고 처리하는 기능이 상대적으로 뒤떨어져 입력된 프롬프트의 맥락 또는 뉘앙스를 오해할 수 있다. 그러나 본 연구에서는 생성형 AI에게 전문 분야에 대한 지식이 아닌 일반적인 품사 구분과 유사어 식별의 기능 및 인과관계 추출 작업을 요구하므로 위 두 가지 한계점이 크게 영향을 미치지 않을 것으로 예상된다.

본 연구에서 제안하는 방법의 성능 평가는 연구자가 제작한 문항에 대해 통합과학, 물리학 I 교과를 학습한 서울 소재의 한 과학고등학교 내 150명 가량의 학생들이 응답한 데이터를 바탕으로 이루어졌다. 이에 따라 다음과 같은 제한점을 가진다.

첫째. 제작한 문항은 교과 지식의 이해·활용 수준을 묻는 문제로, 응답자들은 ‘필수 단어’를 포함하여 답한다. 이러한 유형에 대한 답안들로 모델 성능 평가가 이루어지므로, 모델의 성능을 모든 서술형 문항에 일반화하기에는 어려움이 있다. 그러나 해당 문항의 형식은 단위 학교에서 자주 사용되고 있고, 용어에 대한 지식과 이해가 필요한 과학 교과의 특성을 반영하고 있으므로 평가도구로서 교육적 의미가 충분하다고 사료된다.

둘째. 모델의 성능을 일반화하고, 보편적 활용가능성을 평가하기 위해 학교군, 학년, 성취수준을 다양화하여 학생들의 응답을 수집해 적용해야 한다. 본 연구에서는 한 개의 학교 내 두 개 학년에 대하여 세 개의 물리학 분야에 대한 문항을 사용하여 모델의 성능을 평가했다. 모델의 특성상 특정 교과에 국한되지 않고 범용적으로 사용할 수 있고, 특히 ‘필수 단어’를 노드(node)로 표현하고 있으므로 용어의 중요성이 강조되는 과

학 교과에 확장하여 적용할 수 있다.

본 연구에서 제안하는 방법은 생성형 AI를 이용해 식별한 유사어 결과를 반영하여 그래프의 유사성을 통해 채점한다. 이에 따라 다음과 같은 한계점을 갖는다.

첫째, 생성형 AI가 식별하지 못하거나 잘못 식별한 유사어에 대해서는 정답요소로 간주되지 않는다. 또한 동일한 과학 현상을 설명하기 위해 다른 과학 용어를 사용할 수 있는데, ChatGPT는 의미가 유사한 단어만을 식별하기 때문에 이 또한 정답요소로 간주되지 않는다. 이는 인간이 채점한 결과와 달라져 타당성을 잃을 수 있다. 하지만 간선의 관계를 통해 채점하지는 않았으나 사용자에게 ‘주의어’로써 시각적인 표시를 할 수 있도록 코드를 작성했다. 사용자는 인터페이스를 통해 ‘주의어’를 확인하고 자동채점 결과를 수용하거나 수정할 수 있다.

둘째, 그래프의 유사도 확인은 사용자가 입력한 필수 노드와 간선에 대해서만 이루어진다. 이에 따라 필수 노드와 간선을 제외한 노드와 간선에서 오답이 존재할 경우, 이를 명백하게 식별하기 어렵다. 이를 해결하기 위해 희귀 노드(Unique node)를 도입하여, 답안에서 출현하는 빈도수가 낮은 노드에 대해 사용자가 확인하고 최종 점수를 결정할 수 있도록 코드를 작성하였다.

1.4. 용어의 정의

(1) 서술형 문항(Descriptive assessment)

정책적으로 도입된 용어인 서술형 문항은 어느 정도 답이 정해진 문항에 대해 응답자가 개념 요약, 정리, 사실의 나열, 풀이 과정을 작성하는 문항을 의미한다. 전통적인 문항 분류에 포함되지 않으나, 일반적으로 단답형 또는 완성형보다는 길고, 논술형에 비해서는 분량이 많지 않은 문항을 지칭한다(박도순 외, 2007). 논술형 문항과 달리 피험자의 응답을 제한하는 측면이 있고, 논리적 글쓰기가 아닌 지식, 문제 해결의 관점에서 서술의 깊이와 넓이, 정확성을 판단하는 문항을 의미한다(박혜영, 2019). 본 논문에서는 위와 같은 속성을 갖는 문항을 서술형 문항으로 사용하고, 자동채점의 목표 평가 문항 유형으로 설정한다.

(2) 자동채점(Automated scoring)

컴퓨터 기술을 사용하여 응답을 평가하고 채점하는 작업을 뜻한다(Dikli, 2006). 교육학 분야에서는 학생의 수행 결과물에 대해, 측정과 학습을 위한 채점과 피드백을 목적으로 컴퓨터 알고리즘을 통해 평가하는 과정이다(Foltz, 2016). 또는 연구자 입장에서는 학생이 작성한 문어 텍스트에 대해 인간 채점자의 채점 과정 중 일부 또는 전체를 자동화하고자 하는 연구 분야로 정의할 수 있다(박종임, 2022).

(3) 생성형 AI(Generative Artificial Intelligence)

주어진 데이터로부터 새로운 콘텐츠를 생성하는 인공지능 시스템을 의미한다. 텍스트, 이미지, 음악, 코드 등 다양한 형태의 데이터를 딥러닝 기술을 통해 학습한 모델로, 다양한 분야에 활용되고 있다. OpenAI에서 개발한 ChatGPT, Google에서 개발한 Gemini, Microsoft에서 개발한 Bing, Meta에서 개발한 LLaMA, NAVER에서 개발한 하이퍼클로바 등이 있다. 본 연구에서는 ChatGPT를 이용하여 대량의 텍스트를 분석하였다.

(4) 프롬프트 엔지니어링(Prompt Engineering)

생성형 AI 모델, 특히 대규모 언어 모델이 사용자가 작업을 요청하기 위해 작성하는 입력 텍스트를 프롬프트(Prompt)라고 한다. 이때 프롬프트가 맥락이 있고 명확하며 예시가 포함되는 등 잘 설계될수록 생성형 AI의 성능이 최적화된다. 이렇게 생성형 AI가 요청한 작업에 최적의 결과를 반환하도록 사용자가 프롬프트를 정교화하는 것을 프롬프트 엔지니어링이라고 한다. 본 연구에서 대조군으로 설정한 생성형 AI를 이용하여 직접 채점하는 방법과 이분 그래프를 구축할 때 텍스트 분석 과정에서 프롬프트 디자인 작업이 이루어진다.

(5) 그래프 이론(Graph Theory)

수학과 컴퓨터 과학의 그래프 구조를 연구하는 분야이다. 개별 노드(Node)와 노드들을 연결하는 간선(Edge)으로 구성되어 다양한 문제를 모델링하여 해결하기 위해 활용된다. 컴퓨터 과학에서는 데이터 구조, 알고리즘, 데이터베이스, 네트워크 과학 등에 활용되는데, 소셜 네트워크 분석 추천시스템, 자연어 처리 등에서 응용된다. 본 연구에서는 텍스트를 그래프로 구축하여 텍스트 내의 구조를 파악하고 유사성을 바탕으로 예상 점수를 반환하는 방법에 활용할 것이다.

(6) 이분 그래프(Knowledge Graph)

그래프 이론에서 두 개의 분리된 집합으로 구성된 노드들 간의 관계를 나타내는 특수한 유형의 그래프이다. 동일한 집합 내의 노드끼리는 연결되지 않으며, 모든 간선은 서로 다른 집합의 노드를 연결한다. 매칭 문제, 소셜 네트워크 등에 활용되는데, 데이터 속 복잡한 구조와 관계를 이해하고 분석하는 데 활용된다.

2. 이론적 배경

2.1. 서술형 문항(Descriptive Assessment)

학습자가 교육의 목표에 도달했는지 판단하고, 학습자의 인지능력을 측정하는 도구로써 평가 문항은 교육 평가에서 중요한 기능을 한다. 교육이 일어나는 과정에서 학습자를 지원하고 개선하는 역할뿐만 아니라 교 교사의 교수·학습 방식을 개선하고 추후 자료 제작의 방향을 결정한다(백순근, 2019). 이에 따라 평가 문항의 유형을 이해하고 그 의미와 한계점에 대해 알아볼 필요가 있다. 선행 연구들이 각기 ‘문항’과 ‘평가’를 혼용해서 사용하고 있어, 인용 과정에서 두 용어가 혼재할 수 있다. 참고 문헌의 왜곡을 피하기 위해 선행연구들의 표현을 그대로 인용하되, 해외 논문의 경우 가독성을 위해 의역하였다.

2.1.1. 평가 문항의 유형

평가 문항은 활용 목적에 따라 진단평가, 형성평가, 총괄평가로 구분하거나 결과 해석 및 보고의 관점에 따라 기준참조평가, 준거참조평가, 능력참조평가, 성장참조평가로 구분한다(박혜영, 2019). 또 다른 방법으로는 평가 대상과 목표를 기준으로 선택형(Selected-Response) 평가, 구성형(Constructed-Response) 평가, 수행평가(Teacher Observation), 포트폴리오(Self-Assessment)로 유형화한다(McMillan, 2014). Laverty는 선택형 문항과 구성형 문항의 기준이 되는 4가지 질문을 제시하기도 했다([표2-1]).

[표 2-1] 선택형 문항과 구성형 문항의 기준(Laverty, 2016)

<u>Constructed Response Criteria</u>	<u>Selected Response Criteria</u>
1. Question gives an event, observation, or phenomenon for the student to explain or make a prediction about.	1. Question gives an event, observation, or phenomenon for the student to explain or make a prediction about.
2. Question gives a representation or asks student to construct a representation.	2. Question gives a representation or asks student to select a representation.
3. Question asks student to explain or make a prediction about the event, observation, or phenomenon.	3. Question asks student to select an explanation for or prediction about the event, observation, or phenomenon.
4. Question asks student to provide the reasoning that links the representation to their explanation or prediction.	4. Question asks student to select the reasoning that links the representation to their explanation or prediction.

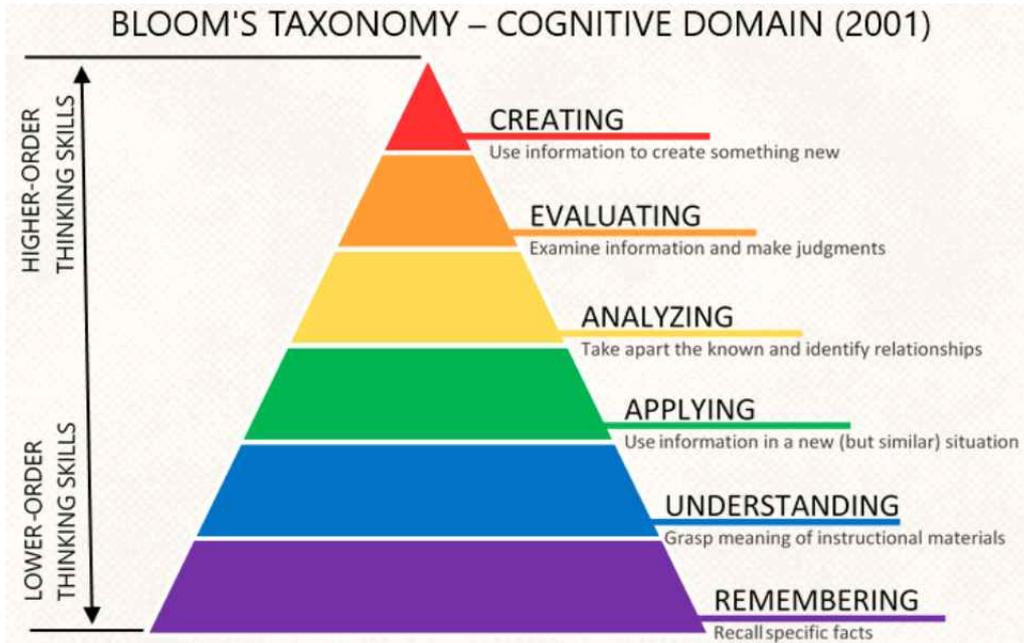
선택형 문항은 선다형, 진위형, 연결형 문항을 포함하는데, 채점이 객관적, 효율적인 특징이 있다. 구성형 문항은 응답자가 답안을 직접 만들어야 하는 문항으로 세부적으로 4가지로 유형화할 수 있다. 첫째. 문항에 대한 짧은 답을 요구하는 단답형과 빈칸을 채우는 완성형과 같은 간략한 답안(Brief-constructed response)이 있다. 둘째. 간략한 답안보다 복잡하고 구체적인 문항에 대해 수행 산출물을 제작하는 수행 평가(Performance tasks)가 있다. 셋째. 질문이나 명제에 대해 긴 글을 통해 논리적으로 답안을 작성하는 논술형 평가(Essay tests)가 있다. 넷째. 구두 질문과 답변을 통해 협력적으로 논쟁하며 토론하는 구두 질문(Oral questioning)이 있다. 다음으로 수행평가는 학생의 행동, 활동, 발언 등을 기록하여 학습자의 특성을 기록하는 방법이고, 포트폴리오는 학습자가 스스로의 학습 활동을 기록하고 평가하는 방법이다. McMillan은 학습의 목표(Target)를 5가지(Knowledge & Simple Understanding, Deep Understanding & Reasoning, Skills, Products, Affect)로 설정하고 목표에 따라 적합한 평가 방법을 사용해야 한다고 주장한다. 각 평가방법마다 5가지의 학습 목표와의 적합성을 1에서부터 5까지 5가지 단계로 수치화하여 제시하였다.([표2-2])

[표 2-2] 평가 유형별 목표 적합도(McMillan, 2014)

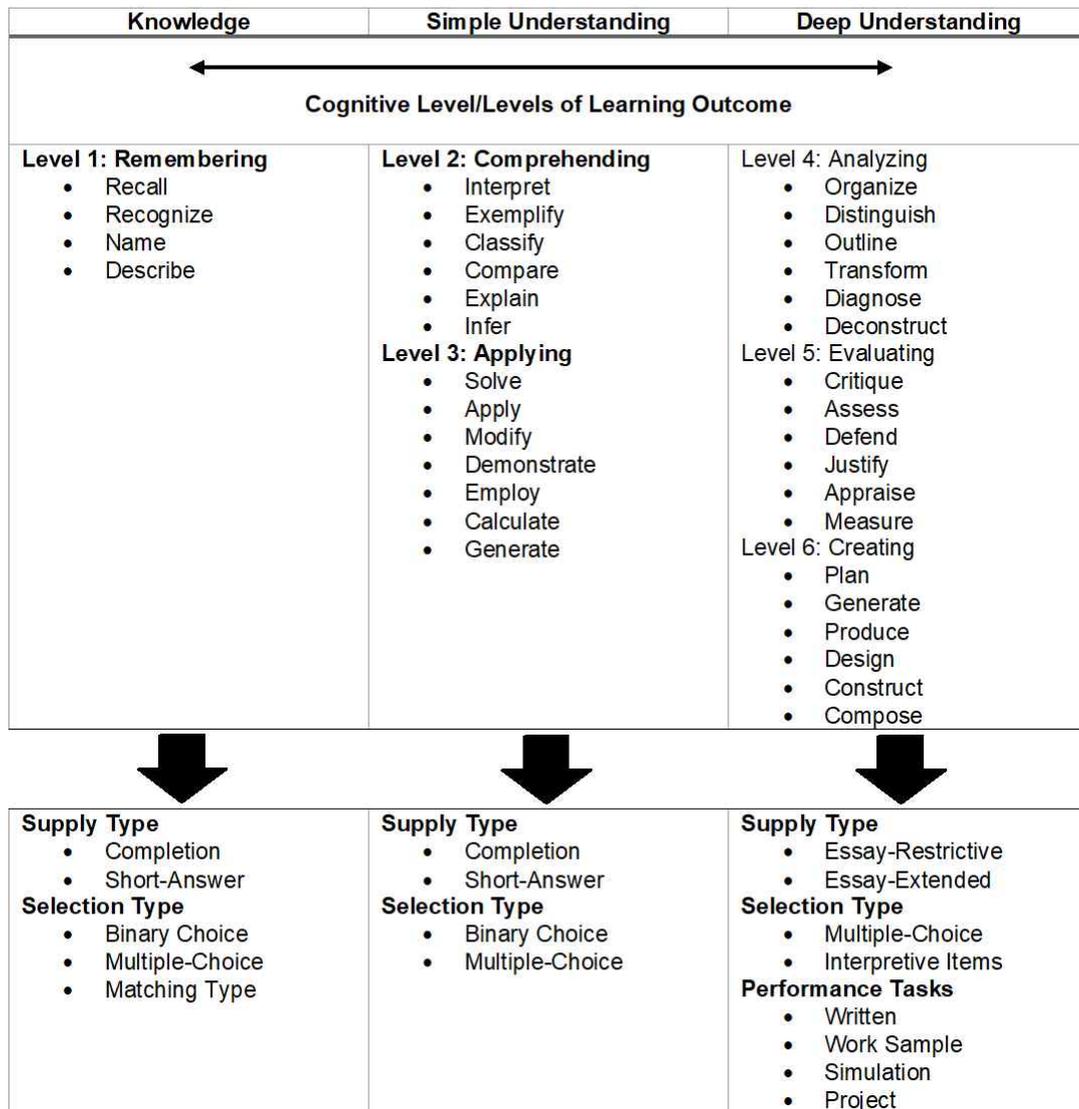
Targets	Assessment Methods					
	Selected-response & brief-constructed response	Essay	Performance Tasks	Oral Questioning	Observation	Student Self-Assessment
Knowledge & Simple Understanding	5	4	3	4	3	3
Deep Understanding & Reasoning	2	5	4	4	2	3
Skills	1	3	5	2	5	3
Products	1	1	5	2	4	4
Affect	1	2	4	4	4	5

벤자민 블룸(Benjamin Bloom)은 1956년에 교육 목표를 체계적으로 분류하고 정의하기 위해 교육목표분류체계를 정의했다. 교육과 학습의 목표를 명확히 하고, 교육과정과 평가 방법을 설계하기 위해 인지적 영역(Cognitive Domain)에 중점을 두며 지식(Knowledge), 이해(Comprehension), 적용(Application), 분석(Analysis), 종합(Synthesis), 평가(Evaluation)의 여섯 가지 단계를 제안했다(Bloom, 1956). 하지만 이러한 지식에서부터 시작되는 누적적인 단일 체계는 실제 적용에 대한 한계가 있었고, 블룸의 교육 목표 분류체계의 문제점을 개선하여 기억(Remembering), 이해(Comprehending), 적용(Applying), 분석(Analyzing), 평가(Evaluating), 창안(Creating)으로 수정된 신교육목표분류체계가 등장했다(Krathwohl, 2002)([그림3-1]³⁾).

3)출처: BLOOM'S TAXONOMY. UNIVERSITY of FLORIDA. 2024.05.23. <https://citt.ufl.edu/resources/the-learning-process/designing-the-learning-experience/blooms-taxonomy/>



McMillan은 신교육목표분류체계의 ‘이해’를 ‘단순한 이해’와 ‘심층적 이해’로 세분화한 뒤 가장 낮은 수준인 ‘기억’과 함께 인지의 수준을 설정했다. 그 후 인지 수준별 평가 유형과의 관계를 [그림 2-2]과 같이 제시했다. [그림2-2]에서는 한 종류의 평가 유형이 한 가지 인지 수준에 국한되거나, 또 다른 평가 유형과 배타적이지 않고 다양한 인지 수준 결과를 확인할 수 있게 제작될 수 있음을 알 수 있다.



[그림 2-2] 기억, 이해 수준과 평가 유형의 관계(McMillan, 2014)

2.1.2. 서술형 평가의 의미와 도입 필요성

서술형 평가(Descriptive assessment)는 서술형 문항을 활용하여 평가자가 학습자가 구성하거나 서술한 것에 대해 평가적인 판단을 내리는 방식을 총칭하는데, 이 용어는 정책적으로 도입된 것이다. 따라서 서술형 문항의 경우 전통적인 문항 분류에 포함되지 않기 때문에, 앞서 분류한 문제 유형과의 관계를 명확히 하고 그 속성을 정의할 필요가 있다. 서술

형 문항을 정의하는 합의된 것은 없으나 기술 요소, 답안의 길이, 응답 방식 및 범위가 제한되어 한 문장 이상의 분량을 요구하는 문항으로 보거나, 대체로 교사들은 개념의 요약, 정리, 사실의 나열, 풀이 과정 등 어느 정도 정답이 정해진 문항으로 간주하는 것으로 사료된다. 서술형 평가는 요구 내용이 포함된 질문이나 문두와 채점기준으로 구성된다. 필요 시에는 답을 하기 위해 필요한 자료가 포함되기도 한다. 이러한 특징에 미루어볼 때 Gronlund(1998)의 분류에 따라 서술형은 제한된 논술형(restricted response essay item), 논술형은 확장된 논술형(extended response essay item)으로 볼 수 있어 서술형 문항은 논술형 문항과 동등한 지위에 있는 문항 유형으로 볼 수 있다. 즉, 학생이 진술해야 할 용소의 가짓수, 문장 수나 글자 수 등을 제한하거나 서술해야 할 내용 범위나 도표에 그리기 등 서술 양식에 제한을 두는 경우를 서술형 문항이라 할 수 있다. 이에 반해 논술형 문항은 지식, 기능과 같은 자원을 복합적이고 종합적으로 활용하여 문제를 해결하는 것을 요구하는 문항 유형이라는 점에서 서술형과 차별된다(박혜영, 2019)⁴. 이러한 논의를 표로 정리하면 [표2-3]과 같다.

[표 2-3] 서술형·논술형 문항 유형 비교(박혜영, 2019)

특성	서술형 문항 (제한된 논술형, restricted response essay item)	논술형 문항 (확장된 논술형, extended response essay item)
응답 길이	한 문장 이상	문단 이상의 한 편의 글
응답 양식의 제한성	기술해야 할 요소, 답안의 길이, 응답 방식 및 범위 제한	응답의 자유도 허용
행동 특성	분석력, 비판적 사고력, 문제해결력, 창의력 등의 고등사고능력	고등사고능력의 복합적 활용, 즉 정보, 지식, 사고 등 자원을 복합적 활용, 통합, 종합하는 능력
글의 구성력, 완성도	서술의 깊이와 넓이, 정확성 등 요구	한 편의 글이나 주장의 완결성, 완성도, 논리적 글쓰기 요구

‘선택형 문항’과 서술형 문항이 포함된 ‘구성형 문항’의 두 유형의 문

4) 서술형 문항의 도입 및 개념화에 대한 자세한 논의를 확인해볼 수 있음

항을 본질적으로 동일하게 평가하므로 시간과 비용의 절감을 위해 선택형 문항을 선호해야 한다(Thissen, 1993), 두 문항의 유형은 blooms의 학습 목표분류체계 중 일부(기억, 이해, 적용, 분석) 영역의 지식은 동일한 수준으로 측정 가능하다는 관점(Kastner, 2011)도 있다. 실제로 채점과 피드백을 빠르게 제공할 수 있어 교육현장에서는 매력적인 ‘선택형 문항’을 만들기에 노력을 기울였다(Bleske-Rechek., 2007).

하지만 선택형 문항은 학생의 창의적인 아이디어와 사고를 포착하기 어렵고 암기나 확인 위주의 학습 방식을 고착화시켜 주체적인 학습 참여, 토론과 논술을 통한 사고의 증진, 능동적이고 탐구적인 학습을 제한하는 한계점이 존재한다(박혜영, 2019). 반면 서술형 평가를 ‘구성형 문항’의 ‘논술형 평가(Essay tests)’ 중 ‘제한된 논술형(Essay-Restrictive)’으로 간주할 수 있다면, [그림 2-2]에 따라 학습자의 ‘심층적 이해’를 평가하기에 적합하다. 서술형 평가는 응답 특성과 문제 해결 과정을 더 자세히 살펴볼 수 있어 학습자의 개념 이해 정도 진단에 효과적이며, 제시된 문항속의 정보로 인지 구조 속의 정보들을 회상해 낸 뒤 조합하여 설명을 구성하는 방법을 활용하므로 사고를 촉진시키며, 학습을 점검하고 부족한 부분을 판단하는데 활용되어 개념 변화에 효과적(하민수, 2019)이며 지식의 확인과 이해를 넘어 학생이 주체적인 해석과 적용, 지식의 소통 능력을 길러 주어(박혜영, 2019) 비판적·창의적·통합적 사고력, 고차원적 분석력, 문제해결력, 의사소통 능력과 같은 역량 교육이 가능하다(박혜영, 2019; 박종임, 2022). 과학 교육 연구자(Lane, 2004; Shepard, 2000)와 차세대 과학 표준(National Research Council, 2012)은 과학적 설명이 과학교육에서 중요하므로 문제 공식화, 조사 수행, 결과 전달과 같은 능력 평가를 위해 서술형 평가는 필수적이다(Liu, 2014). 부가적으로 학습자들은 선택형 평가보다 서술형 평가를 열심히 준비하고(Snow, 1993), 특정 유형의 문제로만 평가할 경우 다른 유형의 평가에 강점을 갖는 학생이 불이익을 받을 수 있다(Lissitz, 2012).

즉, 서술형 평가는 현재 교육현장에서 주요하게 사용되는 선택형 문항이 측정할 수 없는 학습자의 역량을 평가하고 확장한다는 점에서 그 중

요성이 강조된다.

2.1.3. 서술형 평가 도입의 어려움

대규모로 수행되는 평가에서는 수행 과정의 편의와 채점의 타당도 및 신뢰도 확보를 위해 주로 선택형 문항이 구성형 문항보다 선호된다. 서술형 평가가 가지는 분명한 교육적 장점이 있음에도 위와 같은 현상이 나타나는 이유는 첫째, 다수의 응답을 채점하는 과정에서 시간과 노력이 많이 들며 채점 신뢰도를 확보하는 것이 어렵기 때문이다. 채점자가 답안을 읽고 점수를 부여하는 행위 사이에 해석, 비교, 판단, 종합 등의 고차원적 사고를 거치게 된다(박종임, 2022). 채점자 자격 교육 또는 2인 이상의 복수 채점을 통해 신뢰도를 확보하려고 하나, 이는 개인이 부담하는 노력과 채점량이 많아짐을 의미한다. 둘째, 타당한 채점 기준을 개발하고 유사 정답 및 오답 요소를 고려하는 데 많은 시간과 노력이 소요되어 양질의 문항을 제작하는 것이 어렵다. 특히 수식이나 계산이 아닌 텍스트 위주의 답안의 경우 수많은 유사 답안이 등장한다. 또한 단위 학교나 교육기관에서 교육 평가 이후 행정적 처리 마감기한과 같은 시간적 제약으로 인하여 문항의 다양성과 창의성보다는 채점의 용이성을 고려할 수밖에 없다. 셋째, 서술형 평가는 즉각적인 진단과 피드백이 어렵다. 채점에 많은 시간이 소모되는 만큼 학습자에 대한 진단과 피드백이 늦어지고, 이는 서술형 평가의 교육적 효과가 시간에 따라 줄어드는 것을 의미한다(Livingston, 2009; 박혜영, 2019; 박종임, 2023). 위와 같이 서술형 평가는 채점의 일관성, 객관성, 공정성의 문제와 채점에 소요되는 시간과 비용의 문제가 존재한다.

2.2. 자동채점(Automated Scoring)

서술형 평가 도입에서 가장 어려운 점은 채점이다. 문항의 채점은 매번 채점자의 판단에 따라 상당한 시간과 노력을 필요로 하는데, 채점자 특성, 채점 피로도, 채점 시기 등의 기타 제반 요인에 영향을 받을 가능성 또한 잠재되어 있다(노은희, 2016). 이와 같은 한계를 컴퓨터 기술을 활용하여 극복하는 방법이 ‘자동채점’이다. 일반적으로 텍스트로 구성되어 있는 답안을 목표로 하며, 국내의 경우 ‘서술형’, ‘논술형’, ‘서답형’, ‘작문’, ‘글’ 등으로 다양한 용어가 사용되나, 영어권에서는 ‘논술형(Essay)’의 정의로 ‘서술형’과 ‘논술형’을 포함한 한 문장 이상의 글을 지칭하기 때문에, 자동채점을 목표로 하는 문항이 ‘단답형’과 ‘논술형(에세이형)’로 나뉜다(박종임, 2022)([표 2-4]).

[표 2-4] 자동채점의 유형과 특징(박종임, 2022)

특징 \ 구분	단답형 자동채점	에세이형 자동채점
채점 대상 응답 길이	• 어휘, 구 또는 문장	• 문단 이상 한 편의 글
채점 상의 특성에 따른 기술적 쟁점의 차이	• 정답이 명확하므로 정답과의 비교를 통해서 채점함. • 채점 정확성 > 채점 결과의 해석 가능성	• 정답이 존재하지 않기 때문에 정답과의 비교가 아닌 채점 자질(feature)에 근거하여 채점함. • 채점 정확성 < 채점 결과의 해석 가능성 • 기계학습 채점 자질과 알고리즘의 해석 가능성 고려 필요

상대적으로 자동채점 설계에 많은 요소와 절차가 고려되어 도전적인 과제안 국내·외 모두 논술형 자동채점에 대한 연구가 주로 이루어지고 있다. 본 연구에서 [표2-3]에 따라 ‘논술형 문항’ 중 ‘제한된 논술형 문항(이하 서술형 문항)’에 초점을 맞춰 연구를 진행하였다.

2.2.1. 내용 기반 자동채점(Content-Based automated Scoring)

응답자가 주어진 문항에 대해 논리적이고 설득적으로 작성한 글인 ‘확

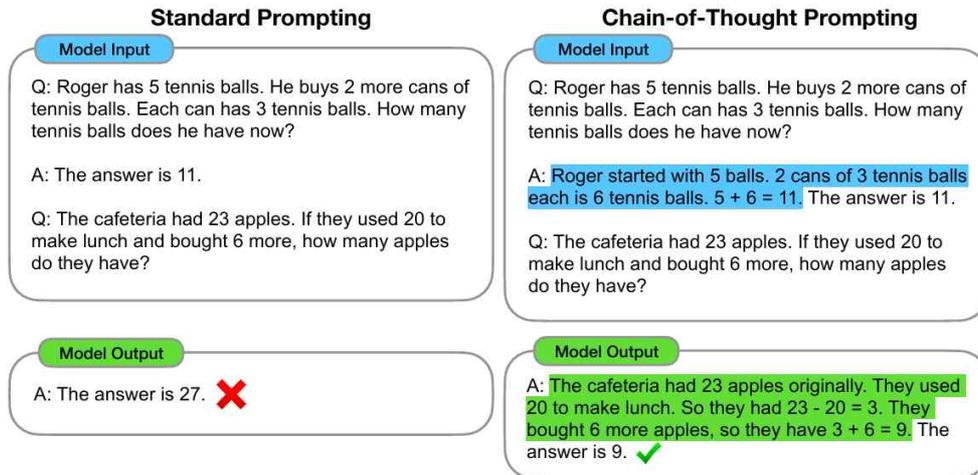
장된 논술형’은 글의 완성도와 창의성을 채점한다. 명확하게 채점요소를 확정할 수 없기 때문에, 연구 초기부터 지금까지 채점 자질을 탐색하는 것이 주요 과제이다. 반면, 기술해야 할 요소와 응답 방식등이 제한되어 있는 ‘서술형 문항’은 정답 요소를 설정할 수 있고, 서술의 깊이와 정확성을 중점으로 두는 목적의 문항이다. 본 연구에서는 컴퓨터 과학, 수학, 생물학 등과 같이 특정 과목에서 학생의 학습 개념에 대한 지식 및 이해 수준을 평가하기 위한 개방형 서술형 문항에 대해 다룰 것이다. 이러한 문항은 영어권에서 서술형(essay) 문항과 구별하여 내용 기반 문항(Content-Based response), 단답형 문항(Short answer), 개방형 문항(Open-Ended Question), 서술형 문항(Descriptive item) 등으로 다양하게 번역되어, SciEntsBank, Beetle, Texas2011, ASAP-SAS 또는 새롭게 수집한 데이터 세트를 바탕으로 답안을 자동채점하기 위한 연구가 진행되고 있다(Haller, 2022). 내용 기반의 자동채점 전략에는 전문가(교사)가 제공한 답안과의 유사성을 바탕으로 채점하는 참조 기반 접근(reference-based approach)(Alfonseca, 2004; Nielsen., 2008; Meurers, 2011; Sukkarieh., 2011; Horbach., 2013; Pado, 2015)과 인간 채점자가 채점한 결과를 이용하여 모델을 훈련하는 응답 기반 접근(response-based approach)(Mohleretal., 2011; Dzikovska, 2013; Ramachandranetal., 2015; Zesch., 2015; Zhu., 2016)가 있다(Madnani., 2017). 인간 채점자가 채점한 데이터가 충분하다면 응답 기반 접근 방식이 정확도가 높으나(Sakaguchi, 2015), 정교한 채점 기준을 개발하고 일관성, 신뢰성이 높은 데이터를 수집하고 적용하는 것에 한계가 있다.

