

생성형 AI 모델을 활용한 요약 성능 평가 연구

최규리^o, 박서윤, 강예지, 김한샘[†]

연세대학교 언어정보학협동과정
{gyuri1345, seoyoon.park, yjkang5009, khss}@yonsei.ac.kr

A Study on Evaluating Summarization Performance using Generative AI Model

Gyuri Choi^o, Seoyoon Park, Yejee Kang, Hansaem Kim[†]

Interdisciplinary Graduate Program of Linguistics and Informatics, Yonsei University

요약

인간의 수동 평가 시 시간과 비용의 소모, 주석자 간의 의견 불일치, 평가 결과의 품질 등 불가피한 한계가 발생한다. 본 논문에서는 맥락을 고려하고 긴 문장 입출력이 가능한 ChatGPT를 활용한 한국어 요약문 평가가 인간 평가를 대체하거나 보조하는 것이 가능한가에 대해 살펴보았다. 이를 위해 ChatGPT가 생성한 요약문에 정량적 평가와 정성적 평가를 진행하였으며 정량적 지표로 BERTScore, 정성적 지표로는 일관성, 관련성, 문법성, 유창성을 사용하였다. 평가 결과 ChatGPT4의 경우 인간 수동 평가를 보조할 수 있는 가능성이 있음을 확인하였다. ChatGPT가 영어 기반으로 학습된 모델임을 고려하여 오류 발견 성능을 검증하고자 한국어 오류 요약문으로 추가 평가를 진행하였다. 그 결과 ChatGPT3.5와 ChatGPT4의 오류 요약 평가 성능은 불안정하여 인간을 보조하기에는 아직 어려움이 있음을 확인하였다.

주제어: 요약 평가, 수동 평가, ChatGPT, BERTScore

1. 서론

인터넷과 디지털 미디어의 폭발적인 증가로 수많은 정보가 생산되고 공유됨에 따라 개인 및 기업에서는 엄청난 양의 정보를 처리해야 한다. 이러한 정보 폭증 상황에서는 필요한 정보를 정확하고 빠르게 추출하는 요약 태스크가 핵심적인 역할을 한다. 자연어 처리와 기계학습 기술의 발전으로 자동 요약 알고리즘이 더욱 더 정교해졌으며, 생성 AI모델의 출현으로 대량의 텍스트 데이터에 대해 핵심내용을 빠르게 추출해 내어 사람이 수동으로 요약문을 작성했던 이전과 비교하여 시간과 노력을 절약할 수 있게 되었다. 한편 생성 AI 모델이 도출해 낸 요약문에 대해 인간의 직관과 부합하는지를 평가하는 과정이 필요하다. 하지만 인간 주석자를 통해 진행하는 수동 요약 평가는 주석자들 간의 평가 일관성, 평가 질의 차이, 시간과 비용의 소요 등 한계점이 있다. 최근에는 기술의 발전으로 요약문에 대한 평가 역시 생성 언어 모델을 활용하여 진행할 수 있게 되었다. 그러나 생성 언어 모델은 토큰 단위로 언어 처리를 수행하는데 한국어는 처리에 영어보다 더 많은 토큰을 필요로 하여 긴 문장을 함께 입력해야 하는 요약 평가에는 적용이 쉽지 않다. 또한 맥락 고려가 어려운 단일 질문-단일 응답으로 처리되는 모델은 요약 평가에 부적합하다. 이러한 한계점을 극복하고자 본 연구에서는 맥락 고려가 가능하고 긴 문장 처리에 적합한 ChatGPT를 활용하여 생성 언어 모델의 평가가 인간의 수동 요약 평가를 대체하거나 보조할 수 있는가에 대하여 연구를 진행하였다. 이를 위해 국립국어원의 신문 말뭉치를 대상으로 ChatGPT3.5와 ChatGPT4를 이용하여 요약문을 생성하도록 하고, 인간 평가 및 모델의 자동 평가 결과를 비교 분석하였다.