2.2.2. 생성형 AI(Generative Artificial Intelligence)

사회의 다양한 분야에서 인공지능(AI)를 활용한 연구가 추진되는 가운데, 생성형 AI(Generative Artificial Intelligence) 기술의 등장은 다양한 연구분야의 혁신적인 도구로 사용되고 있다. 교육학 분야에서도 교육 방법이나 교육 환경에 적용하는 관점에서 ‘도구로서의 AI’를 사용하는

방안에 대해 활발히 연구되고 있다(홍선주 외, 2020). González는 AI를 가장 활용할 수 있는 분야가 교육이라고 말하며, 튜터링(Tutoring), 평가(Assessment), 기타(데이터 관리, 특수교육 등)로 그 방법을 제시하였다. 특히 평가 측면에서 AI는 개인 맞춤 진단 및 피드백(Dai, 2023)과 함께 양적, 질적 평가로 학습자의 성취를 판단하는 도구로 사용될 수 있음을 언급했다(González, 2021; Jeon, 2023). 즉서·논술형 평가를 현장에 도입했을 때 예상되는 문제점을 생성형 AI를 활용한 자동채점 기술개발을 통해 해결하려는 시도가 이루어지고 있다.

ChatGPT로 대표되는 대규모 텍스트로 사전 학습된 생성형 AI는 대형언어모델(Large Language Model)로 불리며 자연어 처리 및 생성 기능이 뛰어나다. 사용자가 작업을 요청하기 위해 작성하는 입력 텍스트를 프롬프트(Prompt)라고 한다. 생성형 AI는 프롬프트를 입력받고, 관련성이 매우 높은 결과를 반환하는 대화형 모델(챗봇)이다. 이때 프롬프트가 맥락이 있고 명확하며 예시가 포함되는 등 잘 설계될수록 생성형 AI의 성능이 최적화되므로, 사용자는 자신의 의도와 모델의 이해 능력 사이의 격차를 고려하여 프롬프트를 정교화하는 작업이 필요한데, 이를 프롬프트 디자인(Prompt Design) 또는 프롬프트 엔지니어링(Prompt Engineering)이라고 한다. 두 용어는 사용자의 목적에 따라 세부적으로 약간의 차이가 있으나, 자동채점의 관점에서 공통적으로 사용될 수 있다고 판단된다. 프롬프트 엔지니어링을 위한 연구도 활발히 진행되어 프롬프트 작성법에 대한 지침이 제안되고 있다(Ekin, 2023; Bsharat, 2024). 그 중 가장 효과적인 것은 Wei가 제안한 ‘생각의 사슬(Chain-Of-Thought)’이다. Wei는 대형언어모델에 예시 및 사고과정을 프롬프트에 입력해주는 생각의 사슬 방법을 활용하여 그 성능을 향상시킨 결과를 발표했다(Wei, 2023)([그림 2-3]). 이는 자동채점 과제에서도 유의미하고, 후술할 선행 연구들의 연구방법에 적극적으로 도입되었다.



[그림 2-3] Chain-of-thought Prompt를 활용한 추론 과제 성능 향상의 예시(Wei, 2023)

2.2.3. 자동채점 연구 동향 및 한계점

자동채점 분야 연구 초기에는 단어의 갯수를 찾고 중요도를 평가하는 통계적 기법으로 채점을 했다. 이후 기계를 학습시켜 분류하는 방법이 등장함에 따라 신뢰도와 정확도가 비약적으로 상승했다. 서술형 문항의 자동채점은 크게 채점 자질을 직접 설정하여 로지스틱 회귀(Logistic regression) 또는 서포트 벡터 머신(Support Vector Machine)과 같은 고전적인 머신러닝(machine learning) 방법, 채점 자질까지도 기계가 설정하는 인공신경망 기반의 딥러닝(deep learning) 방법으로 나눌 수 있다(Haller, 2022). 영어권에서 자동채점은 1966년에 시작된 Project Essay Grader(PEG)로 시작되었다(Ajay, 1973). 주로 교육연구기업들이 주도하여 에세이(essay)로 통칭하는 긴 글 수준의 에세이의 작문 능력을 채점하기 위한 대한 연구(Automated Essay Scoring)가 활발히 진행되었는데, Intelligent Essay Assessor(IEA)를 개발한 Pearson, IntelliMetric를 개발한 Vantage Labs, Rudner가 제안한 BETSY(Bayesian Essay Test Scoring System)등이 있다. 그 중 비교적 짧은 글인 서술형 답안에 대한

채점 모델로는 ETS(Educational Testing Service)가 개발한 E-rater(Electronic Essay Rater)와 C-rater가 대표적이다. 그 중 C-rater은 단답형 답안을 채점하는 모델로, 응답자의 답안을 자연어 처리를 통해 표준화한 형태로 바꾼 뒤 정답과 동일한 의미를 갖는지를 판단하도록 설계되어 교과목의 이해 정도를 평가할 수 있는 프로그램이다(Leacock, 2003; Sukkarieh, 2009). 이러한 C-rater은 미국의 NAEP(National Assessment of Educational Progress) 수학과 문제 논리 시험, 인디애나주 11학년 영어 독해 시험에서 사용되기도 했다. 이외에도 다양한 과목의 단답형 답안을 평가하기 위한 접근법과 기계학습 방법을 제안한 연구(Nielsen, 2008)처럼 기계학습 기술도 자동채점에 적극적으로 도입됨에 따라 단어가방(bag of words)을 바탕으로 SVM(Support Vector Machine)을 이용해 학습한 SIDE 모델을 제안하거나(Nehm, 2012), IBM의 통계 소프트웨어 패키지 SPSS를 이용하여 답안 텍스트를 분석해 핵심 용어를 식별하는 연구(Weston, 2013), 온라인으로 사용 가능한 진화론에 대한 학생들의 응답 채점 결과를 이용해 학습시킨 모델 EvoGrader을 제안하는 연구(Moharreri, 2014), 연속된 단어 n-gram을 이용해 벡터화된 답안을 여러 로지스틱 회귀 모델에 학습시켜 이를 결합하여 만든 스택킹을 활용한 연구(Heilman, 2013), 단어 임베딩 기술을 보완하기 위해 자질 추출과 단어 임베딩을 결합하는 방법으로 전이 학습을 제안한 연구(Roy, 2016), 텍스트 통계, 어휘 유사도, 문장 의미 유사도, 단어 중요도 등의 자질을 랜덤포레스트와 Extreme Gradient Boosting을 사용해 학습한 연구(Lukas, 2019), MOOC에서 단답형 질문 채점을 위해 Word2Vec을 이용해 임베딩한 단어들의 가중치를 tf-idf로 계산하여 랜덤포레스트로 분류한 연구(Kumer, 2019), 학생과 모델의 단어 공통 수를 기반으로 데이터 마이닝 하고, 클러스터링을 통해 유사 답안을 묶어 피드백을 제공하는 방법을 제안한 연구(Suzen, 2020) 등이 있다. 인공지능망을 활용한 딥러닝 기술이 발달하며 나타난 자동채점 방법은 생물학의 신경망을 흉내낸 네트워크 방식을 이용한다. 많은 노드로 이루어진 입력층, 은닉층, 출력층을 활용하여 조정 가능한 매개변수를 통해 비선형 관

계를 모델링 할 수 있는 장점이 있어 텍스트 데이터를 효과적으로 처리할 수 있다. 단어를 임베딩하는 기법으로 대표적인 Word2Vec은 유사한 단어를 잠재적인 공간에서 벡터로 표현하는 것을 목표로 하여 학습된 문장의 정보를 특성화할 수 있다. 다음으로 시퀀스 데이터에 적합한 RNN과 LSTM은 긴 거리의 단어 및 문장간의 관계를 파악하여 보다 정확한 성능을 보였다. 이후 새로운 패러다임으로 트랜스포머(Transformer) 기반 언어 모델인 BERT와 같은 사전 학습된 모델이 등장하며 더 이상 자질을 입력하지 않고, 입력 받은 데이터를 양방향 문맥 이해를 통한 단어 임베딩을 통하여 학습하는 파인 튜닝 과정을 이용하여 자동채점하는 연구가 활발히 진행되었다. Word2Vec, Doc2Vec과 같은 모델은 대규모 코퍼스를 학습한 언어모델로 채점에 기여하는 가중치가 큰 단어를 찾아내기도 했다. 딥러닝을 활용한 연구는 CNN과 LSTM을 결합시켜 단답형 자동채점의 성능을 비교하는 연구(Riordan, 2017), 자질 추출과 LSTM 학습을 결합하여 모델의 성능을 더욱 향상시킨 연구(Saha, 2018), 설정한 언어적 자질(단어 수, 답의 길이 등)을 포함하여 단어를 임베딩하여 LSTM을 학습시켜 컴퓨터 과학 분야 서술형 문항을 채점한 연구(Prabhudesai, 2019), 일본어를 영어로 번역하여 BERT 모델을 이용해 비지도 학습한 연구(Lun, 2020), 단어를 고차원 벡터에 임베딩한 Word2Vec을 기반으로 문장단위의 임베딩을 통해 답안의 유사성을 평가하는 모델 Sent2vec을 제안하는 연구(Gomma, 2020), 직접 정의한 자질과 Word2Vec으로 임베딩한 단어를 BERT에 학습시켜 자동으로 채점하는 모델을 제안한 연구(Beseiso, 2020), 교사와 학생 답안을 임베딩하여 맨해튼 거리가 최소가 되도록 LSTM을 학습시켜 의미적 유사도를 구한 뒤 예상 점수를 반환하는 MaLSTM을 제안하는 연구(Tulu, 2021) 등이 있다.

한편 국내연구로 한국교육과정평가원에서 꾸준히 자동채점 프로그램 연구가 진행되고 있다. 진경애(2008)는 5가지 유형의 영작문을 문단 단위 수준으로 채점할 수 있는 IntelliMetric 모델을 제안했고, 웹서버를 구축하여 활용 가능성을 높였다. 노은희(2016)는 문장 내 단어의 중요도를

수치화하여 자질을 설정한 후, 인간 채점결과를 학습시킨 여러 분류모델의 앙상블을 통하여 문장 수준의 서답형 문항을 자동채점할 수 있는 모델을 제안했다. 이는 기관 수준에서 이루어진 내용 기반 자동채점 모델으로써, 한국어 답안을 채점했고 자동채점 과정에 대한 프레임워크를 제안했다는 점에서 국내 연구에 유의미한 족적을 남겼다. 학생 답안의 자연어 처리 과정을 거친 후, 단어를 n-gram 자질로 설정하여 Logistic Regression, Nearest Centroid, AdaBoost, kNN, Random Forest, SGD와 같은 분류모델을 지도, 비지도 학습시키는 방법을 활용했다. 그 후 성능이 좋은 분류모델을 조합하여 보다 신뢰성 높은 분류결과를 획득하는 구조를 제안했다. 한국어 자연어 처리 기술이 부족하여 답안 인식부터 보조채점까지 인간의 개입을 통해 이를 보완하였다. 박종임(2023)은 38개의 채점 자질을 설정하고, 이를 바탕으로 한국어 논술형 문항을 자동채점하고, 답안에 포함된 단어를 행렬로 변환한 문서-단어 행렬(Document-Term Matrix)을 이용해 수학 교과에서의 서술형 문항을 자동채점하는 연구를 2차 연도 연구까지 수행했다. 채점 결과의 해석 가능성을 확보하기 위해 랜덤 포레스트, 회귀모형과 같은 분류 모형을 학습시키는 방법을 사용했다.

개별 연구자 수준에서도 자동채점에 대한 연구가 활발히 이루어지고 있으나, 박혜영(2019)이 주장한 것처럼 서·논술형 문항이 혼용되어 쓰이고 있고, 주로 한글 에세이(쓰기) 평가(박강윤, 2021; 조희련, 2021; 이용상, 2022)에 관한 연구가 진행되고 있다. 본 논문에서 초점을 맞추고 있는 내용 기반 자동채점 연구 사례 중 머신러닝을 활용한 연구는 학생들의 답안 코퍼스를 구축한 뒤 입력된 문장의 코사인 유사도를 통해 채점하는 LSA 방식을 제안하는 연구(정상목, 2005), 키워드를 포함하여 텍스트, 조합 자질 등 설정하여 결정 나무(Decision Tree), Maximum Entropy, Support Vector Machine을 이용해 기계학습 분류기에 적용한 연구(허정만, 2013), 자연선택 개념 문항에 학생들의 답안을 채점하기 위해 한국어 답안을 기계로 번역한 뒤 EvoGrader을 학습시켜 자동채점의 가능성을 평가하는 연구(하민수, 2016), 31개의 과학 서술형 평가 문항에

대한 채점 결과를 설정한 채점 준거의 존재 여부를 자질로 설정하여 분류 모형인 랜덤 포레스트에 학습시킨 연구(하민수, 2019), 다수의 답안에서 반복적으로 나타나는 용어를 핵심용어로 설정하여 이를 기반으로 비지도 학습 과정을 거쳐 개발한 평가와 피드백을 제공하는 목적의 채점 지원 도구 SAAI 제안한 연구(하민수, 2023)가 있다. 상대적으로 국내에서 딥러닝을 활용한 연구는 드문데, 그 사례로는 초등학교 5학년 학생들이 응답한 과학 서술형 평가 결과를 순환신경망(RNN)에 학습시킨 연구(박세진, 2020), 텍스트 기반의 문항은 아니나 학생들이 그린 그림 또는 그래프를 합성곱 신경망(CNN)에 학습시킨 자동채점 모델을 제안한 연구(Lee, 2022; 한재상 2023), 수학, 과학분야에서 논증과 같은 언어 역량을 측정하기 위한 서술형 문항을 자동채점하기 위한 트랜스포머 아키텍처를 학습 방법을 제안한 연구(배영우, 2024) 등이 있다.

자연어처리와 유사어 식별은 자동채점 연구에서 가장 중요한 문제이다. 방대한 양의 데이터가 쌓이며 딥러닝 기술이 발전하고 트랜스포머와 같은 모델이 더욱 발전하여 AI의 하위 분야인 자연어 처리(Natural Language Processing) 기술이 비약적으로 발달했다. 이로 인해 ChatGPT로 대표되는 생성형 AI(Generative Artificial Intelligence)가 등장했다. 생성형 AI는 대량의 텍스트 데이터를 학습한 사전 학습된 모델로, 별도의 학습의 과정을 거치지 않거나 미세 조정(Fine tuning)을 통해 채점 도구로 활용될 수 있다. 이러한 대형 언어 모델(Large Language Model)의 자연어 처리 및 과제 수행능력을 바탕으로 코드 생성, 이미지 생성, 학습 피드백 등 교육 분야에 다양하게 활용하려는 움직임이 생겼고(Kuchemann, 2023), 자동채점 연구 또한 활발히 진행되고 있다. 대형 언어모델은 오타와 어색한 표현 및 답변과 무관한 요소를 식별하여 평가 및 피드백을 위한 강력한 도구로 사용할 수 있다(하민수, 2023).

대형 언어 모델 또한 딥러닝 모형이므로 딥러닝을 활용한 연구로 볼 수 있으나 본 논문에서 강조할 학습에 필요한 많은 데이터를 요구하지 않는 점을 큰 장점으로 생각할 때 생성형 AI를 활용한 자동채점으로 새롭게 분류한다. 생성형 AI를 활용해 내용 기반 자동채점을 수행한 연구

는 다음과 같다.

영어권에서는 ChatGPT를 이용하여 컴퓨터 공학 분야 개방형 질문 채점 결과를 전문가의 채점결과와 비교하는 연구(Pinto, 2023), 학사 수준의 독일어 및 석사 수준의 영어 시험에서 답안 간 유사성을 통해 서술형 답안을 ChatGPT로 채점하는 연구(Schneider, 2023), ChatGPT의 2가지 버전에 대해 다양한 분야(의학, 생물학, 심리학)의 서술형 답안을 채점시키는데, 예시를 부여하는 원샷 학습 또는 예시가 없는 제로샷 학습 2가지로 나누어 그 성능을 평가한 연구(Chang, 2024), GPT-4를 활용해 예시나 채점기준을 생성하도록 프롬프트를 바꿔가며 서술형 문항 자동채점 성능을 평가한 연구(Jiang, 2024) 등이 있다.

한편 국내에서는 한국교육과정평가원에서는 ‘인공지능 챗봇의 학생 평가 결과 분석 및 활용 방안 탐색(백종호, 2023)’에서 ChatGPT의 교육 평가 활용 방안에 대한 연구를 진행하였다. 과학 교과를 기준으로 평가 상황에서 챗봇 기술을 활용하여 결과를 도출하고 결과에 대한 교사의 인식을 조사했다. 단위 학교의 수행평가에 사용하는 것과 유사한 형태의 8개의 과학 서술형 문항을 제작하고, 교사가 만든 채점 기준을 적용하거나 ChatGPT가 직접 채점 기준을 만들어 ChatGPT가 학생들의 응답을 채점하도록 했다.

개별 연구자로는 ChatGPT를 활용하여 평가기준, 예시 답안을 생성하여 세계지리 서·논술형 평가 결과를 평가한 연구(성정원, 2023), 직접적으로 서술형 평가에 사용한 것은 아니나, 분수 문제에 대한 학생의 응답과 평가 기준을 ChatGPT로 평가한 연구(손태권, 2023), 수학과 ‘확률과 통계’ 분야의 서술형 문제를 ChatGPT를 이용하여 채점하고 피드백을 생성하여 그 영향을 분석한 연구(김욱태, 2024) 등이 있다.

위에서 논의한 내용을 바탕으로 선행연구별 사용한 언어, 프로그래밍, 세부방법론을 내용은 [표 2-5]와 같다.

[표 2-5] 자동채점 기술별 내용 기반 자동채점 선행 연구

구분	관련 선행 연구	언어	프로그램명	세부방법론
머신러닝	Leacock(2003) Sukkarieh(2009)	영어	C-rater	자연어처리, 의미 유사도 평가
	Nielsen(2008)		SCIENTS BANK	교차 검증을 통한 분류 모델(SVM, Tree, boosting, ensemble) 평가
	Heilman(2013)		ETS	n-gram 벡터화를 통한 logistic regression
	Nehm(2012)		SIDE	SVM
	Weston(2013)		Modeler Text Analytics	SPSS 텍스트 분석
	Roy(2016)			logistic regression
	Lucas(2018)			랜덤포레스트, Extreme Gradient Boosting
	Kumar(2019)		AutoSAS	Word2Vec, 랜덤포레스트
	Suzen(2020)			텍스트 마이닝을 통한 공통 단어 수 기반 클러스터링
	Moharreri(2014)			
	하민수(2016)	한국어 → 영어	EvoGrader	Naive Bayes, SVM
	정상목(2005)	한국어		LSA를 이용한 답안의 코사인 유사도 측정
	허정만(2013)			키워드, 텍스트 자질 설정을 통한 결정 나무, SVM
	노은희(2016)			키워드 출현 여부를 통한 분류 모형 학습
	하민수(2019)		WA ³ I	채점 준거(키워드) 존재 여부에 따른 랜덤포레스트 모델 학습
	하민수(2023)		SAAI	다수 답안에 등장하는 핵심용어를 기반으로 한 비지도학습

	박종임(2023)			문서-단어 행렬을 이용한 랜덤 포레스트, 회귀모형 학습
답러닝	Riordan(2017)	영어		LSTM과 CNN을 결합한 자동채점 모델
	Saha(2018)			자질 추출과 LSTM을 결합한 모델
	Prabhudesai(2019)			단어 임베딩 및 언어적 자질을 기반으로 LSTM 학습
	Gomma(2020)		Ans2vec	Word2vec을 문장 단위로 임베딩하도록 학습
	Beseiso(2020)			설정된 자질과 Word2vec으로 임베딩한 단어를 BERT로 학습
	Tulu(2021)		MaLSTM	답안 임베딩 후 맨해튼 거리를 기반으로 LSTM 학습
	Lun(2020)		일본어 → 영어	
	박세진(2020)	한국어		단어별 인덱스를 부여해 순환신경망(RNN) 학습
	Lee(2022) 한재상(2023)			그림 또는 그래프를 합성곱신경망(CNN)으로 학습
	배영우(2024)	한국어		트랜스포머 아키텍처를 학습
생성형 AI	Pinto(2023)	영어	ChatGPT	개방형 질문 자동채점
	Schneider(2023)	독일어, 영어		답안간의 유사성 평가
	Chang(2024)	핀란드어		제로샷, 원샷 등 조건을 바꾸며 자동채점
	Jiang(2024)	영어		프롬프트를 변경하며 자동채점
	백종호(2023)	한국어		채점기준 생성 및 자동채점
	성정원(2023)			평가기준, 예시답안 생성 및 자동채점
	손태권(2023)			학생의 응답 평가 및 피드백
김옥태(2024)				

머신러닝 기반의 분류모델을 학습하는 자동채점 기술은 정밀한 자질 설정이 이루어졌을 때 채점 결과 해석이 용이하고 지도학습 분류 모델을

사용하여 평가 프로세스를 자동화 및 표준화하여 모델의 객관성과 효율성을 높일 수 있다. 하지만 문항마다 채점 자질을 추출하고 학습시키는 과정이 필요하여 일반화할 수 없고, 모델이 신뢰 있는 결과를 도출하기 위해서는 인간 채점자가 평가한 양질의 다수 데이터를 필요로 한다는 단점이 있다. 또한 몇 개의 논문에서 제시한 키워드 중심의 채점 방식이나 문서-단어 행렬을 활용하는 방식은 맥락 파악 없이 단어 출현 여부로만 채점되기 때문에, 동음이의어를 인식할 수 없고(Ramesh, 2021), 절·구·문장간의 논리적 연결을 확인하기 어려우며, 오답 요소나 정답 요소와 관련 없는 단어들은 감점 요소임에도 불구하고 인식되지 않는다. 이를 해결하기 위해 WordNet을 이용한 동의어 반의어 검출 및 품사 태그를 통한 유사도 점수로 모델을 보완하려 했으나 그 정확도는 한계가 있었다.

딥러닝 기반의 모델은 양질의 많은 데이터를 학습할수록 인간 채점 결과와의 일치도가 높아졌다. 특히 사전 학습된 단어 임베딩 모델을 파인 튜닝하는 전이학습은 특정 분야의 자동채점에 높은 성능을 보였다. 하지만 다양한 분야에 활용 가능한 일반화된 모델 개발의 어려움(Haller, 2022), 블랙박스과 같은 은닉층으로 인하여 채점 결과를 해석하기 어렵다는 단점(Ramesh, 2021; 박종임, 2023)이 뚜렷하다. 이는 교육평가의 목적인 학습 과정 전반에 대한 지원과 개선에 적합하지 않다. 무엇보다 기존의 연구방법은 인간이 채점한 결과와의 일치도를 향상시키기 위해서 모델을 학습시킬 많은 데이터가 필요하다.

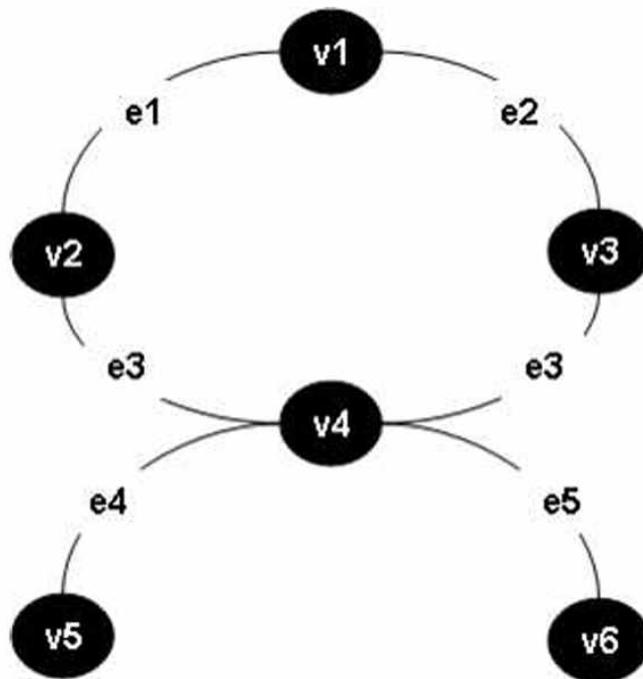
생성형 AI는 학습에 필요한 데이터가 작거나 없어도 된다는 점에서 기존의 모델을 학습시키는 연구에 비해 큰 장점이 있다. 하지만 생성형 AI를 이용한 자동채점은 응답이 가능한 범위나 수준의 측면에서 폭이 너무 넓어, 신뢰도와 타당성을 확보해야 하는 채점의 과정에 적용하기에는 한계가 있다는 의견도 있다(백중호, 2023). ChatGPT와 같은 사전 학습된 대형 언어 모델에 프롬프트 엔지니어링을 통해 자동채점 하는 방식은 직접 채점에 활용했을 경우 일관성이 다소 낮고, 평가 요소와 배점 기준이 복잡할수록 더욱 낮아지는 현상(이동원 외, 2024)이 나타난다. 또한 한계로 지적된다. 이를 보완하기 위한 공식화된 최적의 프롬프트가

없기 때문에 사용자가 과정 또한 부담이다. 뿐만 아니라 ChatGPT와 같은 생성형 AI가 제공하는 피드백은 수업 상황에서 교사와 학생에 대한 배경지식이 없어 일반화된 답변만 반환하고(하민수, 2023) 피드백의 질이 인간 채점자가 제공하는 것보다 부족하다는 문제도 제기된다(Dai, 2023; 백종호, 2023; 성정원, 2023). 대형 언어 모델(LLM)은 실시간으로 바로 사용 가능하지만, 국가간 학습 개념 또는 표현의 차이가 존재할 경우 잘못 채점할 경우(성정원, 2023)가 있기 때문에 완전히 의존할 수 없고 인간 채점자가 항상 같이 존재해야 한다(Schneider, 2023).

따라서, 머신러닝 자동채점 기술은 타당한 자질 추출과 많은 데이터를 필요로 한다는 한계, 딥러닝 자동채점 기술은 많은 데이터를 필요로 하고, 채점 결과를 해석하기 어렵다는 점의 한계, 생성형 AI를 이용한 자동채점 기술은 채점의 성능이 일정하지 않고 프롬프트 작성에 따른 변동성이 크다는 점의 한계가 있다. 이를 극복하기 위한 자동채점 기술은 교과, 문항 유형, 응답자, 데이터 수에 구애받지 않고 다양한 분야에 적용할 수 있는 범용적이고 일반화된 모델이어야 하고, 채점결과를 사용자가 신뢰있게 확인하고 피드백에 사용할 수 있는 타당성을 확보한 자동채점 기술이 필요하다.

2.3. 그래프 이론(Graph Theory)

그래프 이론은 수학자 오일러(Leonhard Euler)가 1736년에 발표한 이론으로, 개별 객체를 뜻하는 정점 또는 노드(node), 정점 또는 객체 사이의 관계를 설명하는 간선(edge), 노드에서 노드로 가는 간선들의 집합인 경로(path)로 구성된다([그림 2-4]). 수학, 컴퓨터 과학, 생물학 분야에서 다양한 문제를 그래프 구조로 모델링하여 해결하기 위해 활용된다. 인터넷 네트워크, 유전자 상호 작용, 데이터 구조, 알고리즘, 데이터 베이스, 네트워크 과학 등에 활용되는데, 소셜 네트워크 분석 추천시스템, 자연어 처리 등에 사용되기도 한다. 본 논문에서는 그 중 이분 그래프(bipartite Graph)를 이용한다.



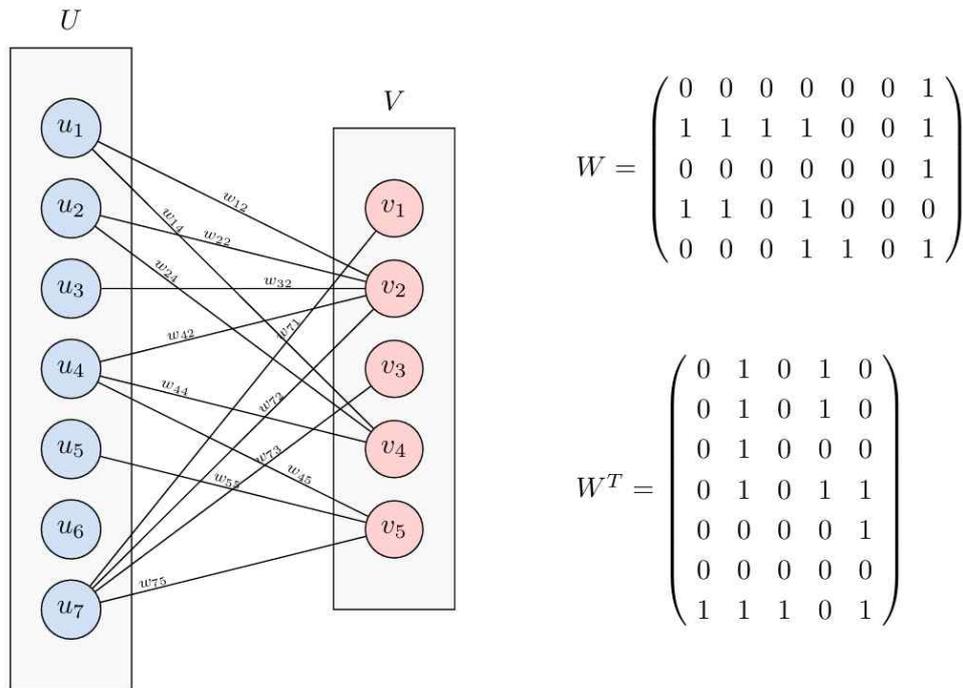
[그림 2-4] 노드(v1~v6)와 간선(e1~e5)으로 이루어진 그래프(Sanfilippo, 2006)

2.3.1. 이분 그래프(Bipartite graph)

이분 그래프는 네트워크가 2개의 집합으로 구분될 때 효과적인 표현 방법이다. 논문과 저자, 영화와 영화배우 등이 있다. 이분 그래프는 서로 다른 2개의 하위집합 내의 노드들과 노드들간의 관계를 간선으로 표현한다. 동일 하위집합 내의 노드들은 간선이 없다. 이분 그래프 $G=(U, V, E)$ 에서 U, V 는 노드의 집합, E 는 간선의 집합이고, $U=\{u_1, u_2, \dots, u_n\}$, $V=\{v_1, v_2, \dots, v_m\}$ 이다. 간선의 집합 $E=\{e_{ij}\}$ 으로 표현할 때 $|A|=n$, $|B|=m$ 이라면 $i=1, 2, \dots, n$, $j=1, 2, \dots, m$ 이고 e_{ij} 는 a_i 와 b_j 를 연결하는 간선이다. 간선에는 a_i 와 b_j 의 연결 강도를 나타내는 가중치 w_{ij} 가 있을 수 있고, 연결 유무에 초점을 둘 경우 연결된 간선은 $w_{ij}=1$, 연결되지 않은 간선은 $w_{ij}=0$ 으로 둘 수 있다. 이분 그래프 U 를 저차원으로 낮추고 수학적 계산을 하기 위해 $W=(w_{ij})$ 와 같이 인접 행렬 W 을 만들 수 있다. 정확히는 W 는 U 의 인접 행렬이고 V 의 인접 행렬은 $W^T \in \mathbb{R}^{m \times n}$ 이다([그림 2-5]5).

이분 그래프를 이용하여 다양한 과제를 해결하려는 연구가 이루어지고 있다. Li(2012)는 이미지의 인접 픽셀을 그룹화한 슈퍼픽셀과 이미지 패치를 노드로 한 이분 그래프 분할 접근법을 통해 패턴 인식 분야에서 이미지 분할 과제의 정확도 효율성은 높이는 방법을 제안했다. He(2017)는 이분 그래프의 구조를 통해 노드의 순위를 계산하는 Birank 알고리즘을 소개하는 연구를 제안했다. Pavlopoulos(2018)는 유전자-단백질 상호작용, 생물학 경로 분석과 같은 생물학 분야, 약물-표적 상호작용, 질병-유전자 연관성과 같은 의학 분야에서 이분 그래프의 활용을 통해 변인간의 복잡한 관계를 구조화하여 분석하는데 활용했다.

5) 이 예시에서는 연결된 간선은 가중치 1, 끊어진 간선은 가중치 0으로 두었다.



[그림 2-5] 이분 그래프와 인접행렬의 예시(Giamphy, 2023)

2.3.2. 그래프를 이용한 자동채점

이분 그래프를 포함한 그래프 이론을 자동채점에 활용하는 연구는 다음과 같다. Wang(2005)은 학생 응답의 아이디어를 평가하기 위해 학생 답안과 참조 답안을 이분 그래프로 구축한 뒤 개념 일치도를 바탕으로 창의성을 자동채점하는 방법을 제안했다. 구체적인 방법이 제시되어 있지 않고, 창의성과 같은 추상적인 개념을 수치화하는 과정에서 한계가 존재한다. 프로그래밍 언어에서 사용하는 변수, 숫자가 제한되어 있는 점을 이용하여 학생들이 작성한 코드를 그래프로 변환하여 참조 코드로 구축한 그래프와의 측정해 이를 기준으로 평가, 피드백을 제공하는 연구(Naude, 2010; Vujosevic-Janicic, 2013)가 있다. 하지만 변수를 다르게 정의 하거나 같은 기능을 다른 코드로 구현했을 때에는 인식하지 못하는 한계가 뚜렷하다. Kaur(2017)은 서술형 답안을 품사 태그를 부착한 후, 이를 바탕으로 그래프를 구축했다. 그리고 대규모 지식 그래프인

Concept map 데이터베이스를 이용하여 유사어를 정량적으로 식별하여 참조 답안과의 유사성을 점수로 반환했다. 그래프를 바탕으로 유사어까지 식별하려는 시도가 있었으나, 품사를 기준으로 비교하는 방식은 언어적 구조(품사)가 다른 유사 답안을 채점할 수 없다는 한계점을 갖는다. Nayak(2019)은 학생의 답안을 채점하는 연구는 아니지만, 다양한 버전의 생물학의 분류학적 계층 구조를 그래프로 구축하고, 그 유사성을 Katz 유사도로 측정하는 연구가 있다. 평가의 정확성과 효율성이 좋으나 이는 유사성이 매우 높은 소수의 그래프를 이용했다는 점에서 일반화하기 어렵다. Bagaria(2020)는 서술형 답안에서 키워드 점수와 개념 그래프 점수를 도입하여 자동채점을 하는 방법을 제안했다. 하지만 자세한 그래프 구축 방법이 제시되지 않았고, 노드와 간선이 완전히 일치했을 때만 개념 그래프 점수를 부여했다는 점에서 한계가 있다. Tan(2020)은 그래프 합성곱 신경망(Graph Convolutional Network)를 이용하여 학생의 단답형 응답을 자동으로 채점하는 방법을 제안했다. 학생들의 모든 응답을 단어, 문장 수준으로 나누고 단어 빈도에 따라 간선의 가중치를 계산하여 반영한 뒤 그래프를 구축하고 GCN을 이용하여 학습했다. 이를 바탕으로 학생들의 응답을 채점했다. 모든 답안을 학습시켰기 때문에 채점자가 설정한 자질이 아닌 답안 전체에서 중요 단어(또는 문장)간의 관계가 고려된다는 점이 장점이나, 딥러닝 기술이 활용되었다는 점에서 많은 데이터를 요구하고, 학습 시간이 오래 걸리며, 그 결과의 해석이 어렵다는 문제가 여전히 존재한다.

따라서, 기존의 그래프를 이용한 자동채점 연구는 다양한 분야의 서술형 문제로 일반화하기 힘들고, 유사어 식별 방법이 고전적이어서 성능이 떨어지며 이를 보완하기 위한 딥러닝 기술 활용시 결과 해석의 어려움이 수반된다. 본 논문에서는 위의 문제와 함께 기존의 머신러닝, 딥러닝을 활용하는 자동채점 모델의 한계점 또한 해결하기 위해 생성형 AI를 활용하여 그래프를 구축하고 자동채점 하는 방법을 제안하고자 한다.

3. 연구 방법

3.1. 연구 절차

본 연구에서는 내용 기반 서술형 문항의 자동채점 분야 선행연구에서 드러났던 문제점을 해결하기 위한 방법으로 그래프 이론을 이용한 자동채점 방법을 제안하고, 평가도구으로써 가치와 실제 교육현장에서의 활용 방법에 대해 논의한다.

먼저, 선행 연구를 분석한다. 현재 서술형 문항 자동채점 연구 분야의 발전 과정과 최신 연구 동향을 파악하고, 각 연구에서 제안된 모형 또는 모델의 장단점을 분석한다. 서술형 문항의 자동채점은 크게 채점 자질을 직접 설정하는 머신러닝을 이용하는 방법, 채점 자질까지도 기계가 설정하는 인공지능 기반의 딥러닝을 이용하는 방법으로 나눌 수 있다. 이때 딥러닝을 활용하는 방법 중 생성형 AI를 활용하여 자동채점을 수행하는 연구가 최근 활발히 진행되고 있다([표 2-5]). 또한 본 연구에서 제안하는 그래프를 이용해 자동채점하는 선행연구도 분석했다. 서술형 답안을 그래프로 구축하여 그 유사성을 바탕으로 채점한다는 점에서 Kaur(2017)과 Bagaria(2020)의 연구가 방법론이 비슷하기는 하나, 두 연구보다 그래프 구조를 세부적으로 비교하고 유사도를 검증에 발전된 모델과 방법을 사용한다는 점에서 차별성이 있다.

다음으로, 자동채점 모델을 구상하였다. 서술형 문항에 대한 응답은 숫자와 텍스트로 이루어져 있다. 논술형과 달리 그 문장이 매우 길지는 않지만, 채점자는 문장을 모두 읽고 채점 기준에 의거하여 채점해야 한다. 하지만, 응답을 노드와 간선으로 구성되어 있는 그래프로 구축한다면 숫자와 텍스트로 이루어진 문장을 압축하여 구조화할 수 있다. 같은 방법으로 모범 답안(참조 답안)도 그래프로 구축한 뒤, 학생들의 응답으로 구축한 그래프와 비교하여 그 구조를 비교할 수 있다. 이렇게 타당하게

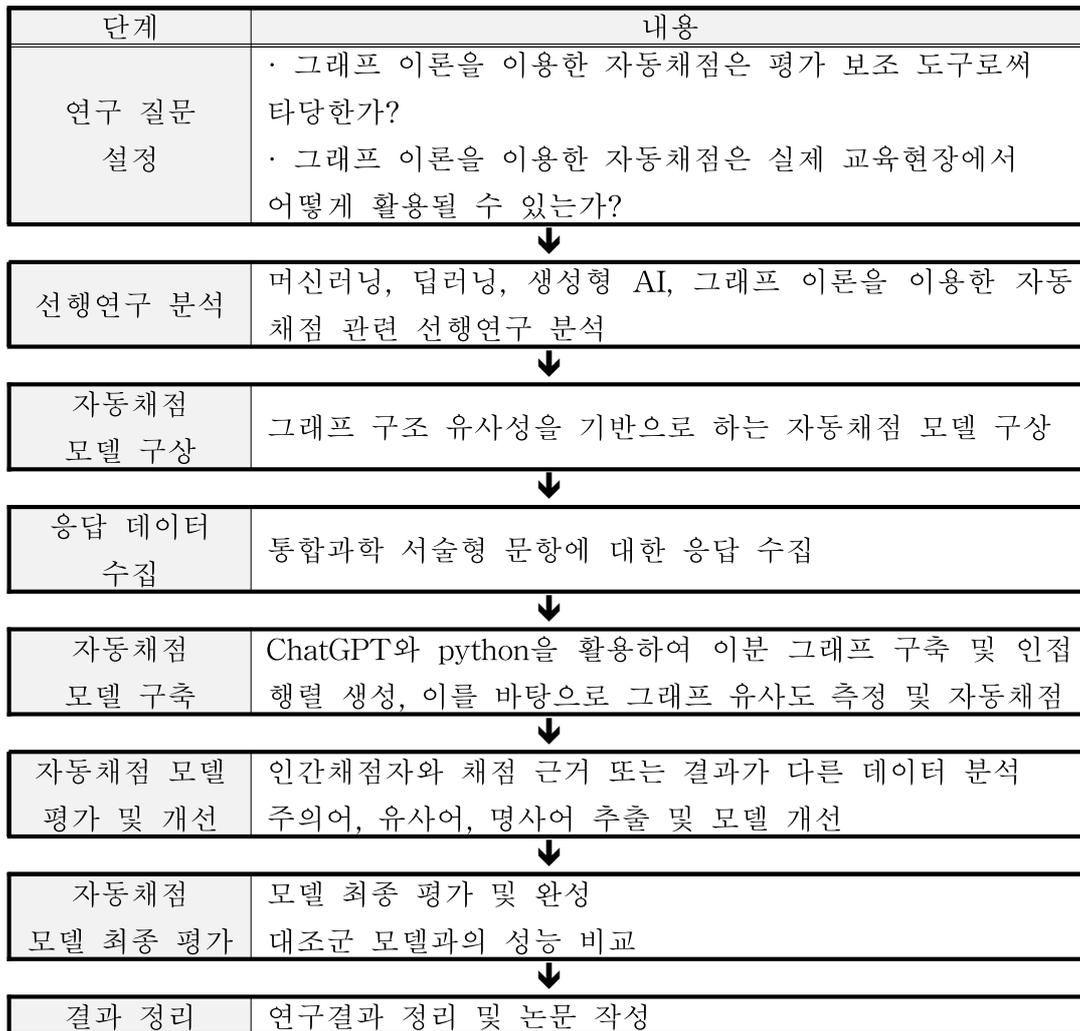
구축된 그래프 간 구조적 유사성을 타당하게 정량화하면 자동채점 도구로써 활용가치가 있다.

다음으로, 아이디어를 시험하고 모델의 성능을 평가하기 위해 필요한 데이터를 수집했다. 재직중인 학교에서 출제하고 응답한 1개의 서술형 문항 답안을 전사하여 데이터화 하였다.

응답 데이터에서 노드와 간선의 역할을 할 주어, 동사, 목적어로 이루어진 삼중항(triple)을 ChatGPT를 활용하여 추출했다. 그리고 이를 그래프로 구축하기 위해 python의 NetworkX library를 이용했다. 모범 답안과 학생의 응답으로 구축한 그래프를 인접 행렬로 변환하여 그래프를 행렬화한 후, 인접 행렬의 차이를 통해 그래프 유사도를 측정하고 자동채점을 수행했다.

자동채점 성능 평가를 위해 인간 채점자가 직접 학생들의 응답을 수동채점하고, 신뢰성을 높이기 위해 연구자 이외의 채점자 1명과의 점수 평균을 적용했다. 인간 채점자와 자동채점 모델의 결과를 혼동 행렬(confusion matrix)를 이용하여 표현하고 정확도, 정밀도, 재현율, f1 score를 기준으로 그 성능을 평가했다. 평가 결과 나타난 문제점 개선을 위해 ChatGPT를 통해 주의어, 유사어, 명사어를 추출했고 이를 반영하여 그래프를 구축하고 인접행렬로 변환하였다.

수정된 모델의 자동채점 성능을 점검하고 평가했다. 객관적 모델의 성능 분석을 위해 대조군 모델들을 구축하고 그 성능을 비교했다. 이 연구의 추진 절차를 요약하면 [그림 3-1]과 같다.



[그림 3-1] 연구 절차

3.2. 분석 자료

본 연구에서는 사용된 데이터셋은 서술형 문항 "발전기의 원리를 '운동 에너지', '전기 에너지', '면적', '유도 전류' 단어를 포함하여 설명하시오."에 대한 답안들이다. 이 문항은 2015 개정 교육과정 고등학교 1학년 통합과학 교과와 '핵발전과 차세대에너지' 단원의 성취기준 '[통과 09-01] 발전기 모형을 활용하여 화석연료, 핵에너지 등이 가정이나 산업에서 사용하는 전기 에너지로 전환되는 과정을 분석할 수 있다.'를 반영한 문항이다. 2023년 서울시 소재의 A과학고등학교 1학년 학생들이 해당 단원을 이수한 뒤, 서술형 문항1에 대한 137개의 응답을 데이터셋으로 사용했다. 데이터셋에 대한 정보를 요약하면 [표 3-1]과 같다.

[표 3-1] 데이터셋 개요

데이터셋	문항	교과	성취기준	데이터 수
데이터셋	발전기의 역할과 원리를 '운동 에너지', '전기 에너지', '면적', '유도 전류' 단어를 포함하여 설명하시오.	고1 통합과학	통과 09-01	137

서술형 문항의 설계는 다음과 같다. [그림 3-2]는 미래엔 통합과학 교과서에는 발전기의 역할과 원리이다. 교과서에서는 전자기 유도 현상의 원인으로 '코일 내부의 자기장 변화'와 함께 '코일 면의 면적의 변화'를 설명하고 있다. 이에 대한 지식·이해 수준 점검과 함께 수업시간에 부연 설명한 발전기의 역할이 '운동 에너지'를 '전기 에너지'로 전환하는 것임을 인지하고 있는지 확인하고자 했다.



④ 그림 IX-2 발전기의 원리 자석 사이에서 코일이 회전하면 자기장(초록색 화살표)이 수직으로 통과하는 코일 면의 면적이 변하여 유도 전류(보라색 화살표)가 흐른다. 90° 회전하는 순간 유도 전류의 방향이 바뀐다.

[그림 3-2] 통합과학 교과서 내 발전기에 대한 설명(김성진, 2019)

이를 기반으로 제작한 평가 문항, 채점기준, 배점은 [표 3-2]와 같다. 핵심 개념으로 ‘면적’, ‘운동 에너지’, ‘전기 에너지’, ‘유도 전류’ 4가지를 제시하고, 발전기의 역할과 원리를 올바르게 설명한 경우 점수를 부여한다. 채점기준은 2가지로 구분하였는데, 문항 제작 의도에 맞게 핵심 개념을 활용하여 발전기의 역할과 원리를 설명했는지 확인하는 채점기준1과 문항 제작 의도와는 다르지만 과학적으로 타당하게 발전기의 역할과 원리를 설명했는지 확인하는 채점기준2가 있다. 이와 같이 구분한 이유는 다음과 같다. 평가 문항에는 평가자의 의도가 포함되어 있고, 평가자의 교육 가치가 반영되어 있다. 평가 문항을 통해 학습자의 이해 점검 및 개념 응용·활용 능력의 발달 또한 일어날 수 있기 때문에, 과학적으로 타당한 답안인지 점검하는 채점기준2와 함께 평가자의 의도에 알맞은 답안인지 점검하는 채점기준1을 함께 설정했다.

[표 3-2] 데이터셋의 문항, 채점기준1,2, 배점

문항	발전기의 역할과 원리를 '운동 에너지', '전기 에너지', '면적', '유도 전류' 단어를 포함하여 설명하시오.
채점기준 1	1. '면적', '운동 에너지', '전기 에너지', '유도 전류' 4가지 단어를 모두 사용했을 경우 2. '운동 에너지', '전기 에너지'를 이용하여 발전기의 역할을 올바르게 설명한 경우 (예시) 운동 에너지를 전기 에너지로 전환하는 장치이다. 3. '면적', '유도 전류'를 이용하여 발전기의 원리를 올바르게 설명한 경우 (예시) 자기장에 수직한 면적이 코일이 회전함에 따라 달라져 유도 전류가 흐른다.
채점기준 2	1. '면적', '운동 에너지', '전기 에너지', '유도 전류' 4가지 단어를 모두 사용했을 경우 2. 발전기의 역할을 올바르게 설명한 경우 (예시) 역학적 에너지를 전기 에너지로 전환하는 장치이다. 3. 발전기의 원리를 올바르게 설명한 경우 (예시) 코일의 단위 면적을 지나는 자기선속의 크기가 달라져 유도 전류가 발생한다.
배점	채점 기준 3가지를 모두 만족했을 경우 3점 채점 기준 3가지 중 2가지를 만족했을 경우 2점 채점 기준 3가지 중 1가지를 만족했을 경우 1점 채점 기준 3가지를 모두 만족하지 못했을 경우 0점

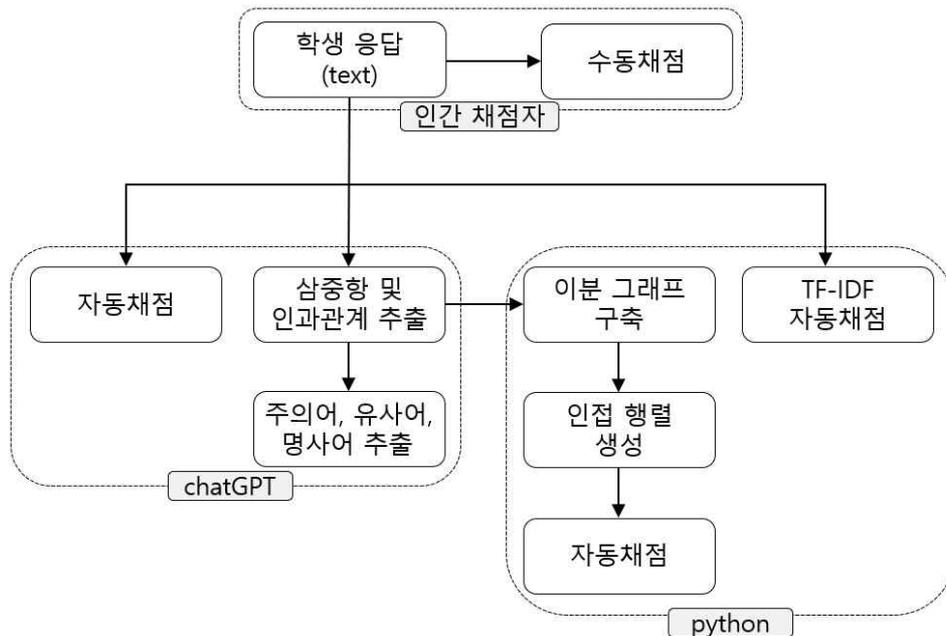
데이터셋의 점수별 데이터 수, 비율, 채점기준은 [표 3-3]과 같다. 응답자가 과학고등학교 1학년이라 과학에 대한 지식과 관심이 많고, 문항 난이도가 높지 않아 최대 점수를 받은 응답이 많다. 평가자의 의도와는 다르나 과학적으로 타당한 답안의 존재로 인해 평가기준2에 의거한 채점 결과에서 상대적인 평균 점수가 높음을 알 수 있다.

[표 3-3] 데이터셋의 점수 척도 및 채점기준, 점수별 데이터 수

점수	데이터 수 (평가기준1)	비율(%)	데이터 수 (평가기준2)	비율(%)
3	61	44.9	76	55.9
2	43	31.6	34	25.0
1	27	19.8	23	16.9
0	5	3.7	3	2.2
합계	136	100	136	100

3.3. 분석 방법

본 연구에서는 학생의 응답과 모범 답안(참조 답안)을 그래프로 구축한 뒤, 두 그래프의 유사도를 바탕으로 예상 점수를 반환하는 모델과 모델을 구축하는 프레임워크를 제안한다. 먼저, 데이터의 레이블(label) 참값으로 인간 채점자의 수동채점을 수행한다. 신뢰도 제고를 위해 연구자 및 과학교육을 전공한 1인의 채점자, 총 2인의 채점 결과의 평균값을 참값으로 설정한다. 이후 응답 텍스트를 그래프로 구축한 뒤 인접 행렬을 생성하여 참조 답안과의 그래프 유사도를 비교하여 자동채점하여 수동채점의 결과와 비교하여 모델의 성능을 테스트 한다. 모델간의 성능 비교를 위해 현재 최신 연구 중 하나인 생성형AI를 활용해 직접적으로 채점하는 방법, 단어 중요도를 통해 통계적 방법으로 자동채점하는 TF-IDF 방법들과 비교하여 본 연구에서 제안하는 모델의 활용 가치를 평가한다. 이러한 연구과정의 개요는 [그림 3-3]과 같다.



[그림 3-3] 그래프를 이용한 자동채점 연구 과정 개요

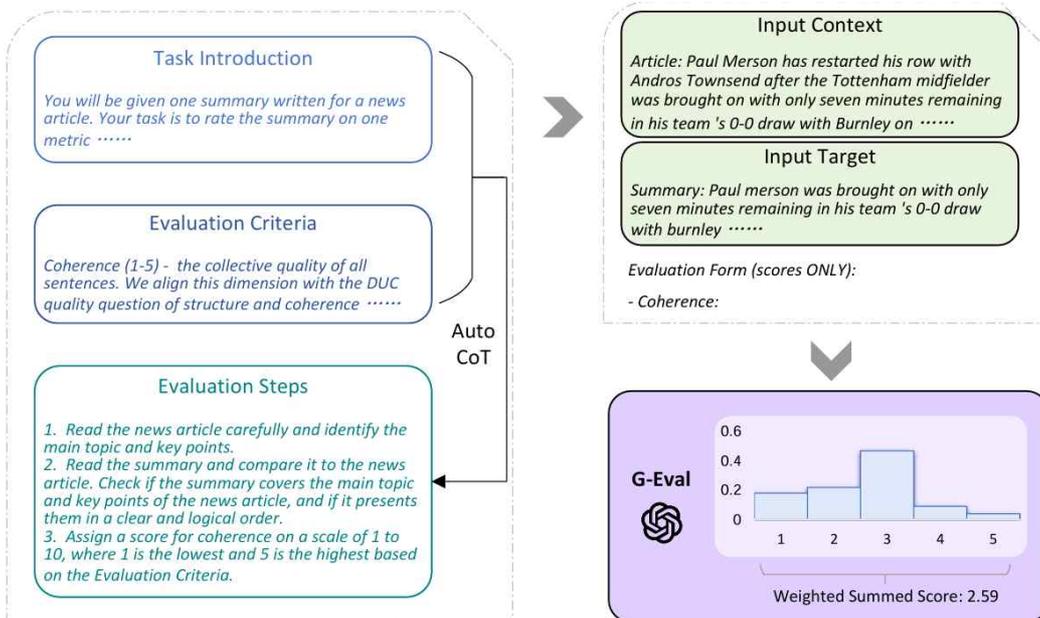
3.3.1. 프롬프트 엔지니어링을 통한 대조군 제작

자동채점 모델의 평가 보조 도구로서의 타당성을 판단하기 위해 대조군 설정이 필요하다. 채점 방식에는 참조 답안과의 유사성을 바탕으로 채점하는 참조 기반 접근(reference-based approach)과 인간 채점자가 채점한 결과를 이용하여 모델을 훈련하는 응답 기반 접근(response-based approach)가 있는데, 응답 기반 접근을 이용하는 기존의 자질 설정을 통한 분류모델을 학습시키는 머신러닝 방법 또는 단어 임베딩을 통해 학습시키는 딥러닝 방법은 대조군으로 설정할 수 없다. 그 이유는 위 방법들은 모델의 성능을 위해 모델을 학습시킬 많은 데이터 요구하는 것에 반해, 본 연구에서 제안하는 방법은 데이터 양에 구애 받지 않는 그래프를 이용한 채점방법을 제안한다. 본 연구에서는 100개~1000개 이하의 소수의 데이터를 사용하므로 기존의 머신러닝 또는 딥러닝 방식의 모델과 비교했을 때 성능이 우수하게 나올 수밖에 없다. 따라서 많은 데이터를 필요로 하지 않고, 최신 자동채점 연구 방법인 ‘생성형AI를 활용한 자동채점’을 대조군으로 설정한다. 이 방법은 2.2.3절에서 설명되어 있고, 최신 연구로 [표 2-5]를 참고할 수 있다.

생성형 AI로는 대규모 데이터를 처리하고, 그래프 구축을 위해 데이터 가공이 용이하고 API가 가능한 ChatGPT를 사용한다. 그 중 2023년 11월에 출시되어 현재 기준 2023년 12월까지의 데이터가 학습되어 있는 GPT-4의 경량화 모델 GPT-4 Turbo 모델을 사용하여 분석하는 도중, 2024년 5월 출시되어 현재 기준 2023년 10월까지의 데이터가 학습되어 있는 OpenAI의 최신 모델 GPT-4o로 변경하여 진행하였다. GPT-4o는 GPT-4 Turbo와 같은 수준의 성능을 보이지만 인출 속도가 빠르고 비용도 저렴하다. 특히 영어가 아닌 언어처리에 대해 현재까지의 모델 중 가장 뛰어난 성능을 보이고 있어 한글 자연어 처리에 적합하여, 현장에서 활용도가 높다고 판단되어 모델을 변경하였다.

이러한 생성형 AI를 자동채점에 활용하기 위해서 정교화된 프롬프트 엔지니어링이 필요하다. 프롬프트 엔지니어링은 알고리즘이 여러 프롬프트

트를 시도하여 최적의 프롬프트를 찾아 선택하는 강화학습을 통한 자동 프롬프트(Automated prompts)와 인간이 설계한 방법에 따라 모델이 최적의 결과를 응답하도록 유도하는 수동 프롬프트(Manual prompts)가 있다(Wang, 2021). 내용 기반 자동채점은 평가기준이 명확하므로 수동 프롬프트를 사용한다. Liu(2023)는 GPT4와 같은 대형언어모델의 자연어 생성품질 향상을 위한 프레임워크를 제안했다([그림 3-4]). 요약하면 과제 도입(Task Introduction), 평가 기준(Evaluation Criteria), 평가 단계(Evaluation Steps)로 구성되는데, 채점 대상, 채점표 양식을 입력하여 자동채점에 활용한다.



[그림 3-4] 자연어 생성 품질 평가 프레임워크 G-Eval의 개요(Liu, 2023)

Liu(2023)가 제안한 G-Eval 프롬프트를 기초로 순서만 조정하여 자동채점을 위한 프롬프트를 구성하였다. 과제 도입 및 평가 단계에서 내용 기반 자동채점 과제를 선언하고, 평가 단계를 설정하여 CoT(Chain of Thought)를 부여했다. 그런 다음 구체적인 평가 기준(rubric)을 선언하고, 답안을 입력한다. 채점기준에는 정답 답안의 예시를 제외하고 입력했다. 예시를 입력해서 실험해본 결과, 예시 이외의 답안을 모두 오답

으로 채점하는 경향이 있어 예시를 제거해주는 것이 근거와 피드백 결과가 타당했다. 마지막으로 출력 규칙을 주지시켜주었다. 이때 채점기준 1,2에 공통적으로 사용된 세부채점기준 1을 제외하고 제시했다. 그 이유는 연구를 진행하는 과정에서 세부채점기준 1에 대한 평가가 수동채점 결과와 비교하여 정확도가 매우 낮았다. 문장안에 단어를 찾아내는 어렵지 않은 과제라고 생각했으나, 예상보다 심한 부정확함이 나타났다. 세부채점기준 1에 대한 채점은 현재의 기술로 완벽히 수행할 수 있기 때문에 수동채점의 결과를 활용했고, 이는 모델의 객관적인 성능 평가에 침해되지 않는다고 판단하였다. [표 3-4]는 데이터셋에 대해 구성한 프롬프트 디자인이다.

[표 3-4] 자동채점을 위한 프롬프트 디자인

구성	프롬프트 내용	
	채점기준1	채점기준2
과제 도입 및 평가 단계	<문항>에 대한 <답안>을 <조건>을 바탕으로 채점해 주십시오. <문항> (문항 제시) <조건> 2가지를 모두 만족할 경우 2점, 2가지 중 1가지를 만족할 경우 1점, 모두 만족하지 못했을 경우 0점으로 평가한 뒤, 근거와 피드백을 <양식>에 맞춰 설명해 주십시오. [근거]에는 조건1, 2에 대해 만족 여부를 {만족, 불만족} 중 하나로 나타내야 합니다. 피드백에는 더 좋은 답변을 하기 위한 제안이 들어가야 합니다. 조건이 부분적으로 만족하거나 불완전 할 때에는 만족하지 않은 것으로 간주합니다.	
평가 기준	<조건> (세부채점기준 1 및 예시를 제외하고 채점기준1 제시)	<조건> (세부채점기준 1 및 예시를 제외하고 채점기준2 제시)
대상 입력	(답안 입력)	
출력 규칙	<양식> [점수] [근거] -조건1: -조건2: [피드백]	

[표 3-5]는 채점기준1, 2 각각에 대하여 ChatGPT4o가 데이터셋의 17번 답안을 채점한 결과 예시이다.

[표 3-5] 17번 답안 ChatGPT4o 자동채점결과(채점기준1, 2)

데이터셋 17번 답안	코일과 자석의 상대적인 움직임에 따라 시간에 따라 코일을 지나는 자기 선속의 변화로 유도전류가 흐르는 것이다. 코일을 지나는 자기선속은 자기장 내 코일의 면적과 동일하다.	
반환결과	<p style="text-align: center;">채점기준1</p> <p>[점수] 0점</p> <p>[근거] -조건1: 불만족 -조건2: 불만족</p> <p>[피드백] 답안에서 '운동 에너지'와 '전기 에너지'를 언급하지 않아 발전기의 역할을 올바르게 설명하지 못했습니다. 또한, '면적'과 '유도 전류'를 언급했지만, 발전기의 원리를 명확하게 설명하지 않았습니다. 발전기의 역할과 원리를 설명할 때, 운동 에너지가 전기 에너지로 변환되는 과정을 포함하고, 코일과 자석의 상대적인 움직임에 의해 유도 전류가 발생하는 원리를 더 구체적으로 설명해야 합니다. 예를 들어, "발전기는 운동 에너지를 전기 에너지로 변환하는 장치입니다. 코일과 자석의 상대적인 움직임에 의해 코일을 지나는 자기선속이 변하면 유도 전류가 발생합니다. 이때, 코일의 면적과 자기장의 변화가 유도 전류의 크기에 영향을 미칩니다."와 같이 작성하면 좋습니다.</p>	<p style="text-align: center;">채점기준2</p> <p>[점수] 1점</p> <p>[근거] -조건1: 불만족 -조건2: 만족</p> <p>[피드백] 발전기의 역할에 대한 설명이 부족합니다. 발전기는 운동 에너지를 전기 에너지로 전환하는 장치입니다. 원리에 대해서는 코일과 자석의 상대적인 움직임에 따라 유도 전류가 발생하는 것을 잘 설명하였으나, 발전기의 역할을 명확히 설명하지 않았습니다. 다음 번에는 발전기의 역할을 명확히 언급하고, 원리에 대해서도 더 구체적으로 설명해 보세요. 예를 들어, "발전기는 운동 에너지를 전기 에너지로 전환하는 장치입니다. 코일과 자석의 상대적인 움직임에 따라 시간에 따라 코일을 지나는 자기선속의 변화로 유도 전류가 발생합니다. 코일을 지나는 자기선속은 자기장 내 코일의 면적과 동일합니다."와 같이 작성하면 좋습니다.</p>

3.3.2. TF-IDF를 통한 대조군 제작

비정형 텍스트 데이터에서 의미있는 정보를 추출하는 기법인 텍스트 마이닝에서 사용되는 방법 중 하나인 TF-IDF는 문서에 출현한 용어의 빈도를 바탕으로 그 중요성을 측정하는 방법이다. 특정 단어가 문서 내에 등장하는 정도를 나타내는 Term Frequency(TF)는 특정 텍스트 d 내 총 단어의 수에 대해 특정 단어 t 가 등장하는 횟수이다.

$$TF(t, d) = d \text{ 내 } t \text{ 출현 횟수} / d \text{ 내 총 단어의 수}$$

단어가 텍스트 내에서 일반화된 정도를 나타내는 Inverse Document Frequency(IDF)는 특정 단어 t 가 포함된 텍스트에 대해 전체 텍스트의 수 D 를 나타낸 것으로 다음과 같이 정의된다.

$$IDF(t, D) = \log(D / t \text{를 포함하는 텍스트의 수})$$

위와 같이 정의된 TF와 IDF를 곱하여 단어의 가중치를 다음과 같이 계산하고, 텍스트 내에서 단어의 중요도를 나타낸다.

$$TF-IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

위와 같이 정의된 TF와 IDF를 곱하여 단어의 가중치를 다음과 같이 계산하고, 텍스트 내에서 단어의 중요도를 나타낸다.