2. 관련 연구

2.1 거대 언어 모델을 활용한 요약 및 평가 연구

거대 언어 모델을 활용한 요약 및 요약평가에 대한 연구가 활발하게 진행되고 있다. 언어 모델을 활용한 요약 연구로는 [1], [2]가 있다. [1]은 언어 모델의 인간 요약 유사성을 확인하고자 10개의 언어 모델의 요약을 분석하여 특징과 요약 선호도를 검토하였으며, 이를 통해 언어 모델의 요약이 인간 요약과 유사함을 확인했다는 데 의의가 있다. [2]는 ChatGPT의 요약 성능에 대해 검토하여 ChatGPT의 장점과 한계를 확인하였고 일관성 부족, 환각 효과, 사실성 탐지 도구의 필요성을 한계점으로 제시하였다. ChatGPT를 포함하여 거대 언어 모델의 장점을 확인했다는 점에 활용 가치가 있다. 언어 모델의 요약 평가에 대한 연구인 [3]은 다양한 인간 평가 지표와 자동 평가 지표를 활용한 점이 의의가 있다. ChatGPT의 요약 평가 능력에 대해 인간과의 차이를 비교하여 인간보다 효율적이고 재현성이 높은 평가 능력을 확인했지만 함께 출력된 설명의 사실성은 고려하지 않은 아쉬움이 있다. [4]는 ChatGPT를 비롯한 거대 언어 모델에게 인간평가와 동일한 프롬프트를 제시하여 인간 평가와 언어 모델의 평가가 유사함을 발견했다. ChatGPT가 인간 평가자와 유사한 평가와 설명이 가능하고 프롬프트의 영향을 많이 받지 않음을 확인한 점이 주목할 만하다. [5]는 ChatGPT의 자연어 생성 지표 가능성을 탐구하였다. 요약, 스토리 생성, 데이터-텍스트 변환 작업 등에 대해 자동 평가와 비교하여 ChatGPT가 평가 지표로 활용 가능함을 확인한 데에 의의가 있다.

2.2. 생성 요약의 정량적 평가

일반적으로 생성 요약에 대한 정량적 평가에는 ROUGE 점수가 사용된다[6]. ROUGE 점수는 정답 요약문과 자동 요약문을 재현율(recall)과 n-gram을 바탕으로 계산되며, ROUGE-1, ROUGE-2, ROUGE-L 등이 있다. ROUGE-1은 단일 단어(uni-gram)의 일치도를 기반으로 하고, ROUGE-2는 연속한 두 단어의 일치도를 기반으로 하며, ROUGE-L의 경우 n-gram 순서를 고려하지 않고 텍스트 안에서 나타나는 최장의 길이 문자열을 바탕으로 일치도를 계산한다. 이처럼 ROUGE 점수는 문장의 구조와 단어 사용에 따라 좌우되기 때문에 추상 요약(abstractive summarization)을 평가하는 데 한계가 있다. 한편 문맥 임베딩을 사용하여 토큰 유사성을 계산하는 BERTScore는 후보 문장의 각 토큰과 참조 문장의 토큰 간 유사도를 계산하는 텍스트 생성 자동 평가 지표이다. [7]은 BERTScore가 기존의 메트릭과 비교했을 때 인간의 평가와 더 높은 상관관계를 가지고 있다고 밝혔다. 이를 참고로 하여 본 연구에서는 모델이 출력한 요약문에 대한 정량적 평가를 위해 문맥 임베딩 기반의 BERTScore를 사용하였다.

2.3. 생성 요약의 정성적 평가

생성 요약문은 정량적 평가 점수가 인간의 언어적 직관에 얼마나 부합하는지 분석하기 위해 정성적 평가가 진행되어야 한다. 이를 위해 연구마다 정의한 정성적 지표가 다른데, [8]에서는 ‘유창성, 일관성, 응집성, 담화 관계, 상호 참조 정보, 의도 식별’을 지표로 설정하고 각 지표에 대해 -1~1까지의 점수를 매기도록 하였다. [9]에서는 두 명의 언어 전문가에게 -1~1로 요약문에 대해 점수를 매기도록 하고, 두 점수 간 관계를 cohen’s kappa coefficient로 검증하였다.

3. 데이터셋

국립국어원 모두의 말뭉치에서 구축한 신문 데이터 말뭉치와 문서 요약 말뭉치[10]¹를 활용하였다. 신문 말뭉치는 신문 기사 원문을 제공하고 문서 요약 말뭉치는 주제 문장과 요약문을 제공한다. 두 말뭉치는 각각 개별적인 말뭉치이나, 신문 기사 원문과 문서 요약문의 ID가 동일한 점을 활용하여 ID가 같은 기사-요약문을 매칭하여 사용하였다. 신문 기사를 원문 텍스트로 설정하였고, 문서 요약 말뭉치의 요약문을 ChatGPT가 생성한 요약문과 비교할 수 있는 정답 요약문으로 설정하였다.

4. 실험

대화형 생성 언어 모델인 ChatGPT3.5와 ChatGPT4로 실험을 하였다. ChatGPT는 대화를 목적으로 제작되어 문맥을 고려한 대화가 가능하고 긴 문장 처리에 용이하여

ChatGPT를 활용하여 실험하였다. [11]에 따르면 ChatGPT 3.5는 텍스트에 대한 처리가 가능하며 영문 기준 3천 단어까지 입력/출력이 가능하다. ChatGPT4는 GPT3.5보다 8배 많은 영문 기준 2만 5천 단어 입력/출력이 가능하다. 이는 더 많은 양의 텍스트 입력이 가능하게 하여 정확성과 관련성이 향상될 수 있다. ChatGPT-3.5보다 ChatGPT4가 정교한 언어 이해와 처리가 가능한 점, 한국어에 대한 성능이 향상된 점 등에 차이점이 있어 두 모델을 비교하였다.