답안 텍스트 데이터 전처리를 위해 Konlpy package를 설치하여 Okt 형태소 분석기를 사용하였다. 형태소 분석기를 통해 각 학생들의 답안을 단어의 집합으로 이루어진 행렬인 문서-단어 행렬(Document-Term Matrix; DTM)으로 생성한 뒤, 훈련과 테스트 셋을 8:2로 나누어 학습시켰다. [표 3-6]은 테스트 셋 28개의 샘플에 대해 수동채점 점수, 자동채점 점수와 함께 MSE, 정확도, QWK의 3가지 평가 지표를 나타낸 것이

다. 수동채점 점수와 자동채점 점수의 차이를 제공하여 평균한 값인 MSE는 2.48로, 0,1,2,3점으로 구성된 수동채점 점수에 대해 자동채점 점수가 1점 이상 차이가 난다는 것을 알 수 있다. 연속형 점수인 자동채점 점수를 범주화하기 위해 반올림 한 후, 정확도와 QWK값을 계산하면 각각 0.32, 0.19로 4가지의 점수 범주에 대해 무작위 확률인 0.25보다 약간 크다는 것을 알 수 있다. 즉, 무작위로 부여하는 방법보다는 정확도가 있으나, 그 신뢰도는 낮다.

[표 3-6] 테스트 셋의 수동채점, 자동채점 점수

연번	수동채점 점수	자동채점 점수
1	2.0	0.70
2	3.0	3.00
3	2.0	3.00
4	3.0	1.16
⋮		
Mean Square Error(MSE)		2.48
정확도(Accuracy)		0.32
Quadratic Weighted Kappa(QWK)		0.19

[표 3-7]은 TF-IDF를 이용해 학습시킨 모델에서 자동채점 점수에 영향이 큰 순서대로 상위 20개의 변수(단어)를 추출한 것이다. 중요도의 절댓값이 클수록 자동채점 점수에 미치는 영향이 크고, 부호는 자동채점 점수에 대해 양, 음의 영향을 의미한다. 추출 결과 정답 요소인 ‘면적’, ‘운동 에너지’, ‘전기 에너지’, ‘유도 전류’ 등과 관련된 단어가 없다. 이는 빈도를 통해 중요도를 확인하는 TF-IDF 방법은 과학적으로 유의미한 개념과 문항의 문맥 및 논리에 대한 정보가 학습에 포함되지 않기 때문이다.

[표 3-7] TF-IDF를 이용한 중요 단어 상위 20개

순위	단어	중요도	순위	단어	중요도
1	상대	-32.92	11	개수	-14.89
2	내	28.39	12	렌즈	13.54
3	공간	-25.89	13	움직임	12.99
4	에너지	24.50	14	그것	12.00
5	대해	24.32	15	주어	-11.73
6	단면	22.38	16	내부	11.69
7	함	-21.22	17	방해	-11.45
8	비스듬	-20.38	18	부분	11.32
9	사이	15.60	19	접근	10.72
10	점	15.48	20	작동	10.72

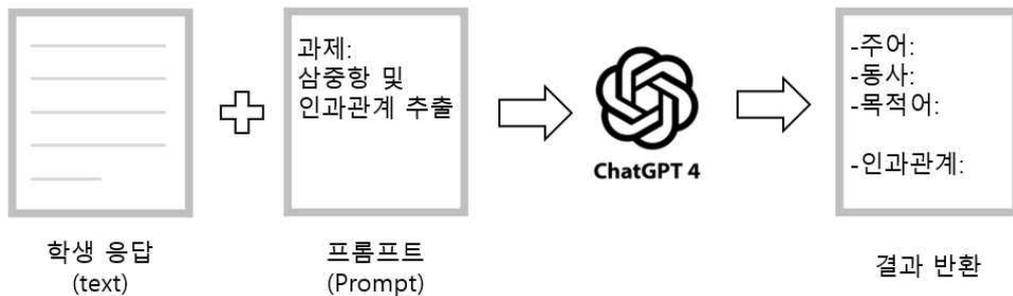
전통적 통계 방법인 TF-IDF를 활용해 보았으나 학습에 필요한 데이터가 충분하지 않아 MSE값과 중요 단어가 타당하지 않았다. 따라서 TF-IDF 방법은 본 연구에서 목표로 하고 있는 데이터 수에 구애받지 않는 평가 모델 구축에 타당한 대조군으로 보기 어렵다고 판단하였다.

3.3.3. 자동채점 모델 구축

텍스트로 이루어진 답안을 구축하는 그래프로 이분 그래프를 선택한 이유는 다음과 같다. 학습자마다 문장의 호흡이 달라 핵심 용어가 동일한 절(clause)에서 등장하지 않을 수 있다. 따라서 구축된 그래프에서 핵심 용어 노드의 위치가 달라질 수 있어 이에 대한 유사성 채점은 어렵다. 하지만 핵심 용어의 기작을 설명하는 동사는 한 개의 절에서 반드시 등장하고, 이는 정오답 여부 확인이 명확하다. 따라서 핵심 용어의 집합인 노드와 용어와 핵심 용어와 동일한 절의 동사를 또 다른 노드로 설정한 이분 그래프를 설정하고, 핵심 용어의 노드와 타당한 동사의 노드가 간선으로 연결되어 있을 때 이를 정답 요소로 간주할 수 있다. 즉, 학습자의 응답을 조사를 제외한 기본형의 형태소로 분리하여 노드에 해당하는 개별 객체를 추출한 뒤, 문장을 분석하여 객체 사이의 관계를 이어 간선을 그린다. 이를 바탕으로 구축한 참조 답안과 학생 답안의 그래프 간 유사도를 측정하면 자동채점이 가능할 것이라 판단하였다. 이를 위해서는 답안 텍스트를 절 단위로 분리하고, 절에 포함된 주어, 동사, 목적어로 이루어진 삼중항(Triplet)을 추출해야 한다. 과거의 자동채점 연구에서는 Komoran, 꼬꼬마(Kkma), 한나눔(Hannanum), khaiii를 사용하여 텍스트의 형태소를 분석했다(김철희, 2021). 하지만 사전 학습된 대형 언어 모델(Pretrained Large Language Model)의 등장 및 기술의 고도화로 인하여 영어를 넘어 한글 자연어 처리 능력이 발달하였고, 그 접근성과 활용성 또한 뛰어나다. 즉, 이분 그래프의 첫 번째 노드 집단이 될 triple에서의 주어(subject), 목적어(object)와 두 번째 노드 집단이 될 triple에서의 동사(verb)를 추출한다.

Opfer(2012)은 평가 문항 개발에서 첫째, 핵심 개념(Core concept)의 사용은 학습자의 장기 기억을 용이하게 한다는 점 둘째, 문항에 대한 인지 수준이 높을수록 인과과정을 바탕으로 현상을 설명한다는 점을 발견했다. 즉, 올바른 핵심 용어와 이에 대한 동사가 잘 추출되었더라도, 글의 맥락이 달라 오답이 될 수 있다. 기존의 연구에서는 이를 해결하기

위해 문장을 벡터화하여 텍스트 유사도, 구문 트리를 구성했으나 이는 문장의 유사성만 검출할 뿐이고 그 정확도 또한 뛰어나지 않았다. 모델 학습을 이용하는 방법 또한 많은 양의 자료가 필요하고 해당 문항에만 적용된다는 한계가 있다. 이 연구에서는 고도로 발달된 생성형 AI를 이용해 절간 관계를 추출한다. 문맥의 흐름에서 변수가 다른 변수를 직접적으로 유발하는 관계가 있을 때에는 인과관계(causality)로 간주할 수 있고, 변수간의 패턴이나 경향이 나타나 동시에 변하기는 하나 직접적인 인과관계가 아닐 때 연관관계(correlation)으로 간주한다. 따라서 서술형 문항의 참조 답안에서 유도하는 핵심 인과관계(Core causality)가 실제로 학생의 답안에도 나타나는지를 확인한다. 수학, 과학, 컴퓨터과학에서는 핵심 용어를 바탕으로 현상을 설명하는 인과관계 설정이 중요하고, 이 요소까지 포함되어야 완전한 정답이 될 수 있다. 따라서 학생 답안에 대해 삼중항 및 인과관계를 생성형 AI를 통해 추출한다([그림 3-5]).



[그림 3-5] 삼중항 및 인과관계 추출

[표 3-6]은 삼중항 및 인과관계 추출단계에서 사용된 프롬프트 디자인이고, [표 3-7]은 ChatGPT4가 반환한 결과의 예시이다. 삼중항 및 인과관계 추출에서 ChatGPT의 반환 결과 성능 향상을 위해 프롬프트내에 예시를 삽입하는 Few-shot Prompting을 사용한다. 대형 언어 모델에게 새로운 작업 수행에 최소한의 예제를 포함시키면 그 패턴을 인식하여 반환 결과의 맥락과 형식을 더 잘 이해하게 된다.

[표 3-8] 삼중항 및 관계 추출 프롬프트 디자인

구성	프롬프트 내용
대상 입력	(답안 입력)
과제 도입	<p>위 문장의 수식어를 포함해 모든 동사가 포함되도록 주어, 동사, 목적어로 이루어진 Triplet을 최대한 많이 추출하시오.</p> <p>목적어가 2개일 경우 모두 추출하고, 주어와 목적어는 명사, 동사는 기본형으로 출력하시오.</p> <p>명사는 수식하는 형용사를 포함하고, 동사를 수식하는 부사는 두 번째 목적어로 추출하시오.</p> <p>추출 후, 문장을 분석하여 Triple 사이의 관계를 찾아내시오. 연관관계는 '&'로, 인과관계는 '→'로 나타내시오.</p>
예시 및 출력규칙 입력	<p>-예시는 다음을 참고하시오.</p> <p>예시 문장: 발전기는 전자기 유도 법칙에 따라 운동 에너지를 전기 에너지로 전환하는 장치이다. 자석이 만드는 자기장에 대해 수직인 코일의 면적은 코일이 회전함에 따라 달라진다. 이로 인해 유도 전류가 흐른다.</p> <p>삼중항 추출:</p> <ul style="list-style-type: none"> -Triplet 1: (발전기, 전환하다, 운동 에너지, 전기 에너지) -Triplet 2: (발전기, 따르다, 전자기 유도 법칙) -Triplet 3: (발전기, 이다, 장치) -Triplet 4: (자석, 만들다, 자기장) -Triplet 5: (수직인 코일의 면적, 대하다, 자기장) -Triplet 6: (코일의 회전, 달라지다, 수직인 코일의 면적) -Triplet 7: (유도 전류, 흐르다) <p>관계 분석:</p> <ol style="list-style-type: none"> 1. Triplet 1 & Triplet 2 2. Triplet 1 & Triplet 3 3. Triplet 2 & Triplet 3 4. Triplet 4 → Triplet 5 5. Triplet 5 → Triplet 6 6. Triplet 6 → Triplet 7

[표 3-9] 2번 답안의 삼중항 및 관계 추출 예시

데이터셋 2번 답안	코일 혹은 자석이 운동에너지를 갖고 이동하면, 코일의 자기장에 수직한 면의 면적이 주기적으로 변화해 전자기 유도가 발생한다. 이 과정에서 운동에너지의 일부가 전기 에너지로 전환되어 유도 전류가 흐르며, 전류의 방향은 렌츠의 법칙을 따른다.
반환결과	<p>삼중항 추출:</p> <ul style="list-style-type: none"> -Triplet 1: (코일 혹은 자석, 갖다, 운동에너지) -Triplet 2: (코일의 자기장에 수직한 면의 면적, 변화하다, 주기적으로) -Triplet 3: (이 과정, 발생하다, 전자기 유도) -Triplet 4: (운동에너지의 일부, 전환되다, 전기 에너지) -Triplet 5: (유도 전류, 흐르다) -Triplet 6: (전류의 방향, 따르다, 렌츠의 법칙) <p>관계 분석:</p> <ol style="list-style-type: none"> 1. Triplet 1 → Triplet 2 2. Triplet 2 → Triplet 3 3. Triplet 3 → Triplet 4 & Triplet 5 4. Triplet 5 & Triplet 6

ChatGPT4o의 일관성을 평가하기 위해 [표 3-6]의 프롬프트에 대한 5회의 추출 결과를 비교했다. 5개 중 2개씩 삼중항 추출 결과를 선택하여 추출된 삼중항의 평균 길이에 대해 동일한 삼중항의 갯수로 일치율을 계산한 결과는 [표 3-8]과 같다. 10개의 일치율 평균 84.0%를 통해 ChatGPT4o의 삼중항 반환결과의 일관성이 충족한다고 판단하였다.

[표 3-10] 추출 횟수간 삼중항 추출 일치율(단위: %)

추출 횟수	1회	2회	3회	4회	5회
1회	100	84.1	81.3	83.9	84.7
2회		100	83.1	85.1	84.2
3회			100	85.0	83.4
4회				100	85.6
5회					100
평균 일치율	84.0				

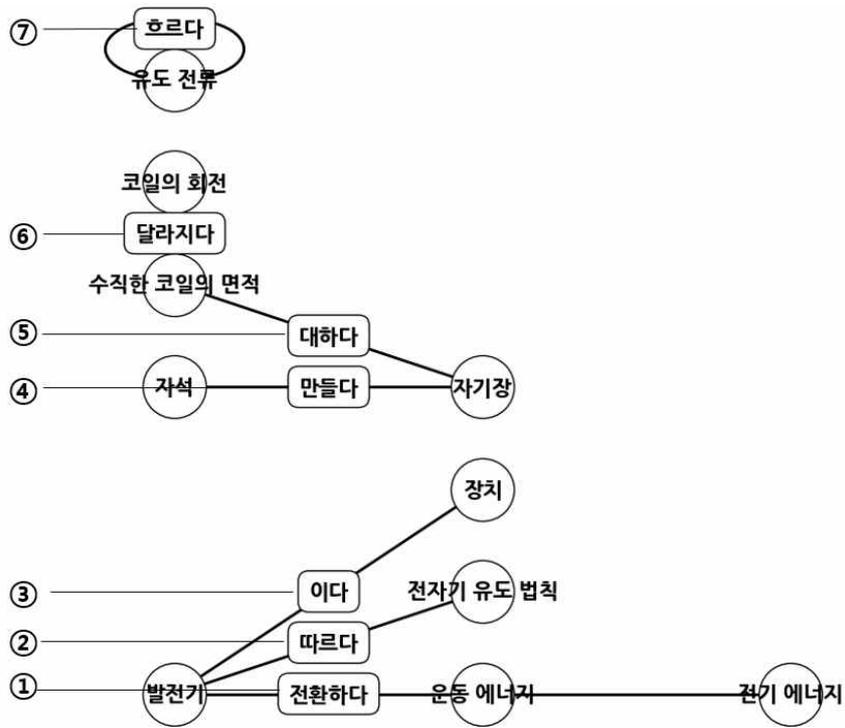
그리고 ChatGPT4o가 반환한 결과의 타당성을 평가하기 위해 5회의 반환 결과 중 최빈 결과를 선정하고, 답안 텍스트와 추출된 삼중항을 비교하며 누락된 부분이 없는지 연구자가 수작업으로 평가했다. 최빈 결과가 없거나 동일한 횟수의 최빈 결과가 존재할 경우, 첫 번째로 나타난 최빈 결과를 선택하였다. 이와 같은 방법을 사용한 이유는 삼중항 추출에 정답이 없고, 모델의 정확도에 영향을 주는 핵심 단어들이 분절 및 누락 여부가 중요하기 때문이다. 그 결과 136개의 답안에서 추출된 641개의 삼중항 중 614개가 올바르게 추출되었고, 27개의 누락된 요소가 존재하였다. [표 3-9]는 삼중항 추출 과정에서 누락된 요소의 예시를 나타낸 것이다. 누락된 요소가 명사를 꾸미는 문장이고, 그 갯수가 적으므로 모델의 정확성에 큰 영향은 미치지 않는다고 판단하였다.

[표 3-11] 12번 답안의 삼중항 추출 결과 및 누락요소

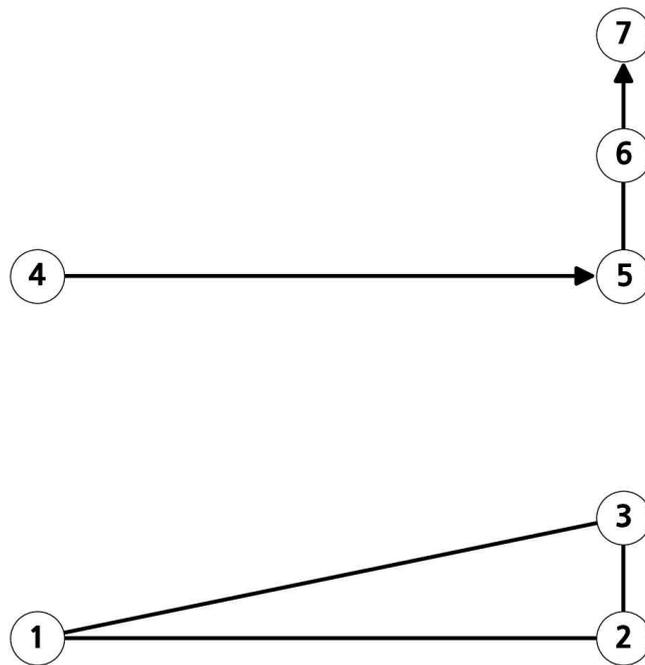
데이터셋 12번 답안	발전기는 자석이나 코일의 운동에너지를 전기에너지로 전환하는 장치이다. 단위 면적당 자기 선속이 변화하면 렌츠의 법칙에 의해 자기선속이 변화하는 것을 방해하는 방향으로 유도 전류가 흐르는 것을 이용했다.
반환결과	삼중항 추출: -Triplet 1: (발전기, 전환하다, 자석의 운동에너지, 전기에너지) -Triplet 2: (발전기, 전환하다, 코일의 운동에너지, 전기에너지) -Triplet 3: (발전기, 이다, 장치) -Triplet 4: (단위 면적당 자기 선속, 변화하다) -Triplet 5: (렌츠의 법칙, 의하다) -Triplet 6: (자기선속이 변화하는 것, 방해하다, 유도 전류의 흐름) -Triplet 7: (발전기, 이용하다, 유도 전류의 흐름)
누락요소	방해하다 → 방해하는 방향

위 논의를 바탕으로 첫 번째로 추출한 삼중항 결과를 바탕으로 자동 채점 모델의 성능을 확인하였다.

다음으로 참조 답안을 생성한다. 평가자 또는 교사가 직접 생성해도 되고, ChatGPT로도 가능하다. 생성한 참조 답안에서도 삼중항 및 인과 관계를 추출하고, 그래프를 구축한다. 이때 구축된 그래프는 이분 그래프 보다는 지식 그래프(Knowledge graph)가 유용하다. 지식 그래프도 이분 그래프와 같이 정보를 노드와 간선으로 표현하는 것은 같으나, 노드 집단이 2개로 나뉘지 않아 노드간 연결이 자유롭다. 즉, 문항에 대한 참조 답안의 맥락을 파악하고 핵심 개념 확인에 유용하다. [그림 3-6]은 데이터셋의 문항에 대해 ChatGPT가 생성한 답안 ‘발전기는 전자기 유도 법칙에 따라 운동 에너지를 전기 에너지로 전환하는 장치이다. 자석이 만드는 자기장에 대해 수직인 코일의 면적은 코일이 회전함에 따라 달라진다. 이로 인해 유도 전류가 흐른다.’를 참조 답안으로 하여 구축한 지식 그래프이고, [그림 3-7]은 참조 답안으로 구축한 삼중항들의 관계 그래프이다. [그림 3-6]의 그래프 내 원형 노드는 주어 또는 목적어이고, 사각형 노드는 원형 노드들 간의 관계를 나타내는 동사이다. 원형 노드들 왼쪽 번호는 추출된 삼중항의 index이며, 이는 삼중항 관계 추출에서 지칭하는 index와 동일하다. [그림 3-7]에서 연관관계의 삼중항들은 무방향 간선(undirected edge), 인과관계의 삼중항들은 유방향 간선(directed edge)로 나타냈다.



[그림 3-6] 참조 답안 지식 그래프



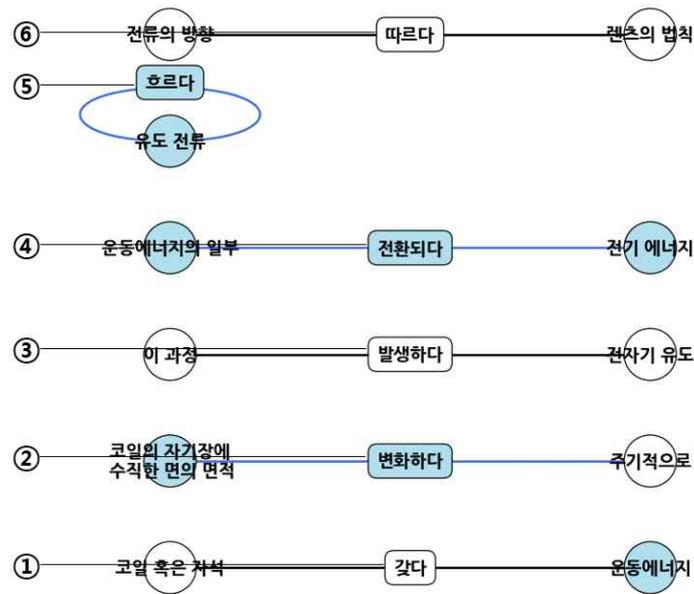
[그림 3-7] 참조 답안 관계 그래프

이렇게 구축된 지식 그래프와 관계 그래프에서 사용자는 핵심 개념과 핵심 개념과 관련된 동사를 결정한다. 문항에서 핵심 개념을 제시하고 있으므로 핵심 개념으로는 {‘운동 에너지’, ‘전기 에너지’, ‘면적’, ‘유도 전류’}로 설정할 수 있고, 정답과 관련된 동사로 {‘전환하다’, ‘전환하다’, ‘달라지다’, ‘흐르다’}로 볼 수 있다. 그리고 핵심 인과관계로는 면적이 달라지는 것(6번 삼중항)이 원인, 유도 전류가 흐르는 것(7번 삼중항)이 결과이다.

이를 바탕으로 학생들의 답안을 그래프로 구축하고 채점하는 과정에서 문제점이 나타난다. 첫째, 유사어를 사용한 답안이다. 핵심 개념은 문항에서 제시했으나, 정답과 관련된 동사의 다른 표현이 존재하므로 정답으로 인식하기 위해서는 유사 동사를 식별하는 과정이 필요하다. 따라서 [표 3-8]과 같이 유사 동사 식별 프롬프트를 디자인한다. [그림 3-8]은 유사 동사를 식별한 뒤, 핵심 개념 및 그와 관련된 동사 노드를 하늘색으로 표시한 것이다.

[표 3-12] 유사 동사 식별 프롬프트 디자인

구성	프롬프트 내용
과제 도입	<p>다음은</p> <p>"발전기의 원리를 '운동 에너지', '전기 에너지', '면적', '유도 전류' 단어를 포함하여 설명하시오."</p> <p>문제에 대한 모범 답안과 학생 답안입니다.</p> <p>-모범 답안: 발전기는 전자기 유도 법칙에 따라 운동 에너지를 전기 에너지로 전환하는 장치이다. 자석이 만드는 자기장에 대해 수직인 코일의 면적은 코일이 회전함에 따라 달라진다. 이로 인해 유도 전류가 흐른다.</p> <p>-학생 답안: (학생 답안 제시)</p> <p>두 문장에서 ['면적', '운동 에너지', '전기 에너지', '유도 전류']의 각 요소와 관련된 동사들을 2차원 list로 추출하면 다음과 같습니다.</p> <p>-추출된 동사(모범 답안): [[전환하다], [전환하다], [달라지다], [흐르다]]</p> <p>-추출된 동사(학생 답안): (학생 답안의 동사 제시)</p> <p>다음 규칙에 따라 추출된 동사(학생 답안)를 수정하여 설명과 함께 반환하십시오.</p> <ol style="list-style-type: none"> 1. 두 2차원 list를 상호 비교하되, 동일한 index의 1차원 list끼리 비교합니다. 2. 1차원 list의 요소가 여러 개 있을 경우, 두 list의 모든 요소(동사)를 상호 비교합니다. 3. 답안의 맥락을 고려하여 추출된 동사(학생 답안)의 요소(동사)의 의미가 추출된 동사(모범 답안)의 요소(동사)와 유사하거나 동일할 경우 요소(동사)를 유지합니다. 4. 맥락상 중요한 정보라도 추출된 동사(학생 답안)의 요소(동사)의 의미가 추출된 동사(모범 답안)의 요소(동사)와 직접적으로 대응되지 않거나 무관할 경우 반드시 삭제합니다. 5. 추출된 동사(학생 답안)은 유지하거나 삭제만 할 수 있습니다. 6. 마지막 줄에 추출된 동사(학생 답안)의 요소(동사) 중 유지된 요소(동사)를 2차원 list로 만들어주십시오.



[그림 3-8] 2번 답안의 핵심 개념 및 동사 노드 표시(유사 동사 식별)

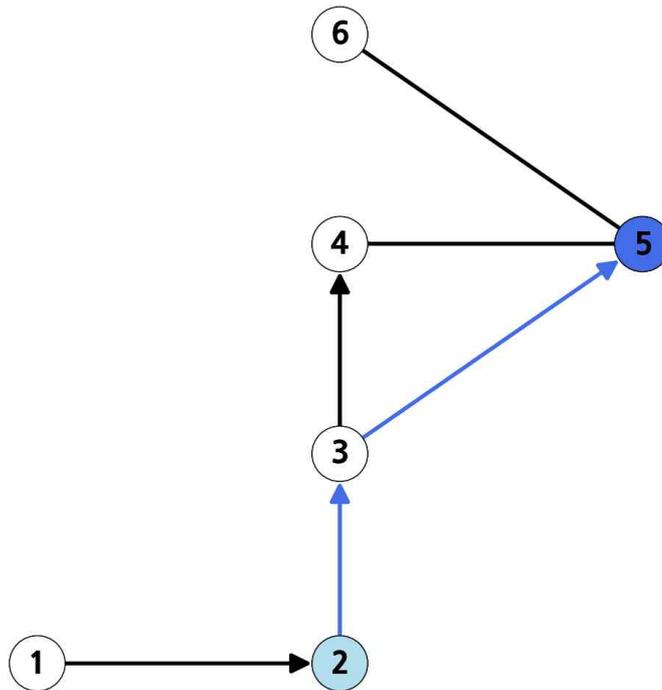
핵심 개념이 포함된 단어의 노드와 함께 ‘전환하다’와 유사한 의미의 ‘전환되다’, ‘달라지다’와 유사한 의미의 ‘변화하다’가 식별된 것을 확인할 수 있다. 또한 문항의 핵심 원인인 면적이 변화하는 것(2번 삼중항)에 해당하는 노드는 하늘색으로, 핵심 결과인 유도전류가 흐르는 것(5번 삼중항)에 해당하는 노드는 파란색으로 나타내어 두 노드가 연결되는 것을 나타낸 것이 [그림 3-9]이다. 2번 답안과 같이 핵심 원인과 핵심 결과 사이에 전자기 유도가 발생하는 것과 같은 관계를 추가할 수 있고, 이를 식별하기 위해 2번 삼중항에서 5번 삼중항으로 이어지는 경로(path)가 존재할 경우 정답으로 채점한다. 삼중항의 노드 집합 V_t 와 삼중항 간의 인과 또는 연관 관계의 간선 E_t 으로 인과 관계 그래프(Causal Graph; CG) $CG = (V_t, E_t)$ 로 표현할 수 있다. 그 중 원인 삼중항 노드 집합 $C(\subseteq V)$ 와 결과 삼중항 노드 집합 $E(\subseteq V)$ 에 대해 인과 그래프 점수(Causal Graph Score; CGS)를 다음과 같이 정의한다.

$$\text{Causal Graph Score(CGS)} = \sum_{i \in C, j \in E} \delta(i, j) \cdot w_1$$

이때 w_1 는 인과 점수에 부여하는 가중치이고, $\delta(i, j)$ 는 다음과 같이 정의한다. w_1 은 문항 내 인과관계 유무의 중요도에 따라 후술할 $w_2 \sim w_7$ 과 크기를 비교하여 채점자가 설정할 수 있다.

$$\delta(i, j) = \begin{cases} 1 & \text{노드 } i \text{에서 노드 } j \text{로 가는 유방향 간선이 존재할 경우} \\ 0 & \text{그외의 경우} \end{cases}$$

경로 내 다른 삼중항과의 관계를 포함하여 노드 i 에서 노드 j 로 가는 인과 관계가 존재할 때 가중치가 반영된 CGS를 부여한다. [그림 3-9]의 2번 답안은 CGS가 w_1 인 것을 확인할 수 있다.

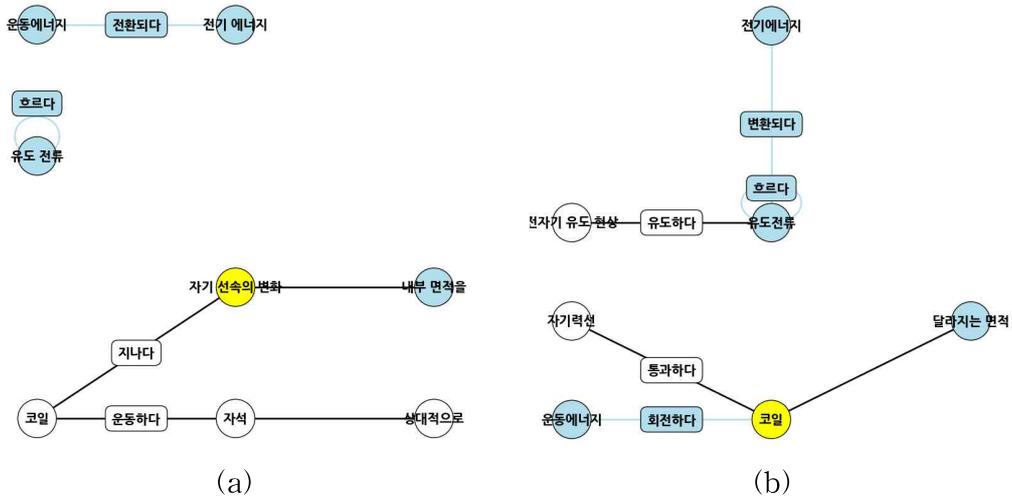


[그림 3-9] 2번 답안의 핵심 인과 관계 표시

둘째, 핵심 개념과 관련된 동사가 핵심 개념에 포함되는 경우가 있다. 따라서 설정한 정답 요소에 해당하는 동사의 명사형과 유사한 동사 또는 목적어 노드가 존재하는지 식별하는 과정이 필요하다. 따라서 [표 3-9]와 같이 명사어 식별 프롬프트를 디자인한다. [그림 3-10]은 명사어를 식별한 뒤, 노란색으로 표시한 것이다. [그림 3-10]은 핵심 개념인 ‘면적’은 관련된 정답 동사와 유사한 동사가 연결되어 있지 않은 두 답안을 나타낸 것이다. (a)의 41번 답안의 경우 유사한 명사인 ‘변화’가 포함된 노드를 표시해주는 것이다. 이는 정답으로 채점되진 않으나, 사용자에게 주의 요소로서 식별된 것이다. (b)의 39번 답안의 경우 유사한 명사가 ‘달라지는 면적’으로 ‘면적’을 포함하면서 존재하므로 이는 핵심 개념과 관련된 동사가 모두 존재하는 것으로 채점된다.

[표 3-13] 명사어 식별 프롬프트 디자인

구성	프롬프트 내용
과제 도입	<pre>list1 = [['달라지다'], ['전환하다'], ['흐르다']] list2 = [학생 답안에서 추출한 주어, 목적어 list 입력]</pre> <p>다음 규칙에 따라 설명과 결과를 함께 반환하십시오.</p> <ol style="list-style-type: none"> list1의 각 동사의 명사형에 대해 그 의미가 동일하거나 유사한 list2의 명사를 list로 저장합니다. 다른 명사들과의 관계를 고려하지 않고, 명사 자체의 의미가 유사해야 합니다. 유사한 동사가 없을 경우, 빈 list를 가져옵니다. 가져온 list를 합쳐 길이가 3인 2차원 list를 만듭니다. 결과를 ``를 이용하여 설명과 구분하여 2차원 list로 정리하여 반환하십시오.

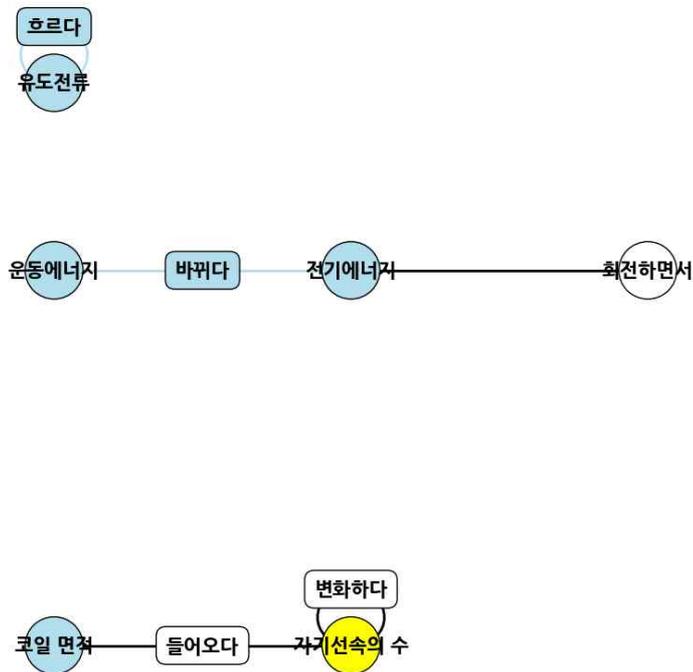


[그림 3-10] 핵심 개념 및 동사 노드 표시(유사 동사, 명사어 식별) (a) 41번 답안, (b) 39번 답안

셋째, 핵심 개념이 아닌 개념을 사용하여 과학적으로 타당하게 작성된 답안의 채점이다. 채점기준2에 의거한 모델인데, 이를 구현하기에는 근원적으로 어렵다. 사용자가 모든 과학적 개념을 핵심 개념으로 설정하는 것은 불가능하고, 이를 ChatGPT와 같은 생성형 AI에게 식별을 맡기는 것은 ChatGPT를 직접적인 채점 도구로 활용하는 것이다. 이는 채점의 주체가 생성형 AI가 되는 것이므로, 해석 가능성을 추구하는 본 연구의 목적과 다르다. ChatGPT는 자연어 처리 도구 및 본 연구에서 제안하는 평가 모델의 대조군으로써만 사용한다.

즉, 본 연구에서 제안하는 모델로 채점기준2에 의거한 채점 기술은 구현할 수 없지만, 사용자가 인지하고 모델의 결과를 수정하도록 인지하는 방법은 있다. 핵심 과학 개념이 다를지라도, 동일한 동사를 사용할 수 있다. 예를 들면 참조 답안은 유도 전류가 흐르는 원인으로 자기장이 통과하는 ‘면적’이 ‘달라진다’로 설정했다. 하지만 ‘자기장’이 ‘자속’, ‘단위 면적당 자속’이 ‘달라진다’도 과학적으로 타당한 원인이다. 이들의 공통점은 핵심 개념은 다르지만, 동사는 같거나 유사하다. 따라서 핵심 개념과 관련된 동사를 식별하고 이와 관련된 개념 노드들을 표시해놓으면, 사용자가 확인하여 정답 또는 오답 요소로 판별할 수 있다. 이를 위한 주의

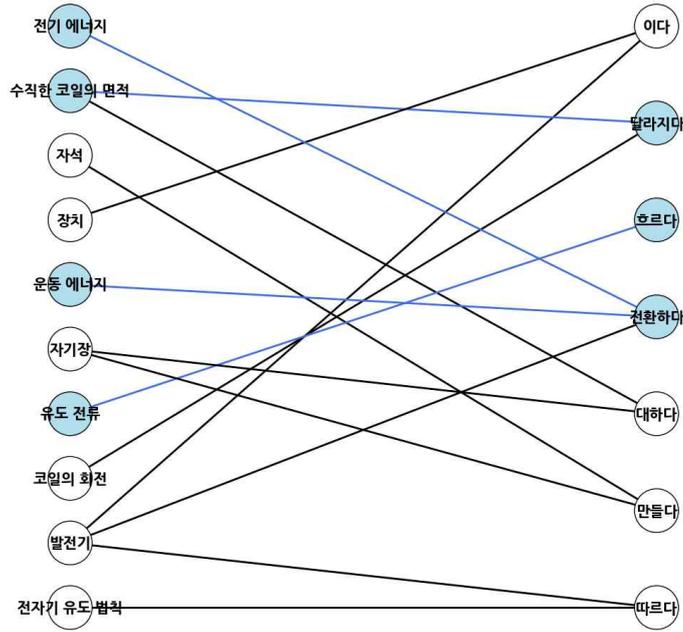
어를 식별하는 과정이 필요하다. [그림 3-11]은 ‘면적’과 관련된 동사가 직접적으로 연결되지 않으나, 관련 동사가 연결된 다른 노드의 존재를 노란색으로 표시한 것이다. 이 노드가 오답 요소일 가능성이 있어 자동 채점에 반영되지 않아 오답으로 처리되지만, 사용자에게 주의어로 제공되어 점수 확정에 주의 대상으로 보여질 수 있다.



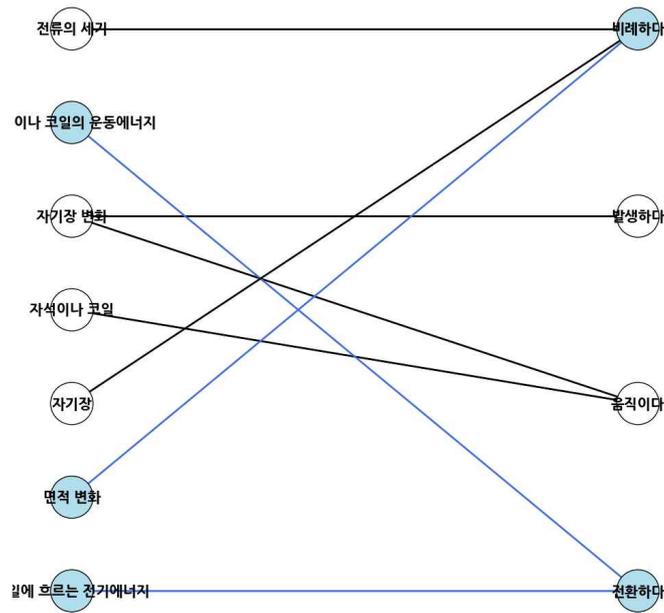
[그림 3-11] 99번 답안의 핵심 개념 및 동사 노드 표시(유사 동사, 명사어, 주의어 식별)

위와 같이 유사 동사, 명사어, 주의어 식별이 끝났을 경우 그래프의 유사도 정량화를 위해 이분 그래프화한다. 지식 그래프 형태로도 채점하거나 정량화도 가능하나, 보다 직관적이고 행렬화가 용이한 이분 그래프를 구축하는 것이 좋다. 삼중항의 명사와 목적어로 이루어진 첫 번째 노드 집단과 삼중항의 동사로 이루어진 두 번째 노드 집단에 대해 관계가 있는 노드끼리 간선으로 연결한다. 지식 그래프와 마찬가지로 핵심 개념과 관련된 동사는 하늘색으로 표시하여 참조 답안과 학생 답안의 이분

그래프를 구축할 수 있다([그림 3-12], [그림 3-13]).



[그림 3-12] 참조 답안의 이분 그래프



[그림 3-13] 22번 답안의 이분 그래프

두 이분 그래프의 유사도를 정량적으로 표현하기 위해서는 각 노드의 집합을 행과 열로 하는 인접행렬(adjacency matrix)을 구성한다. 간선의 가중치는 노드끼리 연결이 되어있을 때를 1, 연결이 되어있지 않을 때를 0으로 두어 인접행렬을 구성하면 참조 답안의 인접 행렬은 [그림 3-14]와 같다. 핵심 개념 및 동사는 파란색 진한 글씨로 축에 표시했고, 핵심 개념과 동사가 만나는 셀이 주요 채점요소이고 파란색으로 표시했다. 나머지 개념과 동사의 연결 셀은 검은색으로 표시했다.

전기 에너지	1	0	0	0	0	0	0	
자기장	0	0	0	1	1	0	0	
장차	0	0	1	0	0	0	0	
전자기 유도 법칙	0	1	0	0	0	0	0	
운동 에너지	1	0	0	0	0	0	0	
유도 전류	0	0	0	0	0	0	1	
코일의 회전	0	0	0	0	0	1	0	
수직인 코일의 면적	0	0	0	0	1	1	0	
자석	0	0	0	1	0	0	0	
발전기	1	1	1	0	0	0	0	
	전환하다	떠나다	이다	만나다	대하다	달라지다	후이다	

[그림 3-14] 참조 답안의 인접 행렬

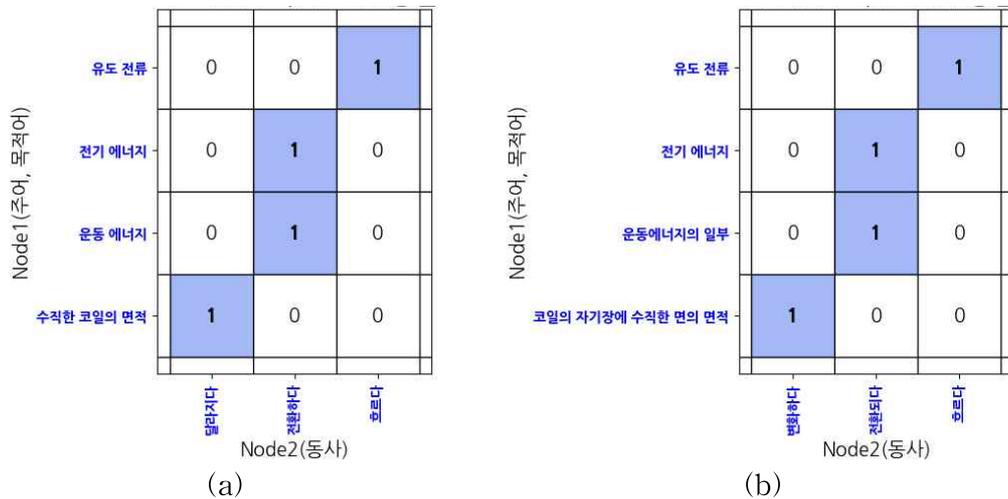
같은 방법으로 학생들 답안으로 구축한 이분 그래프를 바탕으로 인접 행렬을 구성할 수 있다. [그림 3-15]에는 데이터셋의 57번 답안에 대해 인접 행렬을 구성한 것이다. 핵심 개념과 동사를 파란색 진한 글씨로 축에 표시했고, 핵심 개념과 동사가 만나는 해당 셀이 주요 채점요소이고, 존재할 때 파란색으로 표시했다. 핵심 동사와 연결된 주어, 목적어의 노드이거나, 정답 동사의 명사형 노드일 경우 노란색 셀로 표현하여 채점 기준2에 의거한 정답일 수 있음을 사용자에게 주지시켰고, 이외의 셀은 검은색으로 표현했다. [그림 3-15]에서 ‘자기력선의 개수’와 ‘변하다’의 연결이 주의 요소로 표현되어 있다. 이는 실제로 정답 요소로 식별되지 않으나, 사용자가 확인하여 점수를 수정할 수 있다.

	이름	0	0	0	0	1
	전기에너지	0	0	0	1	0
	자기력선의 개수	0	1	1	0	0
	유도전류	0	0	0	1	1
	코일 내부의 면적	0	0	1	0	0
	운동에너지	1	1	0	0	0
	코일	1	0	0	0	0
		가지다	변하다	지나다	생기다	만다
		Node2(동사)				

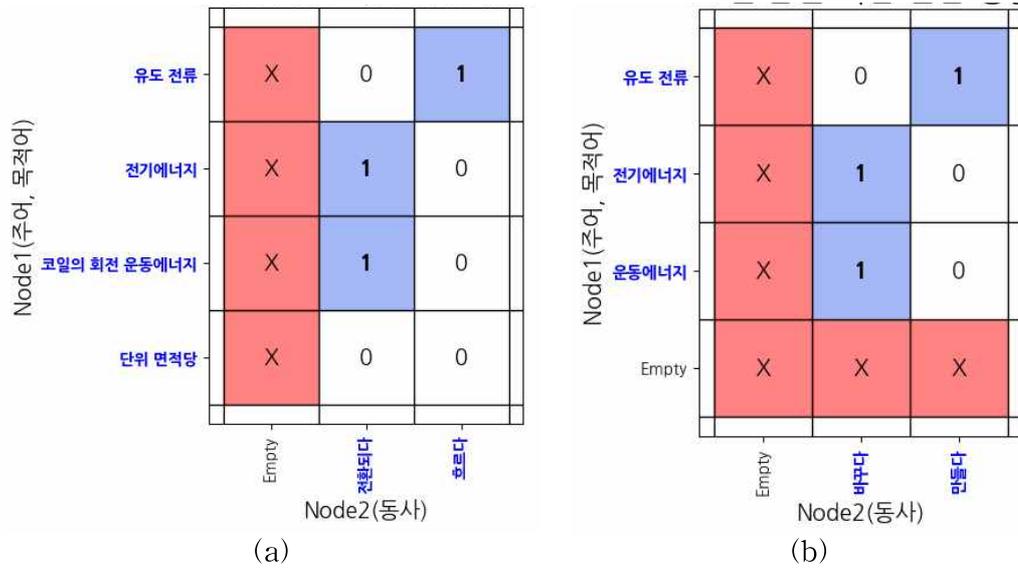
[그림 3-15] 34번 답안의 인접 행렬

두 그래프의 유사도는 핵심 개념과 그에 대한 동사를 통해 비교한다.

핵심 개념과 동사로 이루어진 부분으로 이루어진 부분(sub) 인접행렬을 핵심 인접 행렬이라고 하고, [그림 3-16]에 참조 답안과 2번 의 핵심 인접 행렬이 있다. 유사 동사 식별이 반영한 결과 두 인접 행렬은 일치한다고 간주할 수 있다. [그림 3-17]의 (a), (b)는 각각 34번, 77번 학생 핵심 인접 행렬을 나타낸 것이다. [그림 3-17]의 (a) 인접 행렬에서는 ‘면적’에 해당하는 동사가 존재하지 않고, [그림 3-17]의 (b) 인접 행렬에는 ‘면적’에 해당하는 개념과 함께 이에 해당하는 동사가 존재하지 않는다. 해당하는 행 전체나 열 전체를 붉은색으로 표시했는데, 이는 간선을 연결(1) 하거나 연결하지 않은(0) 정보가 존재하지 않아 (X)로 표기했고, 그래프 유사도 측정에 이를 반영해야 한다.



[그림 3-16] 핵심인접 행렬 (a) 참조 답안 (b) 2번 답안



[그림 3-17] 핵심인접행렬 (a) 34번 답안 (b) 77번 답안

인접행렬의 셀값을 비교하여 그래프 유사도(Graph Similarity Score; GSS)를 정량화하기 위해 기준을 설정한다. 인접행렬 $A_{i,j}, B_{i,j}$ 에 대해 두 행렬의 각 셀값을 비교하여 가중치를 부여하는 함수 $f(A_{i,j}, B_{i,j})$ 를 정의한다.

$$f(A_{i,j}, B_{i,j}) = \begin{cases} w_2 & \text{if } A_{i,j} = 1 \text{ and } B_{i,j} = 1 \\ w_3 & \text{if } A_{i,j} = 1 \text{ and } B_{i,j} = 0 \\ w_4 & \text{if } A_{i,j} = 1 \text{ and } B_{i,j} = -1 \\ w_5 & \text{if } A_{i,j} = 0 \text{ and } B_{i,j} = 1 \\ w_6 & \text{if } A_{i,j} = 0 \text{ and } B_{i,j} = 0 \\ w_7 & \text{if } A_{i,j} = 0 \text{ and } B_{i,j} = -1 \end{cases}$$

참조 답안에서 핵심 개념과 관련된 동사 사이의 간선이 존재함을 표현하는 '1'과 핵심 개념과 직접적인 관련이 없는 동사 사이의 간선이 없음을 표현하는 '0'은 그래프 유사도에서 차지하는 비중이 다를 수 있다. 또한 비교할 학생의 인접행렬에서 간선의 연결 유무(1 또는 0)과 달리 핵심 노드나 동사를 찾지 못하는 경우 해당 셀 값을 '-1'로 부여하여 가중치 함수를 적용했다. 이에 따라 참조 답안에서의 간선 유무(1 또는 0)와 학생 답안에서의 간선 유무(1 또는 0) 및 누락(-1)에 따라 가중치

$w_2 \sim w_7$ 를 채점자가 설정할 수 있다.

이를 통해 참조 답안과 학생 답안의 그래프 유사도(Graph Simillarity Score; GSS)를 수식으로 표현하면 다음과 같다.

$$\text{Graph Simillarity Score(GSS)} = \sum_{i,j} f(A_{i,j}, B_{i,j})$$

이 값에 인과 그래프 점수(Causal Graph Score; CGS)를 더하여 학생 답안의 그래프 점수(Graph Score; GS)를 산출할 수 있다.

$$\text{Graph Score(GS)} = \text{CGS} + \text{GSS}$$

채점기준1, 2의 조건1에 의한 모든 핵심 개념 사용 여부를 Score_1, 그래프 점수로 확인하고자하는 조건인 채점기준 1, 2의 조건 2, 3 점수를 Score_23, 참조 답안의 그래프 점수 GS_ref, 학생 답안의 그래프 점수 GS_stu를 이용하여 그래프 유사도를 비율로 나타내면 학생의 자동채점 결과를 다음과 같이 표현할 수 있다.

$$\text{Automated Score} = \text{Score}_1 + (\text{GS_stu} / \text{GS_ref}) * \text{Score}_23$$

평가자는 가중치 $w_1 \sim w_7$ 를 조절하여 문항마다의 핵심 개념과 동사의 일치 및 누락의 중요도를 정량화하여 그래프 점수를 산출해낼 수 있고, 이를 바탕으로 자동채점 및 피드백에 활용할 수 있다.

3.3.4. 자동채점 모델 평가

서술형 평가 자동채점에서 사용되는 평가 지표로는 크게 세 가지가 있다. 첫째, 혼동 행렬을 통해 모델의 성능을 평가하는 평가하는 방법으로 그 지표(evaluation metric)로는 정확도(accuracy), 재현율(recall), 정밀도(precision), f1-score가 있다. 둘째, 인간 채점과 모델의 채점 결과의 일치도를 0과 1사이의 값으로 산출하는 QWK(Quadratic Weighted Kappa)가 있다. 셋째, 인간 채점과 모델의 채점 결과의 차이를 제공하여 평균한 MAE(Mean Absolute Error)가 있다. 넷째, 인간 채점과 모델의 채점 결과의 상관 계수를 음의 상관관계일 때 -1부터 양의 상관관계일 때 +1까지의 사이값으로 나타내는 PCC(Pearson Correlation Coefficient)가 있다.

본 연구에서는 제안하는 모델은 그래프의 점수를 최소 점수에서 최고 점수까지의 연속형 숫자로 반환한다. 그리고 성능을 비교할 대조군으로 설정한 ChatGPT4o는 수동채점과 같이 이산형 숫자로 반환한다. 따라서 성능을 평가하기 위해 모델이 반환하는 그래프 점수를 이산형 숫자로 변환하고, 이산형 숫자들 간의 평가 지표를 사용한다. 이에 적합한 평가 지표로 혼동 행렬(confusion matrix)를 활용했다. 혼동 행렬의 예시는 [그림 3-18]과 같다. 행은 실제값, 열은 예측값을 의미한다. [그림 3-18]의 (a)와 같은 이진 분류에서는 두 class를 긍정(Positive), 부정(Negative)로 간주하여, 실제값과 예측값과 비교하여 실제로 긍정이고 긍정으로 예측한 TP (True Positive), 실제로는 긍정이나 부정으로 예측한 FN (False Negative), 실제로 부정이나 긍정으로 예측한 FP (False Positive), 실제로 부정이고 부정으로 예측한 TN (True Negative)로 분류 결과를 4가지로 구분할 수 있다.

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

(a)

		Predicted Class			
		C ₁	C ₂	...	C _N
Actual Class	C ₁	C _{1,1}	FP	...	C _{1,N}
	C ₂	FN	TP	...	FN

	C _N	C _{N,1}	FP	...	C _{N,N}

(b)

[그림 3-18] 혼동 행렬 예시 (a) 이진 분류 혼동 행렬 (b) 다중 분류 혼동 행렬(Markoulidakis, 2021)

정확도(accuracy)는 모든 예측 중 올바르게 예측된 경우의 비율을 나타낸 것이다. 재현율(recall)은 실제 Positive 경우 중 모델이 올바르게 Positive로 예측한 비율을 나타내는 것이다. 정밀도(precision)은 모델이 Positive로 예측한 경우 중 실제로 Positive인 경우의 비율이다. f1-score은 정밀도와 재현율의 조화 평균으로, 두 지표의 균형을 나타내는 수치이다([표 3-10]).

[표 3-14] 모델 성능 평가 지표(이진 분류)

평가 지표	산출식
정확도 (accuracy)	$Acc = \frac{TP + TN}{TP + FN + FP + TN}$
재현율 (recall)	$TPR = \frac{TP}{TP + FN}$
정밀도 (precision)	$PPV = \frac{TP}{TP + FP}$
f1-score	$f1\text{-score} = 2 \cdot \frac{TPR \cdot PPV}{TPR + PPV}$

이렇게 다양한 평가 지표를 쓰는 이유는 과제마다 정확도보다 재현율, 정밀도가 중요할 때가 있다. 예를 들면 스팸 메일 검출 과제는 스팸 메일을 정상 메일로 잘못 분류하는 것보다 정상 메일을 스팸 메일로 분류

하는게 문제가 된다. 이런 과제는 정밀도가 중요한 지표이다. 다른 예시로 암 진단에서는 암이 없는 환자를 암으로 분류하는 것보다 암이 있는 환자를 암이 없다고 분류하는 것이 문제가 된다. 이런 과제는 재현율이 중요한 지표이다.

이를 다중 분류로 확장하면 다음과 같다. 분류 Class가 N개일 때, [그림 3-19]의 (b)와 같이 i 행 j 열 값을 $C_{i,j}$ 라고 하자. i 번째 Class에 C_i 에 대하여 $C_{i,i}$ 는 TP 이고 $C_{i,n}(n \in \{1, 2, 3, \dots, N\} \setminus \{i\})$ 는 FP , $C_{n,i}(n \in \{1, 2, 3, \dots, N\} \setminus \{i\})$ 는 FN 이 된다. 이를 바탕으로 다중 분류에서 평가 지표를 나타내면 [표 3-11]과 같다.

[표 3-15] 모델 성능 평가 지표(다중 분류)

평가 지표	산출식
정확도 (accuracy)	$\text{정확도}(C_i) = \frac{\sum_{i=1}^N TP(C_i)}{\sum_{i=1}^N \sum_{j=1}^N C_{i,j}}$
재현율 (recall)	$\text{재현율}(C_i) = \frac{TP(C_i)}{TP(C_i) + FN(C_i)}$
정밀도 (precision)	$\text{정밀도}(C_i) = \frac{TP(C_i)}{TP(C_i) + FP(C_i)}$
f1-score	$\text{f1-score}(C_i) = 2 \cdot \frac{TPR(C_i) \cdot PPV(C_i)}{TPR(C_i) + PPV(C_i)}$

이에 더하여 QWK(Quadratic Weighted Kappa)도 활용한다. 이 평가 지표는 일치 정도뿐만 아니라 그 차이의 정도도 고려한다. 예측값과 실제 값 사이의 오차를 제공하여 가중치를 적용하고, 가능한 최대 오차와 실제 오차 사이의 차이를 통해 계산한다. 예측값과 실제값을 각각 i, j 라 하고 그 차이에 대한 가중치를 $w_{ij} = \frac{(i-j)^2}{(N-1)^2}$, 실제로 관찰된 값으로 혼동행렬을 구성하는 값과 같은 O_{ij} , 무작위로 채점했을 때 기댓값으로

반영되는 기대 빈도 E_{ij} 에 대해 QWK는 다음과 같이 정의된다.

$$QWK = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}}$$

예측값과 실제값의 일치도를 -1부터 1까지 나타내는 지표로, QWK가 1일 때에는 ‘완전한 일치(Perfect agreement)’, 0일 때에는 일치도가 ‘무작위와 같음(Agreement by chance)’, -1일 때에는 평가자 간의 일치도가 ‘무작위보다 나쁨(Worse than random agreement)’으로 표현한다. Landis(1977)은 0과 1 사이의 값에 대해서도 일치의 정도를 등급화했다 ([표 3-12]).

[표 3-16] QWK값에 따른 일치도 해석

QWK값	일치도
0.81 - 1.00	매우 높은 일치도(Almost Perfect)
0.61 - 0.80	높은 일치도(Substantial)
0.41 - 0.60	중간 정도의 일치도(Moderate)
0.21 - 0.40	약한 일치도(Fair)
0.00 - 0.20	아주 약한 일치도(Slight)
< 0.00	무작위보다 나쁜 일치도(Poor)

4. 연구 결과

4.1. 문항별 모델 자동채점 결과

4.1.1. 자동채점 모델 성능 평가 및 지표

논문에서 제안한 Graph로 서술형 문항을 채점한 결과와 ChatGPT4o로 채점한 결과를 수동채점한 결과와 비교하였다. 데이터셋에 대해 산출한 평가 지표 채점기준1, 채점기준2에 대하여 정리한 것은 각각 [표 4-1]의 (a), (b)와 같다. QWK값이 모든 모델이 모든 채점기준에 대해 0보다 크므로 무작위 채점보다 신뢰성 있는 자동채점 모델로 평가할 수 있으나, 채점기준1과 채점기준2 모두 Graph를 사용했을 때 정확도, 재현율, 정밀도, f1-score, QWK값의 모든 지표에서 우수한 결과를 보였다.

[표 4-1] 자동채점 성능 평가 지표 (a) 채점기준1, (b) 채점기준2

점수	Graph					ChatGPT4o				
	정확도	재현율	정밀도	f1-score	QWK	정확도	재현율	정밀도	f1-score	QWK
0	0.941	1.000	0.833	0.909	0.961	0.511	0.444	0.800	0.571	0.660
1		0.815	1.000	0.898			0.833	0.741	0.784	
2		0.930	0.909	0.920			0.393	0.767	0.520	
3		1.000	0.953	0.976			0.650	0.210	0.317	

(a)

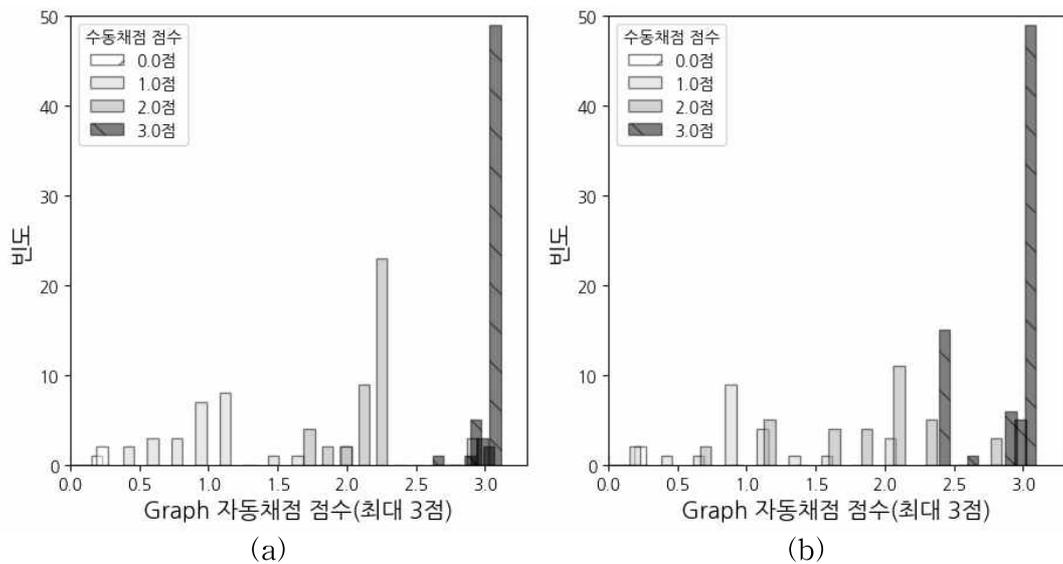
점수	Graph					ChatGPT4o				
	정확도	재현율	정밀도	f1-score	QWK	정확도	재현율	정밀도	f1-score	QWK
0	0.757	1.000	0.500	0.667	0.835	0.431	0.000	0.000	0.000	0.590
1		0.652	0.682	0.667			0.593	0.696	0.640	
2		0.706	0.545	0.615			0.275	0.647	0.386	
3		0.803	0.953	0.835			0.875	0.273	0.416	

(b)

모델의 성능을 평가하기 위해 Causal Graph Score(CGS), Graph Similarity Score(GSS)에 사용할 가중치를 설정한다. 연구자는 데이터셋에 해당하는 문항에서 핵심 개념과 동사간의 연결이 상대적으로 중요하다고 판단하여 다음과 같이 설정하였다.

$$w_1 = 0.5, w_2 = 1.0, w_3 = w_5 = 0.1, w_6 = 0.2, w_4 = w_7 = 0$$

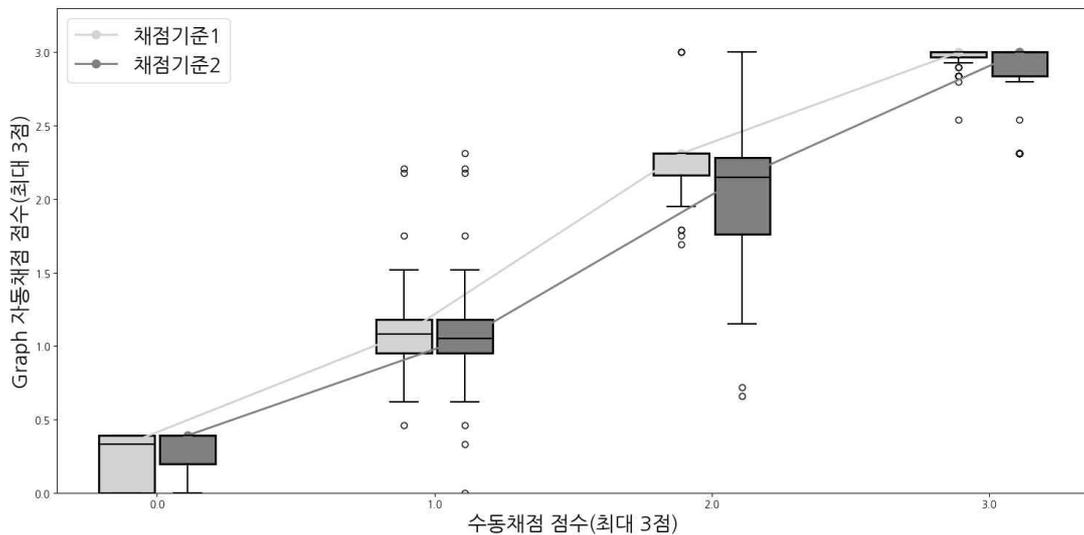
설정된 가중치로 그래프 유사도는 최소 0점에서 최대 5.6점까지의 값을 갖게 되고, 이를 3점 만점으로 규격화하여 사용한다. [그림 4-1]의 (a)와 (b)는 설정한 가중치를 반영하여 산출한 Automated Score를 이용해 자동채점한 결과의 점수 분포를 채점기준1, 채점기준2에 대해 히스토그램(histogram)으로 나타낸 것이다. 핵심 개념과 관련된 동사 사이의 간선 존재를 통해 구조적으로 평가하는 방식으로 인하여, 채점 기준에 핵심 개념이 제시되는 채점기준1에 대한 점수 분포가 실제 수동채점 점수와 유사한 결과가 나타나는 것을 확인할 수 있다.



[그림 4-1] Graph 자동채점 점수 분포 (a) 채점기준1, (b) 채점기준2

이를 상자 수염 그래프(box plot)로 나타내면 [그림 4-2]와 같다. 중간

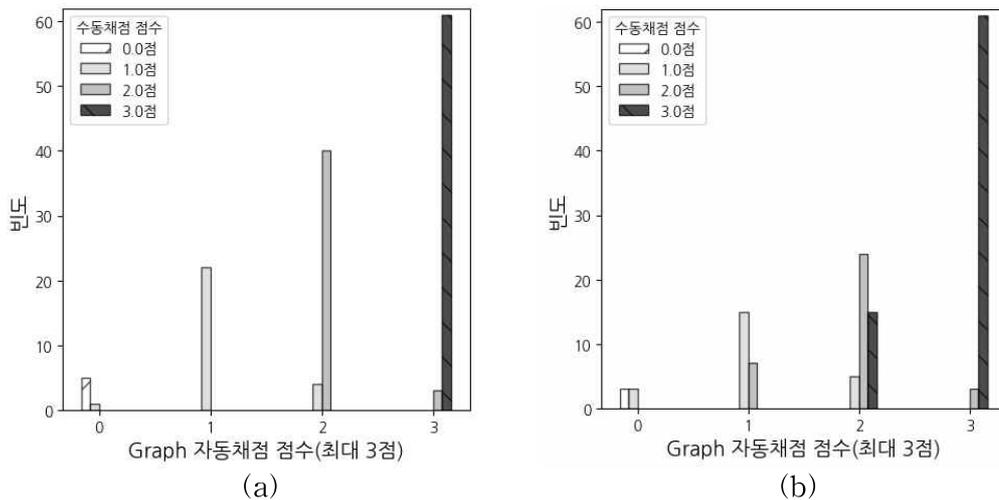
값(Median)을 나타내는 중앙선은 수동채점 점수에 대해 직선 모양을 나타낸다. 수동채점 점수에 따라 Graph 자동채점 점수는 두 채점기준 모두에 대해 대체로 일치하는 경향을 보이고 있지만, 두 채점기준간 비교하면 수동채점 점수가 2점과 3점인 데이터들에 대해 1사분위수부터 3사분위수를 표현하는 상자의 크기가 유의미하게 차이 나는 모습을 확인할 수 있다.



[그림 4-2] 채점기준별 수동채점 점수에 따른 Graph 자동채점 점수

Graph 자동채점 모델의 사용은 평가자 혹은 사용자가 평가를 모두 끝낸 후 검토의 목적으로도 사용될 수 있지만, Graph 자동채점 모델로 사전채점을 한 후 평가자가 재검토하는 것으로 사용할 때 효과가 뛰어나다. Graph 자동채점 점수를 수동채점 점수와 같이 {0, 1, 2, 3}점으로 분류하고 혼동 행렬의 평가 지표를 이용해서 그 성능을 평가할 수 있다. 사전 채점에서는 기준이 존재하지 않으므로 Graph 자동채점 점수를 반올림하는 방법을 사용한다. 반올림 과정에서 Graph 자동채점 점수가 크게 차이 나지 않는 데이터들이 다른 범주로 분류될 수 있어 반올림에 의한 오류(round-off error)가 발생할 수 있다. 하지만 본 연구에서는 정확도를 포함한 평가 지표를 통해 모델의 성능을 제시하기는 하나, 보조 채

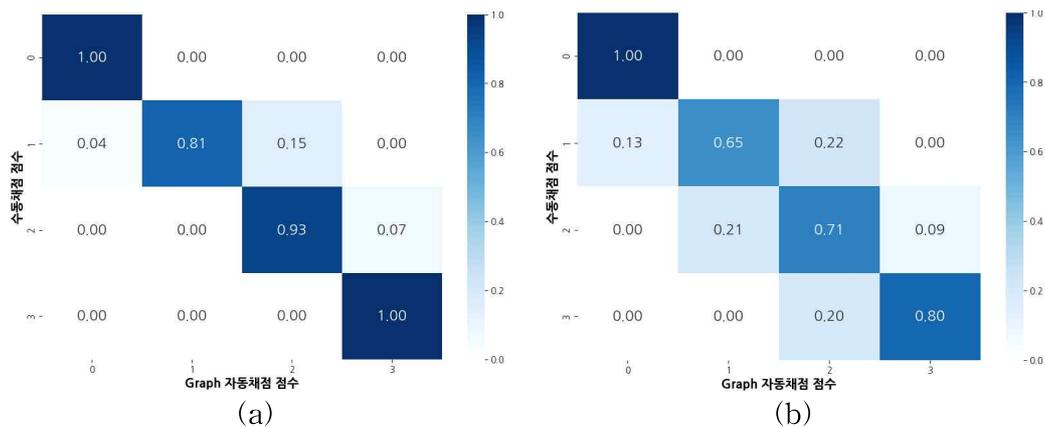
점자 및 사전/사후 검토용 평가 모델임을 강조한다는 점에서 반올림에 의한 오류에 대해서는 논의하지 않는다. [그림 4-3]의 (a)와 (b)는 각각 채점기준별 Graph 자동채점 점수를 반올림한 점수의 분포를 나타낸 것이다. 마찬가지로 채점기준2가 적용되었을 때 핵심 개념을 식별하지 못한 사례로 인하여 잘못 채점된 데이터들의 수가 많아진다.



[그림 4-3] 반올림된 Graph 자동채점 점수 분포 (a) 채점기준1, (b)채점기준2

반올림한 결과를 바탕으로 채점기준별 Graph 자동채점 점수와 수동채점 점수 사이의 혼동 행렬(Confusion matrix)를 생성하면 [그림 4-4]의 (a)와 (b)와 같다. 채점기준1을 적용했을 때에는 수동채점 점수별 80% 이상의 답안이 올바르게 채점되었다. 채점기준2를 적용했을 때에는 65~80%로 정확하게 채점된 비율이 줄어들긴 하나 그 점수 편차가 1점을 거의 벗어나지 않고 있다. 이 결과를 통해 모델의 편향성을 확인할 수 있다. 실제 현장에서 사용되는 채점기준2를 중심으로, 수동채점 점수가 2, 3점인 data에 대해서 각각 1, 2점으로 과소평가(underestimation)되는 경향이 있다. 과소평가된 대부분의 data는 후술할 4.1.2.절에서 주요하게 논의되는 ‘유형3’의 답변들로, 교사가 설정한 핵심 개념과 다른 개념을 사용하여 타당하게 서술된 답안들이다. 이는 참조 답안의 그래프 유사도가 낮으나 수동채점 점수는 높은 경우이다. 수동채점 점수가 1점인

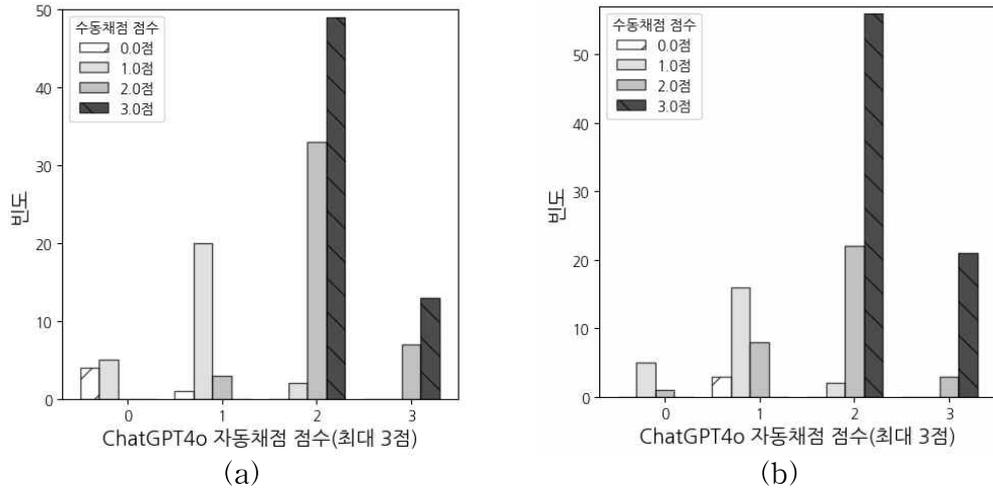
data에 대해서는 과소평가보다 과대평가(overestimation)되는 경향이 있는데, 이는 마찬가지로 ‘유형3’의 답변들로, 교사가 설정한 핵심 개념을 사용하지 않고 타당하게 서술된 답안들이다. 참조 답안에서 설정한 핵심 개념과 연결된 핵심 동사를 찾아 유사도를 산출하는 방식의 모델은, 핵심 개념을 누락했으나 과학적으로 타당하게 서술한 답안들에 대해 올바르게 자동채점하지 못하여 위와 같은 과대평가의 경향성이 드러났다. 위와 같은 이유로 [표 4-1]에서 채점기준2 성능 평가 지표 중 수동채점 점수가 2점인 data들에 대해 정밀도가 .545로 다소 낮게 나타나는 것을 확인할 수 있다. 이를 통해 모델이 최저점, 최고점 data에 대해서는 비교적 높은 성능을 보이나, 중간 영역의 점수에 대해서 비교적 성능이 저하되는 것을 확인할 수 있다. 하지만 수동채점점수와 자동채점점수의 점수차이가 1점 수준의 오차이므로, 모델의 예측 성능이 0~3점의 4단계 분류체계에서 비교적 우수한 성능을 보이고 있음을 알 수 있다. 이는 [표 4-1]에서 채점기준2 성능 평가 지표 중 QWK값이 .835로 나타나는 것을 통해 확인할 수 있다.



[그림 4-4] 반올림된 Graph 자동채점 점수와 수동채점 점수의 혼동행렬 (a) 채점기준1, (b) 채점기준2

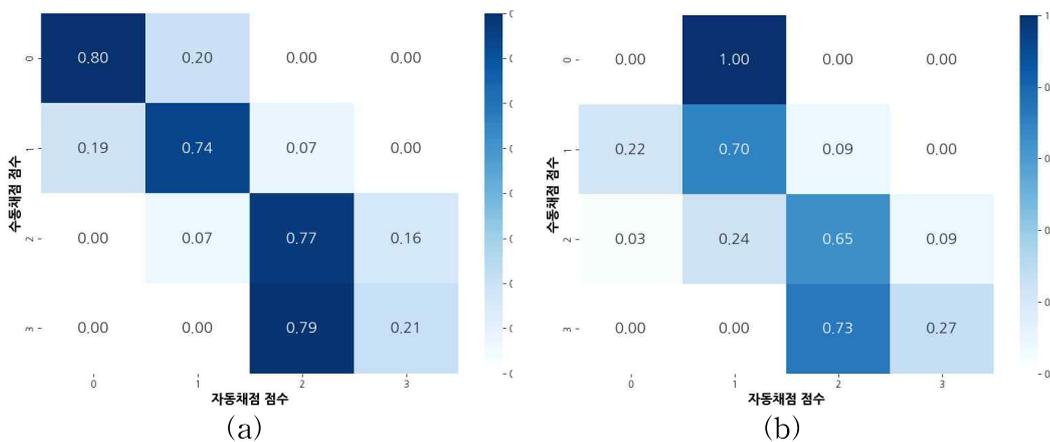
한편, ChatGPT4o를 이용하여 자동채점한 결과의 점수 분포를 채점기준1, 채점기준2에 대해 히스토그램(histogram)으로 나타낸 것은 각각 [그림 4-5]의 (a), (b)이다. 프롬프트에 입력한 채점기준을 기반으로 평가하

는 방식인데, Graph 모델과 달리 두 채점기준 모두에 대해 다수의 답안을 2점으로 채점한 것이 특징이다.



[그림 4-5] ChatGPT4o 자동채점 점수 분포 (a) 채점기준1, (b) 채점기준2

이 결과를 바탕으로 채점기준별 ChatGPT4o 자동채점 점수와 수동채점 점수 사이의 혼동 행렬(Confusion matrix)를 생성하면 [그림 4-6]의 (a)와 (b)와 같다. 2점으로 채점하는 경향성에 의해 수동채점 점수가 2점인 문항의 60% 정도를 정확하게 채점하고 있으나 수동채점 점수가 3점인 문항에 대해 낮은 정확도를 보이고 있다.



[그림 4-6] ChatGPT4o 자동채점 점수와 수동채점 점수의 혼동행렬 (a) 채점기준1, (b) 채점기준2

4.1.2. 오채점 답안 분석

이 장에서는 구체적으로 오채점 답안 분석의 원인과 해결방안을 모색한다. 채점기준2를 적용하면 핵심 개념을 사용하지 않아도 과학적으로 타당한 답안을 정답으로 채점한다. 채점기준1보다 채점기준2를 적용하는 것이 더 현실적이고, 분석결과 채점기준1을 적용한 혼동행렬에서 오분류된 경우가 채점기준2를 적용한 혼동행렬에서 오분류된 경우에 포함되기 때문에, Graph 자동채점 방법에서는 채점기준2에 대해 오채점된 답안을 분석한다. [표4-2]는 채점기준별 데이터셋의 채점기준2를 적용한 Graph 자동채점 오채점의 유형과 유형별 해결방안을 요약한 것이다.

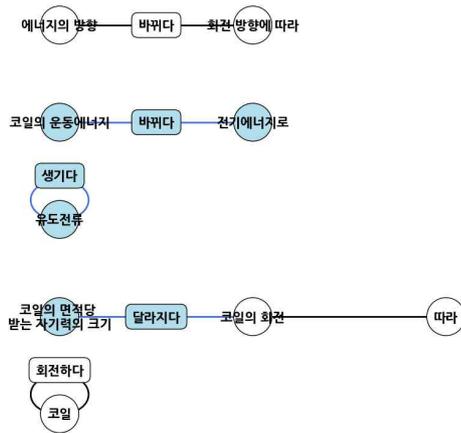
[표 4-2] Graph 자동채점 오채점 유형 및 해결방안

유형	오채점 유형	해결방안	빈도
1	오개념으로 이루어진 구조화된 답안	희귀 노드 식별	5
2	오개념이 혼재하는 구조화된 답안	모델의 경계적 사용	4
3	핵심 개념을 사용하지 않은 타당한 답안	주의어 식별	24
합계			33

유형1은 오개념으로 이루어진 구조화된 답안이다. 유사한 동사 식별 후 그래프의 구조적 유사성을 판단하는 Graph 자동채점 모델에서 핵심 개념을 포함하며 구조화된 답안은 정답으로 채점된다. 이때 오개념이 포함된 핵심 개념을 사용한 답안의 경우, 모델은 모든 답안의 정답 유무를 식별할 수 없다. [그림 4-7]은 7번 답안의 텍스트, 지식 그래프, 인접 행렬을 나타낸 것이다.

데이터셋 7번 답안	코일이 회전하면서 생기는 운동에 따라 코일의 면적당 받는 자기력의 크기가 달라진다. 이 때문에 유도전류가 생긴다. 따라서 코일의 운동에너지가 전기에너지로 바뀌고, 회전 방향에 따라 에너지의 방향이 바뀐다.
---------------	--

(a)



(b)

유도전류	0	0	1
전기에너지로	0	1	0
코일의 운동에너지	0	1	0
코일의 면적당 받는 자기력의 크기	1	0	0
	달라지다	바뀌다	생기다

Node1(주어, 목적어) Node2(동사)

(c)

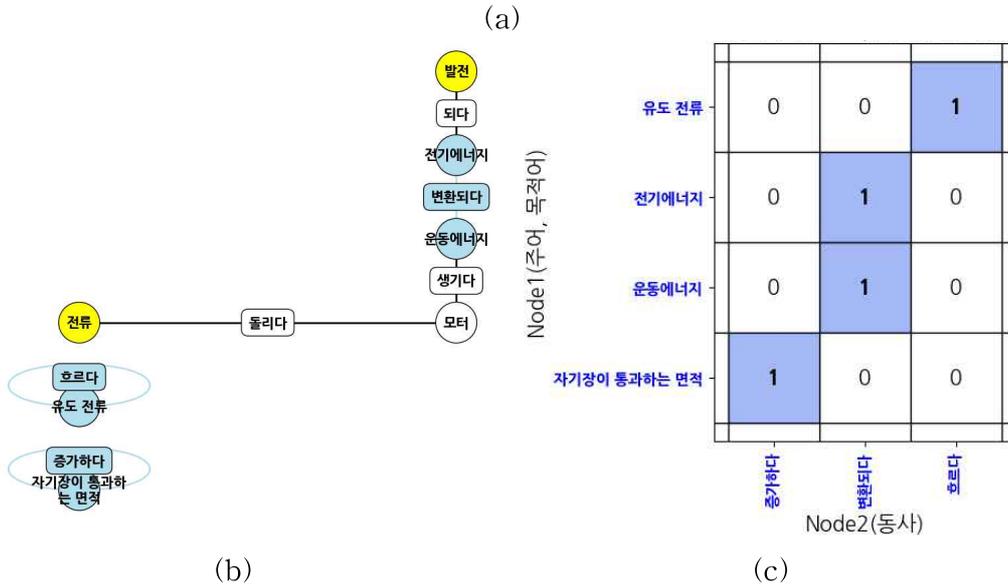
[그림 4-7] 7번 답안 텍스트, 지식 그래프, 인접 행렬

이를 보완할 방법으로는 희귀노드(Unique node) 식별을 제안할 수 있다. 희귀노드는 전체 답안을 삼중항으로 추출한 뒤, 삼중항의 핵심 개념 노드와 관련 있는 노드들 중 등장 빈도가 매우 낮은 node label을 의미한다. 다수의 학생들이 동일한 오개념을 갖고 답안을 작성하지는 않을 것으로 기대하고, 희귀노드를 찾고 이를 식별하는 과정을 거치면 이를 해결할 수 있다. 이는 박종임(2023)의 연구에서 사용된 문서 단어 행렬(DTM) 자동채점 방법, 하민수(2023)의 연구에서 사용된 빈도 기반 자동채점 방법을 역으로 이용한 것이다. 다만, 이 방법이 유의미하게 작동하기 위해서는 대형 언어 모델의 한글 자연어 처리 능력이 더욱 향상되어야 한다. 현재까지는 모든 한글 절(Clause)을 세분하고 분리하고 그 관계를 분석하는 것에 완벽한 성능을 보이지 못하고 있는 것 같다.

유형2는 오개념이 혼재하는 구조화된 답안이다. 이는 유형1과 달리, 모델이 인식한 정답 그래프 구조 이외에 오답 요소들이 그래프로 구축되어 있는 경우이다. 핵심 개념을 식별하고 구조적 유사성을 통해 채점하는 모델의 특성상 핵심 개념 이외의 그래프 구조들을 식별하고 점수화시

키지 못한다. [그림 4-8]은 80번 답안의 텍스트, 지식 그래프, 인접 행렬을 나타낸 것이다.

데이터셋 80번 답안	자기장이 통과하는 면적이 증가하면 유도 전류가 흐르고 이때 전류가 모터를 돌려 운동에너지가 생기고 이것이 전기에너지로 변환되어 발전이 된다.
----------------	--



[그림 4-8] 80번 답안 텍스트, 지식 그래프, 인접 행렬

텍스트를 분석하면, 유도 전류가 발생과 전기에너지 변환에 대한 발전기의 역할과 원리가 올바르게 나타나있다. 하지만 이와 함께 ‘전류가 모터를 돌려 운동 에너지가 생긴다’ 오개념이 혼재하고 있다. 핵심 개념을 바탕으로 구축하는 인접 행렬에는 이와 같은 오답 요소는 반영되지 않은 모습이다. 이에 따라 1번 답안은 정답으로 채점되어 수동채점 결과와 차이가 나타난다.

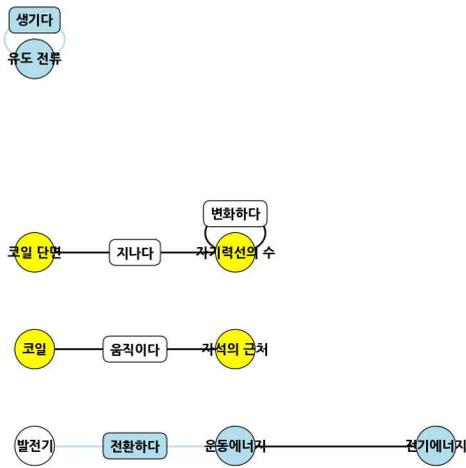
이와 같은 문제점을 해결할 방법은 현재까지 없다. 모델이 모든 오답 요소를 식별하게 하기 위해서는 수많은 규칙을 입력하거나, 대형 언어 모델에게 식별 과정을 맡길 수밖에 없다. 이는 평가자에게 피로도와 노동강도를 증가시키고, 딥러닝 모델의 설명 가능성에 대한 문제가 다시 야기되기 때문이다. 따라서 Graph를 이용한 자동채점 방법은, 인간을 완

전히 대체하는 평가도구가 아닌 채점 보조도구로써 사전채점 또는 채점 결과 검토용으로 사용되어야 한다. 결과를 맹신하지 않고, 인터페이스로 제공되는 주의어 식별을 바탕으로 최종 점수 결정과 책임은 평가자가 짊어져야 한다.

유형3은 핵심 개념을 사용하지 않은 타당한 답안이다. 핵심 개념을 식별하고 관련된 동사와의 연결을 확인하기 위해 제시한 용어를 사용하지 않고 작성한 답안이다. 채점기준의 세부채점기준 1은 불만족하나, 세부채점기준 2,3에 부합하여 채점할 수 있다. 하지만 핵심 개념을 바탕으로 그래프 구조의 유사성을 판단하는 모델의 특성상 핵심 개념을 사용하지 않은 답안에 대해서 정확도가 떨어진다. [그림 4-9]는 80번 답안의 텍스트, 지식 그래프, 인접 행렬을 나타낸 것이다.

데이터셋 45번 답안	발전기는 운동에너지를 전기에너지로 전환한다. 자석의 근처에서 코일이 움직일 때 코일 단면을 지나는 자기력선의 수가 변화하며 유도 전류가 생긴다.
----------------	--

(a)



(b)

	Empty	진행하다	생기다
유도 전류	X	0	1
전기에너지	X	1	0
운동에너지	X	1	0
Empty	X	X	X
	Empty	진행하다	생기다

Node2(동사)

(c)

[그림 4-9] 45번 답안 텍스트, 지식 그래프, 인접 행렬

텍스트를 분석하면 ‘면적’이 아닌 ‘단면’과 ‘자기력선의 수’를 이용하여

답안을 작성했다. 세부채점기준 1에서 제시한 ‘면적’은 사용하지 않았으나, 작성된 답안은 발전기의 역할과 원리가 올바르게 나타나 있다. 핵심 개념을 바탕으로 구축하는 인접 행렬에는 이와 같은 요소는 반영되지 않은 모습이다. 이에 따라 45번 답안은 오답으로 채점되어 수동채점 결과와 차이가 나타난다.

이와 같은 문제점을 해결할 방법은 연구 방법에 제안한 것과 같이 주의어 식별이다. 유형2에서 모든 오답요소를 모델에게 학습시키기 어려운 점과 마찬가지로 모든 정답요소를 모델에게 학습시키기 어렵다. 따라서 선불리 채점요소로 반영할 수는 없으나, 사용자에게 주의어로 표시할 수는 있다. 과학적 현상을 설명할 때 타당한 다른 용어를 사용하더라도 관계된 동사는 다르지 않은 경우가 많다. 따라서 참조 답안에서 사용한 핵심 개념의 동사와 유사한 동사를 식별하고, 이와 연결되어 있는 노드들을 주의 노드로 표시한다면 사용자가 채점 요소인지 판단할 수 있게 된다. 이는 유형2와 동일하게 Graph를 이용한 자동채점 모델이 인간 평가자의 대체물이 아닌 평가 보조 도구로써 사용되어야 하는 또 다른 이유이다.

한편, Graph 자동채점 방법과 달리 ChatGPT4o에서는 채점기준1, 채점기준2 각각에 대해 오채점된 답안을 분석한다. 그 이유는 여러번 프롬프트를 재작성했으나, 채점기준1을 적용할 때와 채점기준2를 적용할 때 판단 기준이 달랐다. 연구를 진행하며 반환된 결과가 매번 달라지거나, 채점기준 1, 2를 적용한 채점 판단 근거와 결과가 일치하지 않는 등 신뢰도가 부족한 결과가 나타났다. 수동채점과 비교하여 발전기의 역할과 원리에 대한 채점이 모두 일치하지 않아 총점이 같아지는 사례가 있어 혼동 행렬로 구한 정확도를 포함한 평가 지표도 과대평가된 것으로 간주할 수 있다. 자동채점 작업에 대한 ChatGPT4o의 성능은 아직 미흡한 것으로 판단된다. [표4-3]은 ChatGPT4o를 활용한 자동채점 작업에서 나타나는 대표적인 오채점 예시를 나타낸 것이다.

[표 4-3] ChatGPT4o 자동채점 오채점 근거

예시	오채점 근거
1	답안별 채점기준 적용 수준의 차이
2	오개념 미식별
3	프롬프트 작업의 불안정성

예시1은 ChatGPT4o가 제시된 답안별 채점기준 적용 수준의 차이가 나타나 오채점된 경우이다. 이는 발전기의 원리에 대한 채점에서 주로 드러나는데, ‘자기력선(자기선속)의 변화’, ‘자속의 변화’, ‘코일을 통과하는 자기장의 면적’ 등 답안별로 다른 용어의 사용을 제안하며 조건을 만족하지 못한 것으로 채점한다. 이 중 몇몇 요구는 교육과정의 수준을 벗어나고, 동일 용어를 사용했음에도 정답으로 채점된 경우도 있고, 오답으로 채점된 경우도 있어 신뢰있는 결과로 보기 어려웠다. [표 4-4]의 (a), (b)는 각각 5번, 32번 답안의 텍스트, ChatGPT4o의 채점 근거 및 피드백, 점수를 나타낸 것이다.

[표 4-4] 텍스트, 채점 근거 및 피드백, 점수 (a) 5번 답안, (b) 32번 답안

데이터셋 5번 답안	발전기는 운동에너지와 같은 역학적에너지를 전기에너지로 전환시켜주는 장치이며, 전자기 유도를 이용해 코일면의 단면적에 수직으로 통과하는 자기력선의 수의 변화에 비례하고, 이 변화를 방해하는 방향으로 유도전류가 흐르는 원리이다.
ChatGPT 4o 채점 근거 및 피드백	...(생략) ... 코일의 면적이 아니라 코일을 통과하는 자기력선의 변화가 중요합니다. 더 나은 답변을 위해서는 코일의 회전과 자기력선의 변화를 명확히 설명하는 것이 필요합니다.
점수	- 수동채점: 조건1, 2, 3 만족(3점) - ChatGPT4o: 조건 1, 2 만족(2점)
(a)	
데이터셋 32번 답안	발전기는 운동에너지(자석이 회전)를 전기에너지로 바꿔준다. 자석의 자기장이 코일에 대해 수직하게 들어오는 세기가 변하면서 코일에 유도 전류가 흐르게 된다.
ChatGPT 4o 채점 근거 및 피드백	...(생략) ... 더 나은 답변을 위해서는 면적의 변화와 유도 전류의 관계를 명확히 설명해 주세요. 예를 들어, "자석이 회전하면서 코일을 통과하는 자기장의 면적이 변하게 되고, 이로 인해 코일에 유도 전류가 발생한다"와 같이 설명할 수 있습니다.
점수	- 수동채점: 조건2, 3 만족(2점) - ChatGPT4o: 조건 2 만족(1점)
(b)	

예시2는 Graph를 이용한 자동채점에서도 드러났었던, 오개념이 포함된 답안을 식별하지 못한 경우이다. 답안 데이터에는 모터, 전류량의 변화, 전기장의 작용 등 많은 오개념 요소가 존재함에도 불구하고 식별하지 못하거나, 오개념이 근거가 아닌 다른 근거로 오답으로 채점하는 경우가 대다수였다. [표 4-5]의 (a), (b)는 각각 7번, 32번 답안의 텍스트, ChatGPT4o의 채점 근거 및 피드백, 점수를 나타낸 것이다. 7번 답안의 채점은 동일한 총점을 부여했지만 세부채점기준이 다른 경우로, ChatGPT4o의 평가지표가 과대평가되어있음을 알 수 있는 예시이다. 32번 답안의 채점은 발전기의 원리가 틀린 이유에 대해 오개념을 제시하지 않고 있다.

[표 4-5] 텍스트, 채점 근거 및 피드백, 점수 (a) 7번 답안, (b) 54번 답안

데이터셋 7번 답안	코일이 회전하면서 생기는 운동에 따라 코일의 면적당 받는 자기력의 크기가 달라진다. 이 때문에 유도전류가 생긴다. 따라서 코일의 운동에너지가 전기에너지로 바뀌고, 회전 방향에 따라 에너지의 방향이 바뀐다.
ChatGPT 4o 채점 근거 및 피드백	...(생략) ... 학생의 답안은 발전기의 원리를 잘 설명하고 있습니다. 코일의 회전으로 인해 유도전류가 발생하고, 운동에너지가 전기에너지로 변환된다는 점을 잘 언급했습니다. ...(후략)
점수	- 수동채점: 조건1, 2만족(2점) - ChatGPT4o: 조건 1, 3 만족(2점)
(a)	
데이터셋 54번 답안	코일과 자석이 있을 때 코일이 운동에너지를 지녀서 움직이면 움직임을 방해하는 면적에 비례하는 유도전류가 생긴다. 그래서 전기에너지가 생산된다.
ChatGPT 4o 채점 근거 및 피드백	...(생략) ... 발전기의 원리는 코일이 자석의 자기장 내에서 회전할 때 유도 전류가 발생하는 것입니다. 이 과정에서 코일의 면적이 자기장과 상호작용하여 유도 전류가 생성됩니다.. ...(후략)
점수	- 수동채점: 조건1, 2 만족(2점) - ChatGPT4o: 조건 1, 2 만족(2점)
(b)	

예시3은 대규모 언어 모델을 이용한 자동채점 모델 자체의 불안정성이다. 프롬프트를 작성하며 여러번 실험을 할 때마다, 그 결과가 조금씩 달라지며 채점기준1과 채점기준2를 적용했을 때 채점 근거와 피드백이 타당하지 않았다. [표 4-6]은 70번 답안의 텍스트, ChatGPT4o의 채점기준1, 채점기준2 채점 근거 및 피드백, 점수를 나타낸 것이다. 채점기준에 따라 조건2에 대한 결과가 달라지는 것을 확인할 수 있다.

[표 4-6] 70번 답안 텍스트, 채점기준1, 2 채점근거 및 피드백, 점수

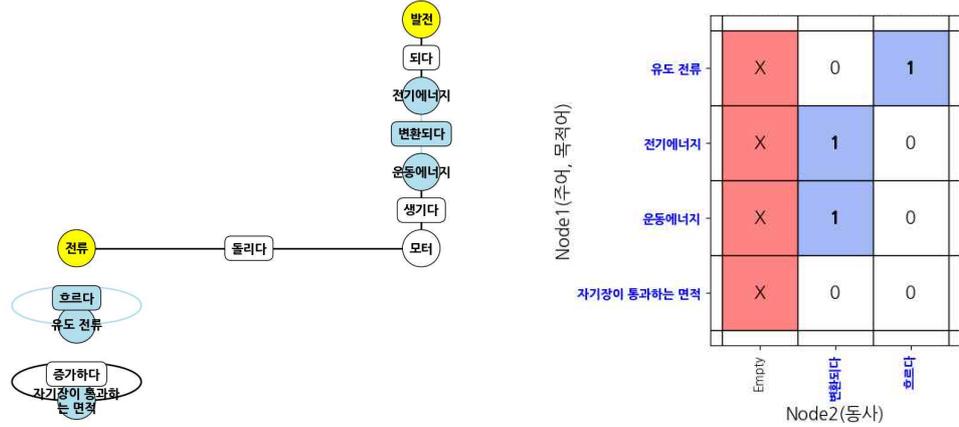
데이터셋 70번 답안	발전기는 면적에 따라 비례하며, 자석이나 코일이 움직이게 되면 유도전류가 생긴다. 이는 운동에너지가 전기에너지로 전환된 것이다.
ChatGPT 4o 채점 근거 및 피드백(채 점기준1)	...(생략) ... 조건1에 대해서는 '운동 에너지'가 '전기 에너지'로 전환된다는 설명이 포함되어 있어 만족합니다. 그러나 조건2에 대해서는 '면적'과 '유도 전류'를 이용하여 발전기의 원리를 올바르게 설명하지 않았습니다. ...(후략)
ChatGPT 4o 채점 근거 및 피드백(채 점기준2)	...(생략) ... 발전기의 역할에 대한 설명이 부족합니다. 발전기의 역할은 운동 에너지를 전기 에너지로 변환하는 것입니다. 원리에 대해서는 자석이나 코일이 움직일 때 유도 전류가 생긴다는 점을 잘 설명 ...(후략)
점수	<ul style="list-style-type: none"> - 수동채점: 조건1, 2만족(2점) - ChatGPT4o(채점기준1): 조건 1, 2 만족(2점) - ChatGPT4o(채점기준2): 조건 1, 3 만족(2점)

4.2. 자동채점 모델의 활용 방법

그래프 이론을 이용한 자동채점 방식은 첫째, 학습과정이 없어 데이터 수에 의존하지 않으므로 대규모 시험뿐만 아니라 단위 학교, 학급 수준에서도 활용 가능하다. 둘째, 그래프의 유사도를 통해 채점하기 때문에 사용자 및 평가자가 채점 결과의 근거를 이해하고 이를 응답자의 피드백에 활용할 수 있다. 셋째, 그래프 구축을 통한 채점 방식이 핵심 개념이 사용되는 내용 기반 문항에 일관적으로 적용할 수 있으므로 교과, 문항에 대한 활용 가능성이 높다. 넷째, 모범 답안의 수준을 설정하여 그래프를 구축할 수 있으므로 응답자의 연령, 학력 수준을 고려한 채점이 가능하다.

Graph를 이용한 자동채점은 Graph의 구조를 정량화하여 채점하는 방식이기 때문에, 수동채점에서 같은 점수를 받은 학생이라도 그 구조의 유사도가 달라 다른 피드백을 부여할 수 있다. [그림 4-10]은 80번 답안의 텍스트, 지식 그래프, 인접 행렬을 나타낸 것이다. 80번 답안에는 핵심 개념 ‘면적’에 대해 ‘증가한다’의 관련 동사를 사용하여 정답 요소로 식별되지 않았다. 이는 ChatGPT4o가 ‘증가한다’를 ‘달라지다’의 유사어로 식별하지 않았기 때문이다. 이는 수동채점 평가자별로 의견이 다를 수 있다. 따라서 평가자는 해당 동사의 정/오답에 대한 판단과 함께 학생에게 동사 사용에 대한 피드백을 제공할 수 있다.

데이터셋 80번 답안	자기장이 통과하는 면적이 증가하면 유도 전류가 흐르고 이때 전류가 모터를 돌려 운동에너지가 생기고 이것이 전기에너지로 변환되어 발전이 된다.
----------------	--



[그림 4-10] 80번 답안 텍스트, 지식 그래프, 인접 행렬

또한 [그림 4-11]은 23번, 114번 답안의 텍스트, 인접 행렬, Automated Score, 수동채점 점수를 나타낸 것이다. 두 답안은 수동채점과 Graph를 이용한 자동채점 모두 조건 1, 2를 만족하여 2점으로 채점되었지만 Graph를 이용한 자동채점 결과 Automated Score는 각각 2.31점, 1.79점이다. 즉, 114번 답안보다 23번 답안이 참조답안과 그래프 유사성이 높다는 것이다. 이는 텍스트의 맥락에서도 드러난다. 23번 답안은 면적을 지나는 자기장의 변화로 인한 유도 전류 발생 과정에 자기장의 작용이라는 부정확한 요소를 사용하여 조건3을 만족하지 못했다. 반면 114번 답안은 면적의 증가와 유도 전류의 크기를 비교하며 유도 전류 발생 과정을 설명하지 못했다. 즉, 조건 3을 만족하지 못하여 2점으로 채점된 결과는 같으나 유도 전류 발생 과정에 대한 이해도가 다르고, 그 정도가 그래프 유사도 점수로 반영됨을 의미한다.

답안	23번 답안	114번 답안
텍스트	코일이 움직이는 운동에너지가 전기에너지로 바뀌며 유도전류가 흐른다. 면적을 지나는 자기장이 변하여 자기장이 작용한다. 이때 유도전류가 발생하고 방향은 반대이다.	발전기는 운동에너지를 전기에너지로 변환시키는 장치이다. 발전기는 면적이 증가할수록 유도 전류가 커진다.
인접행렬		
Automated score (최대 3점)	2.31	1.79
수동채점 점수	조건1, 2만족(2점)	

[그림 4-11] 23번, 114번 답안의 텍스트, 인접행렬, Automated score, 수동채점 점수

[그림 4-12]은 54번, 134번 답안의 텍스트, 인접행렬, Automated score, 수동채점 점수를 나타낸 것이다. 두 답안 모두 핵심 개념 4개를 사용하여 수동채점 결과가 1점인 답안이나, Automated score를 보면 그래프 유사도는 차이가 크다. 134번 답안은 핵심 개념의 유기적인 관계가 적합하지 않아 그래프 유사도가 0인 반면, 54번 답안은 ‘면적에 비례’, ‘유도 전류가 생긴다’, ‘전기 에너지가 생산된다’ 등의 유의미한 핵심 개념간의 관계가 존재한다. 이는 그래프 점수로 반영되어 두 답안의 문항에 대한 이해도 차이를 알 수 있다. 이를 바탕으로 교사는 학생에게 필

요한 피드백을 제공할 수 있다.

답안	54번 답안	134번 답안
텍스트	코일과 자석이 있을 때 코일이 운동에너지를 지녀서 움직이면 움직임을 방해하는 면적에 비례하는 유도전류가 생긴다. 그래서 전기에너지가 생산된다.	운동에너지를 면적에 비례해서 유도 전류를 이용하여 전기에너지를 생산한다.
인접 행렬		
Automated score (최대 3점)	1.52	0
수동채점 점수	조건1만족(1점)	

[그림 4-12] 54번, 134번 답안의 텍스트, 인접행렬, Automated score, 수동채점 점수

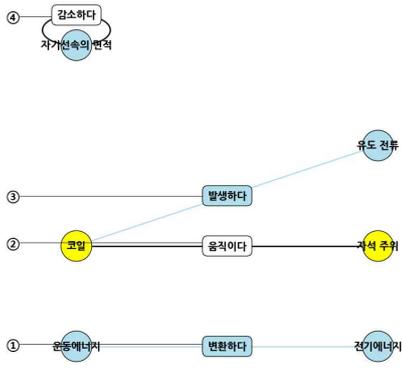
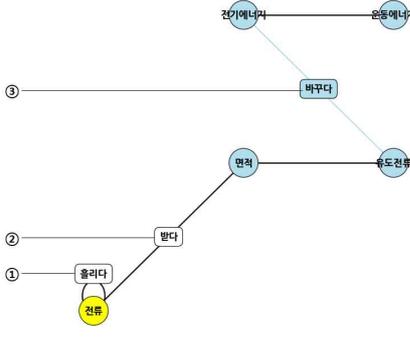
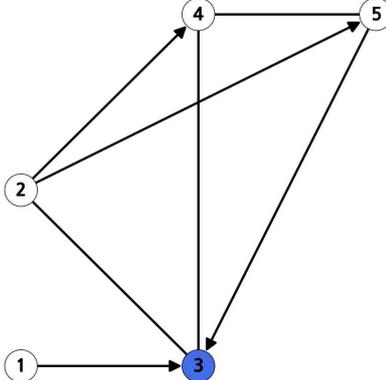
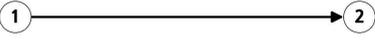
또 다른 활용 방법으로 명사어 노드와 주의어 노드에 대한 활용 예시이다. [그림 4-13]은 21번, 76번 답안의 텍스트, 지식 그래프, 인접행렬, Automated score, 수동채점 점수를 나타낸 것이다. 21번 답안을 보면 수동채점 점수(2점)에 비해 Automated score(0.72)가 낮은 경우이다. 핵심 개념을 사용하지 않고 답안을 작성하여 조건 2, 3을 만족한 경우인데, 이는 Graph를 이용한 자동채점에서는 앞 절에서 논의한대로 정답 요소로 식별하기 힘들다. 하지만 이를 보완하기 위해 지식 그래프에서 노란색으

로 표현된 것과 같이 주의 요소를 사용자에게 표시한다. 핵심 개념과 관련된 동사와 유사한 동사가 학생 답안에서 존재할 때, 이와 연결된 노드들에 대해서는 정답 요소일 수 있으므로 사용자에게 주의를 요하는 것이다. 또한 76번 답안에서는 핵심 개념인 ‘면적’에 대해 관련된 동사가 삼중항으로 추출되지 않았음에도 Automated score(2.83점)가 수동채점 점수(3점)과 유사함을 확인할 수 있다. 이는 핵심 개념과 관련된 동사가 명사형으로 존재하고, 핵심 개념과 한 개의 노드로 결합되어 있을 시 채점 요소로 반영했기 때문에 76번 답안에 대해 올바르게 자동채점할 수 있다.

답안	21번 답안	76번 답안																																								
텍스트	코일이 움직일 때 코일의 단면적을 지나는 자기장의 세기가 변하므로 코일의 운동을 방해하는 방향으로 전류가 흐른다. 이때 코일의 운동에너지가 전기에너지로 전환된다.	면적에 수직으로 들어오는 자기선속의 변화로 인해 유도 전류가 생성되는 것을 이용하여 운동에너지를 전기에너지로 바꾸는 방법이다.																																								
지식 그래프																																										
인접 행렬	<p>21번 답안 핵심 인접 행렬</p> <table border="1"> <tr> <td>Empty</td> <td>X</td> <td>X</td> <td>X</td> </tr> <tr> <td>전기에너지</td> <td>X</td> <td>1</td> <td>X</td> </tr> <tr> <td>코일의 운동에너지</td> <td>X</td> <td>1</td> <td>X</td> </tr> <tr> <td>단면적</td> <td>X</td> <td>0</td> <td>X</td> </tr> <tr> <td></td> <td>Empty</td> <td>전환되다</td> <td>Empty</td> </tr> </table> <p>Node1(주어, 목적어) Node2(동사)</p>	Empty	X	X	X	전기에너지	X	1	X	코일의 운동에너지	X	1	X	단면적	X	0	X		Empty	전환되다	Empty	<table border="1"> <tr> <td>유도 전류</td> <td>0</td> <td>0</td> <td>1</td> </tr> <tr> <td>전기에너지</td> <td>0</td> <td>1</td> <td>0</td> </tr> <tr> <td>운동에너지</td> <td>0</td> <td>1</td> <td>0</td> </tr> <tr> <td>면적에 수직으로 들어오는 자기선속의 변화</td> <td>1</td> <td>0</td> <td>0</td> </tr> <tr> <td></td> <td>면적에 수직으로 들어오는 자기선속의 변화</td> <td>바꾸다</td> <td>생성하다</td> </tr> </table> <p>Node1(주어, 목적어) Node2(동사)</p>	유도 전류	0	0	1	전기에너지	0	1	0	운동에너지	0	1	0	면적에 수직으로 들어오는 자기선속의 변화	1	0	0		면적에 수직으로 들어오는 자기선속의 변화	바꾸다	생성하다
Empty	X	X	X																																							
전기에너지	X	1	X																																							
코일의 운동에너지	X	1	X																																							
단면적	X	0	X																																							
	Empty	전환되다	Empty																																							
유도 전류	0	0	1																																							
전기에너지	0	1	0																																							
운동에너지	0	1	0																																							
면적에 수직으로 들어오는 자기선속의 변화	1	0	0																																							
	면적에 수직으로 들어오는 자기선속의 변화	바꾸다	생성하다																																							
Automated score (최대 3점)	0.72	2.83																																								
수동채점 점수	조건2, 3만족(2점)	조건1, 2, 3만족(3점)																																								

[그림 4-13] 21번, 76번 답안의 텍스트, 지식 그래프, 인접행렬, Automated score, 수동채점 점수

인과 그래프를 통해 답안에서 핵심 개념끼리의 인과 관계가 존재하는지 확인하고 피드백에 활용할 수 있다. [그림 4-14]는 73번, 123번 답안의 텍스트, 지식 그래프, 인과 그래프, automated score, 수동채점 점수를 나타낸 것이다. 두 답안 모두 핵심 개념인 면적, 유도 전류를 사용했다. 하지만 두 답안 모두 핵심 개념 ‘면적’이 ‘유도 전류’와의 인과적 관계가 없다. 이는 삼중항 추출과 삼중항간의 관계 추출에서 식별되고, 인과 그래프에서 원인 삼중항과 결과 삼중항이 추출되지 않은 모습이다. 이는 그래프 유사도 점수로 반영되고, Automated score과 수동채점 점수의 차이로 이어진다. 이를 반영해 학생에게 인과적 관계 포함의 피드백을 할 수 있다.

답안	73번 답안	123번 답안
텍스트	운동에너지를 전기에너지로 변환하는 것이 발전기의 원리. 코일이 자석주위를 움직이면 유도 전류 발생. 자기선속의 면적이 증가/감소	전류를 흘리면 전류를 받는 면적에 따라 받는 유도전류를 이용해 운동에너지를 전기에너지로 바꾼다.
지식 그래프		
인과 그래프		
Automated score (최대 3점)	2.54	1.75
수동채점 점수	조건1, 2, 3만족(3점)	조건1, 2만족(2점)

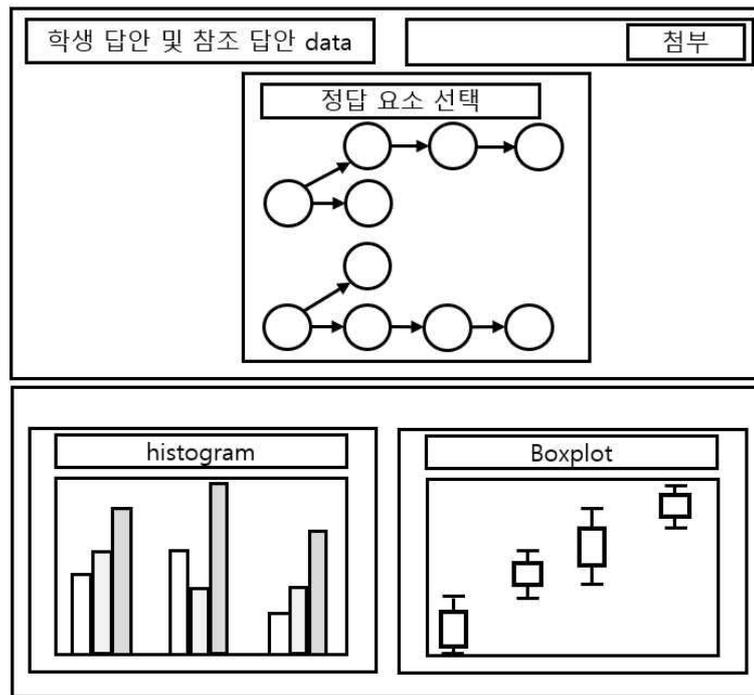
[그림 4-14] 73번, 123번 답안의 텍스트, 지식 그래프, 인과 그래프, Automated score, 수동채점 점수

이와 같은 방식으로 평가자가 가중치를 조절하여 Automated score를 산출하고, 이 값을 참조하여 평가 보조도구로써 선행 또는 후행채점에 활용할 수 있다. 연속형 점수인 Automated score를 범주형으로 구분하

여 채점하기 위해 단순한 반올림과 같은 방법을 사용하면, 그래프 점수가 매우 유사함에도 불구하고 다른 범주로 구분될 수 있다(2.49점과 2.51점). 이를 보완하기 위해서 분류하고자 하는 점수 범주와의 차이(예를 들면 1점)를 설정하고, 그 값이 점수 범주 사이의 간격과 유의미한 차이가 나타난다면(예를 들면 2.5점 이상) 특정 범주로 분류하지 않고 최종 점수 결정을 유보하도록 할 수 있다. 이러한 답안들에 대해서만 평가자가 다시 한번 확인하고 최종 평가하게 된다면 높은 신뢰도의 자동채점이 가능할 것으로 판단된다. 앞서 비교한 ChatGPT4o를 사용한 자동채점 방법은 정량적인 채점도구로써의 기능보다는 제시된 답안에서 부족한 점과 함께 모범 답안을 제시받는 것에 사용 의의가 있을 것으로 판단된다. 따라서 Graph를 이용한 자동채점을 기반으로 두 채점방법을 결합하여 사용하면 정량적 평가와 피드백 관점에서 상호 보완 가능하다.

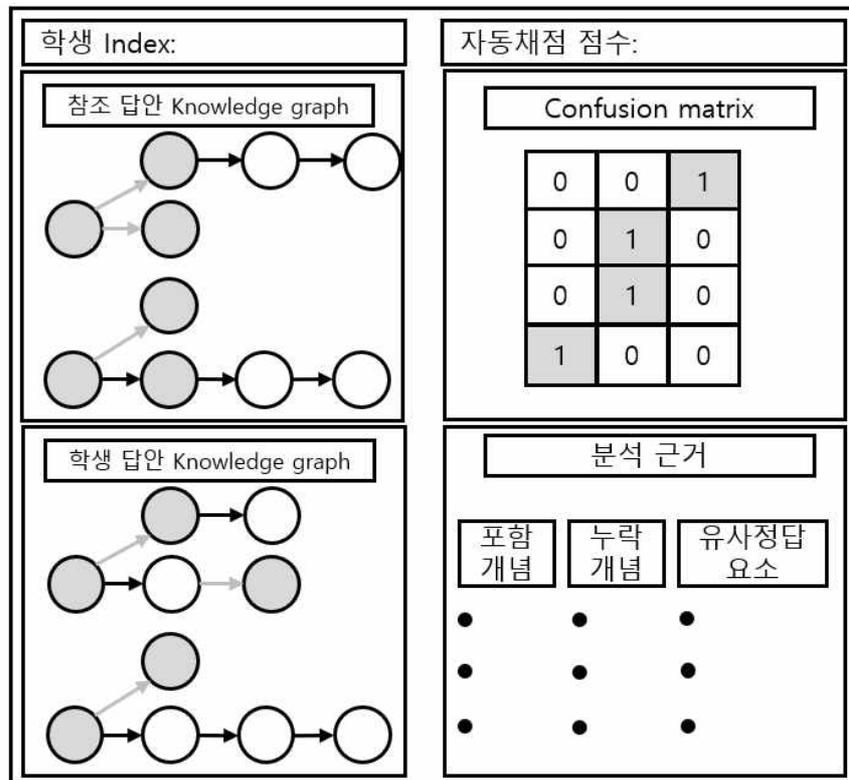
4.3. 자동채점 모델 프로그램

앞서 살펴본 자동채점 모델의 활용방법을 통해 학생 답안을 채점하고 피드백할 요소를 확인할 수 있다. 이렇게 구축한 평가 모델은 프로그램으로 전환하여 다수의 교수자들이 범용적으로 활용할 수 있다. [그림 4-15]는 본 연구에서 제안하는 모델을 프로그램으로 제작했을 때의 전체 학생 분석 화면 예시이다. 화면에는 학생 답안 및 참조 답안의 data를 첨부했을 때, 참조 답안을 바탕으로 구축된 지식 그래프 내에서 사용자가 정답 요소를 선택하는 화면을 나타내고 있다. 정답 요소를 선정하고 나면, 학생 답안으로 구축한 지식 그래프 및 이분 그래프에서 해당 요소가 존재하는지 분석 한 뒤, 전체 학생 답안에 대해 자동 채점 점수의 분포(histogram)와 수동 채점 점수와의 관계를 수염 그래프(box plot)로 나타내는 것을 볼 수 있다.



[그림 4-15] 자동채점 모델 프로그램의 전체 학생 분석 화면 예시

[그림 4-16]은 자동채점 모델 프로그램의 학생별 분석 화면의 예시를 나타낸 것이다. 학생 Index를 입력하면 자동채점 점수가 나타나고, 참조 답안과 학생답안의 지식 그래프, 혼동 행렬을 나타낸다. 이를 바탕으로 참조 답안과 일치하는 노드와 간선을 나타내는 ‘포함 개념’과 참조 답안에는 있으나 학생 답안에 존재하지 않는 노드와 간선을 나타내는 ‘누락 개념’과 참조 답안과 일치하는 간선과 연결되어 있는 노드를 나타내는 ‘유사정답 요소’를 나타내줌으로써 사용자가 모델의 채점 결과와 근거를 함께 확인할 수 있어야 한다. 이를 바탕으로 사용자는 학생의 인지상태와 누락된 개념을 확인하고, 피드백 요소를 확인하여 수업 설계 및 평가에 반영할 수 있다.



[그림 4-16] 자동채점 모델 프로그램의 학생별 분석 화면 예시

모델을 사용할 때에는 대용량의 data를 처리해야하기 때문에

ChatGPT API를 활용한다. 학생 데이터를 입력하고 전처리, ChatGPT의 ‘gpt-4o’ 모델을 활용한 삼중항·주의어·유사어·명사어 추출, 통계처리 및 학생별 그래프 구축에 필요한 시간과 비용을 계산하기 위해 동일한 작업을 5회 수행한 후 평균하였다. 그 결과 본 연구에 사용된 136개의 데이터 기준 평균 시간 58분 33.56초, 평균 비용 1.862달러이다.

5. 결론 및 제언

5.1. 결론

서, 논술형 문항을 현장에 도입하는 것의 어려움은 다음과 같다. 첫째, 평가자가 다수의 응답을 채점하는 과정에서 시간과 노력이 많이 들어, 대규모 시험에서는 행정처리까지의 인력과 시간이 턱없이 부족하다. 둘째, 채점 결과의 타당도 및 신뢰도 확보를 위해 다수의 채점자와 다중 검토 작업이 필요한데, 이 또한 많은 비용이 들고 평가 오류의 가능성을 내재하고 있다. 셋째, 모든 답안의 유사 정답 및 오답 요소를 고려한 문항을 제작하기에 부담과 위험성이 크다. 다섯째, 서, 논술형 평가는 채점 과정에 드는 시간으로 인해 평가 시점과 피드백 시점 간의 시간 간격이 크다. 이는 학습자에 대한 진단과 피드백이 늦어지고, 교육적 효과가 감소함을 의미한다. 이상의 문제로 인하여 서술형 평가는 채점의 일관성, 객관성, 공정성의 문제가 나타나고, 이와 같은 한계를 극복하기 위해 컴퓨터 기술을 활용하여 채점하는 방법인 자동채점 연구가 활발히 진행되고 있다. 자동채점 모델을 이용해 일관적인 기준으로 빠르게 답안을 채점한다면, 채점의 타당도와 신뢰도를 확보하여 서술형 문항 현장 도입에 수반되는 다수의 어려움을 극복할 수 있다. 본 연구에서는 2~3 문장 수준의 학습 개념의 지식·이해 수준을 점검하기 위한 문항인 서술형 문항에 초점을 맞추어 자동채점 할 수 있는 방법을 제안했다.

기존의 자동채점 모델들은 크게 통계 모델, 머신러닝 기반 모델, 딥러닝 기반 모델, 생성형 AI를 활용한 모델로 구분할 수 있다. 최초의 자동채점 모델인 통계 모델은 학생 답안 내에 정답 요소에 해당하는 Keyword의 출현 여부만을 채점하는 단순한 모형이었고, 문장 내부의 논리 관계나 오답 요소를 확인할 수 없어 그 한계는 뚜렷했다. 머신러닝 기술이 발달하며 텍스트 답안에서 채점 요소인 자질(feature)을 추출하고

다수의 데이터로 자질(feature)에 따른 자동채점 점수를 반환하는 모델이 나타났다. 하지만 이 방법은 자질 설정의 난해함과 다수의 양질의 데이터를 필요로 하는 단점이 있다. 딥러닝 기술이 발달하며 자질 설정 없이 인공신경망 기반의 모델을 활용하는 방법이 등장했으나, 이 또한 양질의 많은 데이터가 필요했고 자질이 없는 장점이 채점결과를 해석할 수 없는 단점으로 이어졌다. 이를 ChatGPT4를 대표로 하는 대규모 언어 모델의 뛰어난 자연어처리능력을 활용하여 학습과정 없이 채점결과를 반환하는 연구가 최근 이루어지고 있다. 하지만 반환 결과의 변동성이 크고 신뢰도와 타당성이 아직 미흡한 것이 본 연구에서 확인되었다. 즉, 기존의 자동채점 방법은 타당도와 신뢰도 확보를 위해서는 많은 양의 데이터와 학습시간이 필요했다. 이를 회피하기 위해서 사용되는 통계기반, 생성형 AI를 활용하는 방법은 그 타당도와 신뢰도가 낮았다. 따라서 자동채점 모델은 교과, 문항 유형, 응답자, 데이터 수에 구애받지 않고 다양한 분야에 적용할 수 있는 범용적이고 일반화되며, 채점결과를 사용자가 신뢰 있게 확인하고 피드백에 사용할 수 있는 타당성을 확보해야하는 지향점이 존재한다.

본 연구에서는 그래프 이론을 이용하여 서술형 문항을 채점하는 방법을 제안하고, 데이터셋에 적용해보며 그 성능을 평가했다. 성능 평가를 위해 최근 연구가 활발히 진행되고 있는 생성형 AI를 이용한 자동채점과 비교하여 덜 개방된 채점기준(채점기준1)과 개방형 채점기준(채점기준2)에 적용했다. 그 결과 그래프 이론을 적용한 채점 모델은 데이터셋에 대해 채점기준1에서 정확도 .941, 채점기준2에서 .757을 기록하였다. 이는 데이터셋에 대해 채점기준1에서 .511, 채점기준2에서 .431인 생성형 AI를 이용한 자동채점의 결과에 대해 높은 성능을 보였다. 즉, 모든 데이터셋과 모든 채점기준에서 정확도를 포함한 모든 평가 지표에서 우수한 성능을 보였다. 실험 결과에서 논의한것과 같이 ChatGPT4o의 성능이 과대평가(overestimated)된점을 고려하면, 더 우수한 성능이라고 판단할 수 있다. 평가 기준으로 가장 많이 활용되는 학습 기반의 머신러닝, 딥러닝 모델을 선정하지 않은 이유는 학습을 위해 많은 데이터를 필요로

하고, 다수의 문항에 적용할 수 없는 단점이 본 연구에서 제안하는 방법과 다르기 때문이다. 그래프를 이용한 자동채점 모형은 다음과 같은 장점이 있다. 첫째, 학습과정이 없어 특정 교과, 문항에 일반화되지 않는다. 노드와 간선을 이용해 핵심 그래프 구조의 유무를 채점하는 방식이므로 다양한 문항에 일반적으로 활용할 수 있는 모델이다. 둘째, 학습과정이 없으므로 양질의 다수의 데이터를 필요로 하지 않으므로 단위 학급부터 대규모 시험까지 확장할 수 있다. 셋째, 참조 답안의 수준과 형식 및 가중치 조절을 통해 응답자의 답안에 따른 수준별 채점이 가능하다. 고학년 응답자에 대해서는 핵심 개념과 점수의 기준을 엄격하게 부여하고, 저학년 응답자에 대해서는 핵심 개념의 수준을 낮추고 점수의 기준을 낮출 수 있다. 넷째, 채점 근거와 결과를 평가자가 확인하고 피드백에 활용할 수 있다. 핵심 그래프에서 핵심 노드와 간선의 존재를 통해 그래프 유사도를 산출하는 방식을 이용하여, 정답 요소와 감점 요소를 명확히 확인할 수 있다. 뿐만 아니라 연속형 점수로 반환되는 그래프 유사도 점수는 최종 점수가 동일한 학생들끼리의 이해 수준을 판단할 수 있고, 이에 맞게 피드백을 제공하며 학습자의 성장에 맞춤형 평가가 가능하다.

위와 같은 장점을 바탕으로 모델을 활용할 수 있는 방법은 다음과 같다. 첫째, 선행채점의 기능이다. 교사가 생성한 답안과 구조적으로 유사한 답안에 대한 정확도가 높으므로, 고득점 답안에 대해서 교사의 채점 부담을 덜어줄 수 있다. 이는 교사가 문항에 대한 이해수준이 낮은 학생에 대한 답안 채점과 피드백에 집중할 수 있다는 점에서 활용가치가 높다. 둘째, 후행채점의 기능이다. 교사의 수동채점이 끝난 후, 검토를 위한 후행채점으로 활용할 수 있다. 자동채점 점수와 수동채점 점수의 차이를 확인하고, 피드백 요소를 그래프를 통해 확인하면 보조 평가자 및 교육평가 도구로써 활용가치가 있다. 셋째, 수업 성찰에 활용할 수 있다. 교사가 사용하는 용어 및 표현, 구조가 유사한 답안과 그렇지 못한 답안의 분포를 통해 문항 답안의 다양성을 확인할 수도 있고, 수업에서 사용한 용어와 표현이 학생 답안에 대한 반영 정도를 확인할 수도 있다. 이를 통해 교사는 수업이 학생들에게 미치는 영향을 해석하고, 성찰과 반성을

통하여 이후의 수업 설계 및 평가도구 제작에 반영할 수 있다는 점에서 활용가치가 있다. 과학교과에서는 정의, 정리와 같은 엄밀함이 요구되는 문항에서 교사 및 교과서에 기반한 참조 답안과의 구조적 유사성이 중요하고, 현상을 설명하고 해석하는 개방형 문항에서는 구조적 유사성 보다는 다양하고 창의적인 답안이 중요시되므로 이를 고려하여 평가 및 교사의 문항 성찰에 반영해야 한다.

5.2. 제언

구축한 그래프의 핵심 그래프를 추출하여 유사도를 측정하는 자동 채점 방법의 한계는 다음과 같다. 첫째, 핵심 그래프 이외의 그래프 요소에 대한 채점이 어렵다. 이는 핵심 그래프에서 식별하지 못한 정·오답 요소를 식별하지 못해 채점 정확도가 떨어질 수 있음을 의미한다. 이를 극복하기 위해서 생성형 AI를 활용하여 참조 답안의 핵심 개념과 관련된 동사와 유사한 동사를 학생 답안에서 식별하고, 이와 연결된 노드를 주의 노드로 표시한다. 오답 요소 식별을 위해 학생 전체 답안에서 자주 등장하지 않는 희귀 노드를 선정하여 오답 요소로 표시할 수 있다. 이들은 정·오답 요소로 채택되지 않아 채점 정확도는 떨어질 수 있으나 사용자가 확인하고 최종 점수에 반영할 수 있다. 둘째, 유사 동사 및 삼중항 추출 과정에서 생성형 AI를 활용하므로, 모델의 채점 정확도가 생성형 AI의 자연어처리능력에 크게 의존된다. 현재의 한글 자연어처리능력은 비약적으로 상승했기는 하나, 그 결과의 변동성이 존재한다. 모델이 꾸준히 발전하고 있고, 이를 교육적으로 활용하기 위해 미세 조정 기술 등에 대한 연구가 활발히 이루어지고 있어 이 또한 해결가능할 것으로 기대한다.

그래프를 이용한 자동채점 모형에 대해 필요한 후행 연구는 다음과 같다. 첫째, 타당한 참조 답안을 다양하게 생성할 수 있는 연구가 필요하다. 다양한 학년, 교육수준 따라 참조 답안을 단계화할 수 있다면, 그래프 유사도를 더욱 정밀화할 수 있다. 이는 정성적인 학생의 이해 수준을 정량화하여 학생의 현재 이해수준을 가늠하고 피드백에 활용하여 목표 이해 수준 도달을 위한 도구로 활용될 수 있다. 둘째, 오답 요소를 식별하는 더욱 정밀한 방법이 필요하다. 텍스트 답안의 특성으로 인해 전체 그래프를 비교할 수 없어, 핵심 그래프만 비교하는 본 연구의 방법은 핵심 그래프에 포함되지 않는 오답 요소를 식별할 수 없다. 이를 위해서는 교과 내용에 대한 학습이 이루어진 생성형 AI를 활용할 수 있다. 답안에 포함되는 여러 오답 요소는 굉장히 독특한 개념이 아니라, 교과 내 또는

교과 간 학습 과정에서 생성된 종류가 한정된 것들이다. 만약 생성형 AI가 대표적인 오개념을 학습하고 있다면, 이를 식별하는데 활용할 수 있을 것이다. 셋째. 그래프를 이용한 자동채점 모형을 사용자 친화적인 환경으로 구성하여 활용 가능성을 높여야 한다. 기존의 모형들은 연구자 또는 개발자만 사용할 수 있도록, 수정이 불가능하거나 그 과정이 매우 어려웠다. 생성형 AI를 활용해 코딩의 진입장벽이 많이 낮아졌기는 하나, 범용적이고 일반적으로 사용하기 위해서는 컴퓨터 언어의 규칙을 몰라도 수정 및 보완이 가능해야 한다. 따라서 생성형 AI가 코드를 해석하고, 자연어로 된 프롬프트를 입력하면 수정 및 보완하기 위한 코드를 제시하는 등 사용자의 편의성을 확보하고 진입장벽을 낮추어야 한다.

본 연구에서 제안하는 방법은 학습할 데이터가 필요하지 않아 대량의 답안이 확보되지 않아도 사용할 수 있는 장점이 있다. 문항에 대한 다수의 참조 답안이 존재한다면 주의어 사용을 줄이고, 자동채점의 정확도를 높일 수 있다. 따라서 타당한 참조 답안을 자동으로 생성할 수 있는 방법이 필요하다. 이를 대형 언어 모델을 통해 해결할 수 있을 것으로 기대된다. 연구결과에서 확인할 수 있듯이 현재의 생성형 AI가 사용자의 요구에 맞는 정확한 채점 작업을 수행한다고 보기에는 어렵다. 앞으로 교육 목적의 특정 교과나 분야에 미세 조정 모델(Fine-Tuning Model)이 등장한다면 다양한 문항에 대해 활용할 수 있을 것이다.

참 고 문 헌

- 교육부 (2015). 2015 개정 교육과정 총론 및 과학과 교육과정. 교육부 고시 제2015-74호. [별책 1], [별책 9]
- 교육부 (2023). 미래 사회를 대비하는 2028 대학입시제도 개편 확정안
- 김성진, 안형수, 박가영, 최미화, 서인호, 한문정, 김혜경, 오현선, 구향모, 장희정, 김대준, 이진우, 류형근, 문무현, 이인순 (2017). 고등학교 통합과학. 미래앤
- 김육태 (2024). 수학과 ‘확률과 통계’ 영역에서 ChatGPT를 활용한 서답형 평가 피드백이 학생들의 피드백 리터러시에 미치는 영향. 한국교원대학교 교육대학원. 석사학위논문
- 김철희, 정지호, 한혜정, 김동현 and 신명선. (2021). 한국어 교육을 위한 형태소 분석기의 조건과 향후 과제. 새국어교육, 129, 635-672
- 박혜영, 김성숙, 김경희, 이명진, 김광규, 김지영 (2019). 수업-평가 연계 강화를 통한 서·논술형 평가 내실화 방안. 한국교육과정평가원 연구보고 RRE 2019-6
- 노은희, 심재호, 김명화, 김재훈 (2012). 대규모 평가를 위한 서답형 문항 자동채점 방안 연구. 한국교육과정평가원 연구보고 RRE 2012-6
- 노은희, 김명화, 성경희, 김학수 (2013). 대규모 평가를 위한 서답형 문항 자동채점 프로그램 정교화 및 시범 적용. 한국교육과정평가원 연구보고 RRE 2013-5
- 노은희, 이상하, 임은영, 성경희, 박소영 (2014). 한국어 서답형 문항 자동채점 프로그램 개발 및 실용성 검증. 한국교육과정평가원 연구보고 RRE 2014-6
- 노은희, 송미영, 성경희, 박소영 (2015). 한국어 문장 수준 서답형 문항 자동채점 프로그램 개발 및 적용. 한국교육과정평가원 연구보고 RRE 2015-9
- 노은희, 송미영, 박종임, 김유향, 이도길 (2016). 한국어 문장 수준 서답형 문항 자동채점 프로그램 고도화 개발 및 적용. 한국교육과정평가

원 연구보고 RRE 2016-11

- 박강윤, 이용상, 신동광 (2021). 순환신경망 장단기 기억(LSTM)을 이용한 자동채점의 가능성 탐색. *교육과정평가연구*, 24(4), 223-238
- 박도순, 권순달, 김명화, 김석우, 김영애, 김옥남, 김종필, 김진규, 김진하, 김형주, 남명호, 남현우, 박완성, 백순근, 성태제, 양길석, 어윤경, 이기종, 이나현, 이원석, 이종승, 이현주, 장은아, 조지민, 최대호 (2007) *교육평가-이해와 적용-*, 서울: 교육과학사
- 박세진, 하민수 (2020). 순환신경망을 적용한 초등학교 5학년 과학 서술형 평가 자동채점시스템 개발 및 활용 방안 모색. *교육평가연구*, 33(2), 297-321
- 박소영, 이병윤, 함은혜, 이유경, 이성혜 (2023). ChatGPT-4의 과학적 탐구 역량 평가 가능성 탐색: 인간평가자와의 비교를 중심으로. *Korean Journal of Educational Research*, 61(4), 299-332
- 박인용, 강혜진, 곽민호, 김미림, 김완수, 박도영, 박민호 (2023). 2022년 국가수준 학업성취도 평가 결과: 고등학교. 한국교육과정평가원 연구자료 ORM 2023-78-2
- 박종임, 이상하, 송민호, 이문복, 이민정, 최숙기 (2022). 컴퓨터 기반 서·논술형 평가를 위한 자동채점 방안 설계(I). 한국교육과정평가원 연구보고 RRE 2022-6
- 박종임, 송민호, 이상하, 박상욱, 최숙기, 이민정 (2023). 컴퓨터 기반 서·논술형 평가를 위한 자동채점 방안 설계(II). 한국교육과정평가원 연구보고 RRE 2023-7
- 배영우 (2024). 트랜스포머 아키텍처를 이용한 서술형 평가 자동채점 모델의 가능성 탐색. 서울대학교 교육전문대학원, 석사학위 논문
- 백순근 (2007). *교육측정의 이론과 실제*. 서울: 교육과학사.
- 백순근, 김혜숙, 함은혜, 최인희, 상경아, 이재열, 유예림, 황은희, 김지은, 길혜지 (2019). *교육평가의 이론과 실제*. 경기도 파주: 교육과학사
- 백종호, 심현표, 이동원 (2023). 과학 교과에서 인공지능 챗봇의 학생 평

- 가 결과 분석 및 활용 방안 탐색. 한국교육과정평가원 연구보고 ORM 2023-30-7
- 성정원, 신병철 (2023). ChatGPT를 활용한 서·논술형 평가 자동 채점 가능성 탐색: 세계지리 서·논술형 평가를 중심으로. 한국지리학회, 12(3), 415~432
- 손태권 (2023). ChatGPT의 수학교육 활용 가능성 탐색: 분수 문제에 관한 학생의 산출물과 예비교사의 담화 사례를 중심으로. 한국수학교육학회, 26(2), 99-113
- 송진웅, 강석진, 광영순, 김동건, 김수환, 나지연, 도종훈, 민병곤, 박성춘, 배성문, 손연아, 손정우, 오필석, 이준기, 이현정, 임혁, 정대홍, 정종훈, 김진희, 정용재 (2019). 미래세대를 위한 ‘과학교육표준’의 주요 내용과 특징. 한국과학교육학회지 39(3): 465-478
- 이동원, 심현표, 백종호 (2024). 과학 교과의 학생 평가에서 ChatGPT의 활용 가능성 및 교사 인식 탐색. Journal of the Korean Association for Science Education, 44(1), 119~130
- 이용상 (2015). 수능 체제 개편의 쟁점 및 현황 - 2021학년도 수능 체제 개편 관련 쟁점을 중심으로. 한국교육과정평가원 연구자료 ORM 2015-50-15
- 임중환, 임중현 (2024). ChatGPT API 활용 인공지능 논·서술형 평가 자동채점 프로그램 개발 실행연구. 교육혁신연구, 34(1), 349-370
- 장지현 (2021). 머신 러닝 기법을 활용한 영어 에세이 자동채점 방안 연구. 서울대학교 대학원, 박사학위 논문
- 정상목, 송기상, 한병래 (2005). LSA를 이용한 서술형 주관식 평가 시스템의 설계 및 구현. 한국정보교육학회, 9(2), 289-298
- 조희련, 이유미, 임현열, 차준우, 이찬규 (2021). 딥러닝 기반 언어모델을 이용한 한국어 학습자 쓰기 평가의 자동 점수 구간 분류. 한국언어문화학, 18(1), 217-241
- 진경애, 남명호, 김명화, 오상철, 김민정, 주형미, 신호필, 반재천, 김수경 (2006). 서답형 문항 자동 채점 프로그램 도입 방안 연구(I). 한

- 국교육과정평가원 연구보고 RRI 2006-6
- 진경애, 이병천, 주형미, 신동광 (2007). 서답형 문항 자동 채점 프로그램 도입 방안 연구(Ⅱ). 한국교육과정평가원 연구보고 RRE 2007-4
- 진경애, 이병천, 신동광, 박태준, 주현우 (2008). 서답형 문항 자동 채점 프로그램 도입 방안 연구(Ⅲ). 한국교육과정평가원 연구보고 RRE 2008-6
- 최숙기, 박종임 (2024). 생성형 AI를 활용한 현직 국어교사의 서·논술형 평가 문항 개발 양상 분석. 청람어문교육, 97, 243-270
- 하민수 (2016). 영어기반 컴퓨터자동채점모델과 기계번역을 활용한 서술형 한국어 응답 채점-자연선택개념평가 사례. 한국과학교육학회지, 36(3), 389-397.
- 하민수, 이경건, 신세인, 이준기, 최성철, 주재걸, 김남형, 이현주, 이종호, 이주림, 조용장, 강경필, 박지선 (2019). 학습 지원 도구로서의 서술형 평가 그리고 인공지능의 활용: WA3I 프로젝트 사례. 현장과학교육, 13(3). 271-282
- 하민수, 최진선, 임다니엘, 최성철, 주재걸, 신철균, 박지선 (2023). Development of an Unsupervised Learning-Based Automated Evaluation System of Descriptive Assessment. Brain, Digital & Learning. 13(4), 339-351
- 한재상, 김현주 (2023). 합성곱 신경망 기반 채점 모델 설계 및 적용을 통한 운동학 그래프 답안 자동 채점. 한국과학교육학회, 43(3), 237~251
- 허정만, 박소영 (2013). 투표 기반 서술형 주관식 답안 자동 채점 모델의 설계 및 구현. 한국컴퓨터정보학회논문지, 18(8)
- 홍선주, 조보경, 최인선, 박경진, 김현진, 박연정, 박정호 (2020). 학교 교육에서의 인공지능(AI) 활용 방안 탐색, 한국교육과정평가원 연구보고 RRI 2020-2
- Ajay HB, Tillett PI, Page EB (1973) Analysis of essays by computer (AEC-II) (No. 8-0102). Washington, DC: U.S. Department of

Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development

- Bagaria, V., Badve, M., Beldar, M., and Ghane, S. (2020). An intelligent system for evaluation of descriptive answers. In Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020 19 - 24.10.1109/ICISS49785.2020.9316110
- Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement (pp. 1-27). Hillsdale, NJ: LEA
- Beseiso M, Alzahrani S (2020) An empirical analysis of BERT embedding for automated essay scoring. Int J Adv Comput Sci Appl 11(10):204 - 210. <https://doi.org/10.14569/IJACSA.2020.0111027>
- Bleske-Rechek, Zeug, N., Webb, R. M. (2007). Discrepant performance on multiple-choice and short answer assessments and the relation of performance to general scholastic aptitude. Assessment and Evaluation in Higher Education, 32 (2), 89-105.
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., Krathwohl, D. R. (1956). Taxonomy of educational objectives: Handbook 1: Cognitive domain, New York: David McKay
- Bsharat, S. M., Myrzakhan, A., Shen, Z. (2024). Principled Instructions Are All You Need for Questioning LLaMA-1/2, GPT-3.5/4. <https://arxiv.org/abs/2312.16171v1>, last accessed 2024/01/14
- Dai, W.; Lin, J.; Jin, H.; Li, T.; Tsai, Y.-S.; Gašević, D.; Chen, G. Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In Proceedings of the 2023 IEEE International Conference on Advanced Learning Technologies

(ICALT), Orem, UT, USA, 10 - 13

- David R. Krathwohl (2002) A Revision of Bloom's Taxonomy: An Overview, *Theory Into Practice*, 41:4, 212-218
- Deng, H., Lyu, M. R., King, I., (2009). A generalized Co-HITS algorithm and its application to bipartite graphs, in: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, pp. 239 - 248
- Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment*, 5(1).
- Dong F, Zhang Y, Yang J (2017) Attention-based recurrent convolutional neural network for automatic essay scoring. In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* p 153 - 162
- Ekin, S. (2023). Prompt Engineering For ChatGPT- A Quick Guide To Techniques, Tips, And Best Practices. *TechRxiv*. May 04, 2023
- Foltz PW, Laham D, Landauer TK (1999) The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1, 2, [http:// imej.wfu.edu/articles/1999/2/04/ index.asp](http://imej.wfu.edu/articles/1999/2/04/index.asp)
- Foltz, P., Leacock, C., Rupp, A. A., Zhang, M. (2016). Understanding Automated Scoring: Theory and Practice. Paper presented at the National Council on Measurement in Education, Washington, D. C
- Gomaa WH, Fahmy AA (2020) Ans2vec: a scoring system for short answers, vol. 921, pp 586 - 595. [https:// doi.org/ 10.1007/ 978-3- 030- 14118- 9_59](https://doi.org/10.1007/978-3-030-14118-9_59)
- González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021).

- Artificial intelligence for student assessment: A systematic review. *Applied Sciences*, 11(12), 5467–5482
- Gronlund, N. E., (1988). *How to construct achievement tests*, NJ: Prentice–Hall, Inc
- Gustavo Pinto, Isadora Cardoso-Pereira, Danilo Monteiro, Danilo Lucena, Alberto Souza, and Kiev Gama. 2023. Large language models for education: Grading open-ended questions using chatgpt. In *Proceedings of the XXXVII Brazilian Symposium on Software Engineering*. 293 - 302
- Haller, S (2022) Survey on Automated Short Answer Grading with Deep Learning: From Word Embeddings to Transformers. arXiv:2204.03503
- He X, Gao M, Kan M-Y. (2017). Birank: towards ranking on bipartite graphs. *IEEE Trans Knowl Data Eng* 29(1):57 - 71
- Heilman, M., & Madnani, N. (2013). ETS: Domain adaptation and stacking for short answer scoring. *Second Joint Conference on Lexical and Computational Semantics(*SEM)*. *Proceedings of the seventh international Workshop on semantic evaluation (SemEval 2013)*, 2(SemEval): Vol. 2, (pp. 275 - 279)
- Jeon, J. H. and Lee, S. (2023). *Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT*. Springer Science Business Media. vol. 28. no. 12
- Jiang, L., Bosch, N. (2024) Short answer scoring with GPT-4, ACM ISBN. <https://doi.org/10.1145/3657604.3664685>
- Kastner, M., Stangl, B. (2011) Multiple Choice and Constructed Response Tests: Do Test Format and Scoring Matter?. *Procedia Social and Behavioral Sciences*, 12, 263–273
- Kaur, A. Sasikumar, M., (2017). A comparative analysis of various

- approaches for automated assessment of descriptive answers, International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, 10.1109/ICCIDS.2017.8272650
- Küchemann, S., Steinert, S., Revenga, N., Schweinberger, M., Dinc, Y., Avila, K. E., et al. (2023). 'Can ChatGPT support prospective teachers in physics task development?', *Physical review. Phys. Educ. Res.* 19:020128. doi: 10.1103/PhysRevPhysEducRes.19.020128
- Kumar, Y., Aggarwal, S., Mahata, D., Shah, R. R., Kumaraguru, P., Zimmermann, R. (2019). Get IT Scored Using AutoSAS-An Automated System for Scoring Short Answers. In *AAAI*
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.+
- Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice*, 23(3), 6 - 14
- Laverty, J. T.; Underwood, S. M.; Matz, R. L.; Posey, L. A.; Carmel, J. H.; Caballero, M. D.; Fata-Hartley, C. L.; Ebert-May, D.; Jardeleza, S. E.; Cooper, M. M. Characterizing College Science Assessments: The Three-Dimensional Learning Assessment Protocol. *PLoS One* 2016, 11, e0162333
- Lee, J. (2023). Exploring the Possibility of Automatic Scoring for Graphical Responses Using a Convolutional Neural Network. *New Physics: Sae Mulli*, 73(2), 138-149
- Lee, H. S., Liu O. L., Linn M. C. (2011). Validating Measurement of Knowledge Integration in Science Using Multiple-Choice and Explanation Items. *Applied Measurement in Education* 24(2): 115-136
- Leacock, C. & Chodorow, M. (2003). C-rater: Automated Scoring of

- Short-answer Questions. *Computers and the Humanities* 37. pp. 389-405
- Li, Z., Wu, X.-M., Chang, S.-F., (2012). Segmentation using superpixels: A bipartite graph partitioning approach. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 789 - 796
- Li-Hsin Chang and Filip Ginter. 2024. Automatic Short Answer Grading for Finnish with ChatGPT. In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38. AAAI Press, Palo Alto, CA, 23173 - 23181
- Lissitz, R. W., Huo, X., Slater, S. C. (2012). The contribution of constructed response items to large scale assessment: Measuring and understanding their impact, *J. Appl. Test. Tech.* 13, 1
- Liu, O.L., C. Brew, J. Blackmore, L. Gerard, J. Madhok, and M.C. Linn. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice* 33(2): 19 - 28 .
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). Gptheval: Nlg evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.
- Livingston, S. A. (2009). Constructed-Response Test Questions: Why We Use Them; How We Score Them. R&D Connections. Number 11. In Educational Testing Service. Educational Testing Service
- Logan IV R. L., Liu, N. F., Peters, M. E. (2019). Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In: Proceedings of the Conference of the Association for Computational Linguistics. Florence, 5962 - 5971

- Lucas B. Galhardi, Helen Senefonte, Rodrigo de Souza, and Jacques Brancher. 2018. Exploring Distinct Features for Automatic Short Answer Grading. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*. SBC, Porto Alegre, RS, Brasil, 1 - 12
- Lukhele, R., Thissen, D., Wainer, H. (1993). On the Relative Value of Multiple-Choice, Constructed Response, and Examinee Selected Items on Two Achievement Tests. *Program Statistics Research Technical Report No. 93-28*
- Lun J, Zhu J, Tang Y, Yang M (2020) Multiple data augmentation strategies for improving performance on automatic short answer scoring. In: *AAAI 2020 – 34th AAAI conference on artificial intelligence*, pp 13446 - 13453
- Madnani, N., Loukina, A., & Cahill, A. (2017a). A large scale quantitative exploration of modeling strategies for content scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 457 - 467)
- Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021, June). Multi-class confusion matrix reduction method and its application on net promoter score classification problem. In *Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference* (pp. 412-419)
- McMillan, J. H. (2014). *Classroom Assessment: Principles and Practice for Effective Standards-Based Instruction*(7th Ed.). Boston, MA: PEARSON.
- Melnyk, I., Dognin, P., Das, P. (2022). Knowledge Graph Generation From Text. *Findings of the Association for Computational Linguistics: EMNLP 2022*, 1610-1622

- Moharreri, K., Ha, M., & Nehm, R.H. (2014). Evograder: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach*, 7(1), 1 - 14
- Naudé, K. A., Greyling, J. H., Vogts, D. (2010). Marking student programs using graph similarity. *Computers & Education* 54, 2, 545 - 561
- Nayak, G., Dutta, S., Ajwani, D., Nicholson, P., & Sala, A. (2019). Automated assessment of knowledge hierarchy evolution: Comparing directed acyclic graphs. *Information Retrieval Journal*, 22(3), 256 - 284
- Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, 21(1), 183 - 196.
- Nielsen, R. D., Ward, W. Martin J. H. (2008). Classification errors in a domain-independent assessment system. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 10 - 18, Columbus, Ohio, June. Association for Computational Linguistics
- Opfer JE, Nehm RH & Ha M (2012) Cognitive foundations for science assessment design: Knowing what students know about evolution. *Journal of Research in Science Teaching* 49:744-777
- Page. E. B (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 47(5), 238 - 243
- Parker, J. L., Carroca, C. (2024). ChatGPT for Automated Writing Evaluation in Scholarly Writing Instruction, 62(12), 721-727
- Pavlopoulos GA, Kontou PI, Pavlopoulou A, (2018). Bipartite graphs in systems biology and medicine: a survey of methods and

applications. GigaScience

- Prabhudesai A, Duong TNB (2019) Automatic short answer grading using siamese bidirectional LSTM based regression. In: TALE 2019 - 2019 IEEE international conference on engineering, technology and education
- 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>
- Riordan, B., Horbach, A., Cahill, A., Zesch, T., Lee C. M. (2017). Investigating neural architectures for short answer scoring. In BEA@EMNLP.
- Roy, S., Bhatt, H. S., Narahari, Y. (2016) An Iterative Transfer Learning Based Ensemble Technique for Automatic Short Answer Grading. ArXiv abs/1609.04909 (2016)
- Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, 1(2).
- Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of intellimetric essay scoring system. *The Journal of Technology, Learning and Assessment*, 4
- Saha, S., Dhamecha, T. I., Marvaniya, S., Sindhgatta, R., Sengupta, B. (2018). Sentence Level or Token Level Features for Automatic Short Answer Grading?: Use Both. In *Artificial Intelligence in Education*, Carolyn Penstein Rosé, Roberto Martínez-Maldonado, H. Ulrich Hoppe, Rose Luckin, Manolis Mavrikis, Kaska Porayska-Pomsta, Bruce McLaren, and Benedict du Boulay (Eds.). Springer International Publishing, Cham, 503 - 517.

- Sakaguchi, K., Heilman, M., & Madnani, N. (2015). Effective feature integration for automated short answer scoring. In Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies (pp. 1049 - 1054)
- Sanfilippo A. (2006). Graph Theory. *Encyclopedia of Language & Linguistics* (Second Edition), 311(3):140-142. doi: 10.1016/B0-08-044854-2/01600-X
- Schneider, J., Schenk, B., Niklaus, C., Vlachos, M. (2023). Towards llm-based autograding for short textual answers. arXiv preprint arXiv:2309.11508
- Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29(7), 4 - 14.
- Shermis MD, Mzumara HR, Olson J, Harrington S (2001) On-line grading of student essays: PEG goes on the World Wide Web. *Assess Eval High Educ* 26(3):247 - 259
- Snow, R. E. (1993). Construct validity and construed response tests. In R. E. Bennett, & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates
- Süzen, N., Gorban, A. N., Levesley, J., & Mirkes, E. M. (2020). Automatic short answer grading and feed back using text mining methods. *Procedia Computer Science*, 169, 726 - 743
- Sukkarieh, J. Z. & Blackmore, J.(2009). C-rater: Automatic Content Scoring for Short Constructed Response. *Proceedings of the Twenty-Second International FLAIRS Conference*. pp. 290-295
- Tan, H., Wang, C., Duan, Q., Lu, Y., Zhang, H., & Li, R. (2020). Automatic short answer grading by encoding student responses via a graph convolutional network. *Interactive*

- Learning Environments, 0, 1 - 15.
<https://doi.org/10.1080/10494820.2020.1855207>
- Tulu, C. N., Ozkaya, O., & Orhan, U. (2021). Automatic Short Answer Grading With SemSpace Sense Vectors and MaLSTM. *IEEE Access*, 9, 19270–19280.
<https://doi.org/10.1109/ACCESS.2021.3054346>
- Vujosevic-Janjic, M., Nikolic, M., Tomic, D., Kuncak, V. (2013). Software verification and graph similarity for automated evaluation of students' assignments. *Information and Software Technology* 55, 6, 1004 - 1016
- Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2005). Automated scoring for creative problem-solving ability with ideation-explanation modeling. In *Proceedings of the Thirteenth International Conference on Computers in Education* (pp. 522 - 529). Singapore: IOS Press
- Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., ... & Zhang, S. (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E. Le, Q., Zhou, D (2023). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*
- Weston, M., Parker, J., & Urban-Lurain, M. (2013). Comparing formative feedback reports: Human and automated text analysis of constructed response questions in biology. In *NARST annual conference, Rio Grande, Puerto Rico*.
- Zhang, M. (2013). Contrasting Automated and Human Scoring of Essays. *ETS R&D Connections*, 21. Retrieved 27, Jan, 2016, from http://www.ets.org/Media/Research/pdf/RD_Connections_21.pdf

Abstract

Automated Scoring of Descriptive Assessment Using Graph Theory

JinWoo Kim

Department of Science Education(Physics Education)

The Graduate School

Seoul National University

Creating questions that are appropriate to the goals of educational assessment, such as checking the achievement of educational goals, measuring learners' cognitive abilities, improving the learning process, and providing feedback, is a major area of interest in educational research. Among them, evaluation using descriptive assessment is very important because it provides a more detailed view of the response characteristics and problem-solving process, which can effectively check the level of learners' understanding and provide customized feedback to learners. Nevertheless, the main reasons for the predominant use of selective assessment in the field are the effort and cost of the grading process and the reliability of the assessment results.

Automated scoring research is actively underway to solve the problem of scoring descriptive assessments with computer technology. Automated scoring research, which started with statistical-based technology that checks the occurrence of correct keywords, can be divided into two main types: machine learning technology that extracts features and learns classification models, and deep learning technology that learns models based on artificial neural networks that do not use features. Recently, there has been an active research effort to utilize generative AI (Generative Artificial Intelligence), which can be used immediately without training, for automated scoring. However, previous studies require a large amount of high-quality data to ensure model accuracy and have difficulties in extracting valid qualities. In addition, the built models can only be applied to the training assessment, resulting in poor generalization ability. Additionally, machine learning models struggle to extract valid qualities, and deep learning models have difficulty interpreting scoring results. Studies utilizing generative AI are also limited by the high level and variation of responses, making it difficult to secure reliability. Therefore, it is necessary to study a model that does not require a large amount of data with high reliability and validity, is easy to extract qualities and interpret scoring results, and can return results and feedback immediately.

In this study, this paper proposes an automatic scoring model using graph theory that meets the above goals. It utilizes generative AI with excellent natural language processing capabilities to extract the triplets of text answers, and then transforms the bipartite graph into an adjacency matrix to quantify the similarity of the graph to the reference answer. In this process, similar words and nouns that are characteristic of the text are identified using generative AI and

reflected as correct answer factors to improve the accuracy of the model. In addition, the relationship between the extracted triplet clauses is expressed as a causal graph, and the logical relationship between clauses can be visualized as a graph. Based on the similarity of the two graphs, the similarity of the reference answer and the student answer is expressed as a continuous score, and based on this, the automatic grading score is returned according to the grading criteria. We categorized the scoring criteria into two types. Scoring criterion 1 is the scoring method of the model proposed in this study, which is scored by matching the reference answer, and scoring criterion 2 is the scoring method of whether the answer is scientifically valid including the reference answer.

To evaluate the performance of the graph model proposed in this study, ChatGPT4o was set as a control group, and 137 data from first-year students at a science high school in Seoul were automatically scored on the role and principle of generators. We evaluated the performance with five indicators, including accuracy compared to manual grading, and found that the graph model, which showed an accuracy of .934 and .763 for grading criteria 1 and 2, respectively, outperformed ChatGPT4o, which showed an accuracy of .511 and .431 for grading criteria 1 and 2, respectively. It can be seen that the Graph model performs very well on criterion 1, which is scoring by similarity to the reference answer, but performs poorly on criterion 2, which is scoring by scientifically valid answers. To compensate for this, we added a cautionary word to the user's score, which was then factored into the final score.

Graph models can quantify the understanding of concepts related to a question as a continuous number through graph similarity, even among students with the same score, enabling concept checks and

personalized feedback, and evaluate and provide feedback on whether students have written answers with scientifically valid logical expressions through the presence of logical relationships.

In this study, we propose a model for automatic scoring of content-based descriptive assessment using graph theory and suggest educational applications. This model is composed of three parts. Answers are quantified as continuous score, and the scoring factors are clear, making it easy for users to interpret the scoring results. Second. Teachers can recognize the concepts that need to be supplemented or the relationship between concepts and immediately reflect them in the feedback. Third. It is significant in that it does not require a learning process and does not require a large amount of high-quality data for learning, so it can be used from the classroom level to large-scale exams. However, it is necessary to elaborate the graph model by developing a method for identifying incorrect answer elements contained in correct answers and adopting them as deduction factors, and further research to broaden the scoring possibilities of similar answers through various reference answers.

keywords : Graph theory, Automated scoring , Descriptive Assessment, Causal relationships, Generative Artificial Intelligence

Student Number : 2022-26856