[1]에서는 CNN/DM, XSUM을 대상으로 100건씩의 영어 신문 데이터를 분석한 바 있고, [2]에서도 뉴스 기사를 포함해 100건의 데이터를 분석하였다. 본 연구에서도 선행 연구의 데이터 규모에 맞추어 100건의 한국어 신문 데이터를 분석하였다. 요약문 생성 과정에서 ChatGPT3.5와 ChatGPT4에게 제공한 프롬프트는 동일하다. 요약문 생성을 위해 신문 기사 원문을 ‘서론-본론-결론’ 3단락으로 구분하여 제시하고 zero-shot으로 ChatGPT가 각 단락을 한 문장으로 요약하도록 했다. 요약한 문장을 토대로 전체 기사 원문을 3-4문장으로 한정하여 요약하도록 했다. ChatGPT-3.5와 4를 통해 생성한 요약문의 예시는 표 1과 같다.

표 1 ChatGPT 생성 요약문 예시

ChatGPT3.5	두 명의 미국 여기자가 중국-북한 국경 두만강에서 억류되었으며, 현재 평양에서 조사를 받고 있습니다. 이들의 석방 여부는 협상 결과에 따라 결정될 것으로 보입니다. 또 다른 여기자 두 명과 함께한 미국인 카메라기자와 중국인 가이드는 중국 국경수비대에 체포되지 않았으며, 자신들의 판단에 따라 신고를 한 후 자유로운 상태입니다. 이들이 국경을 넘은 것이 실수가 아니라고 현지 주민과 관계자들이 주장하고 있습니다.
ChatGPT4	중국 지린 성 투먼의 북한 접경지역에서 억류된 두 미국 여기자는 평양으로 옮겨져 조사를 받는 중이며, 이 사안은 미국 국무장관 힐러리 클린턴의 관심을 받고 있다. 변경 취재를 동행한 카메라기자와 가이드는 중국 공안당국에 자진 신고하였고, 현재 자유의 몸이다. 미국 여기자들이 실수로 국경을 넘었다는 주장은 현지 상황과 일치하지 않는다.

5. 요약 성능 평가

5.1 요약문 평가

[7]을 참고하여 ChatGPT가 생성한 요약문을 평가하기 위한 정량적 평가 지표로 BERTScore를 사용하였다. 일반적으로 ROUGE가 자동 요약 평가에서 문서의 품질을 측정하는 지표 중 하나로 원문과 요약본 사이의 유사성을 판

¹ 국립국어원, 국립국어원 문서 요약 말뭉치(버전 1.0), <https://corpus.korean.go.kr>, 2020

단하는데 활용된다. 하지만 ROUGE 점수의 경우 n-gram을 기반으로 연속된 단어 간의 일치도를 기반으로 계산되기 때문에 문장의 구조 및 어휘 사용의 영향이 점수에 큰 영향을 미친다. 따라서 패러프레이징된 추상 요약 결과의 평가에 적용하기 부적절하다. BERTScore는 사전 훈련된 임베딩을 활용하여 토큰의 유사성에 따라 상관관계를 분석하므로 문장이 내포하고 있는 의미를 고려하여 평가할 수 있다. 본 연구에서 ChatGPT가 생성한 요약문은 추상 요약으로 생성된 데이터이기 때문에 BERTScore의 문맥 임베딩을 활용하여 정답 요약문과의 의미적 유사도를 측정하였다.

표 2 ChatGPT BERTScore 비교

	정답요약&GPT3.5요약	정답요약&GPT4요약
BERTScore	0.714	0.716

ChatGPT3.5와 ChatGPT4의 요약문 100건에 대한 두 모델 각각의 평균 BERTScore 점수인 표 2에서 ChatGPT3.5와 ChatGPT4의 점수 차이가 크지 않음을 확인할 수 있다. 두 요약 모두 BERTScore 0.71 이상으로 정답 요약과 어느 정도의 높은 유사성을 가진 내용으로 생성되었음을 정량적 평가를 통해 확인할 수 있다.

ChatGPT가 생성한 요약문에 대한 정성적 평가는 리커트 척도를 활용하여, 0점부터 5점까지 주관적으로 점수를 부여하는 방식으로 진행한다. 평가 지표는 [4], [12]를 참고하여 설정하였다. [12]는 4가지 지표를 설정하였다. 문장의 전체적인 품질을 문장 간의 관계가 이어져 주제에 대한 일관된 정보로 구성이 되었는지에 대한 Coherence(일관성)으로 평가하고, Consistency(일관성)는 환각효과에 대한 평가를 포함하며 원본 문서에서 언급된 내용만으로 일관되게 요약이 되었는지에 대해 평가한다. Fluency(유창성)는 문장의 이해가 어렵거나, 대소문자, 문장 구성 요소의 누락에 대해 평가한다. Relevance(관련성)는 과도하거나 중복된 정보가 없는지, 원문의 중요한 정보가 요약문에 포함되어 있는지에 대해 평가한다. 본 연구에서는 사전연구를 참고하여 한국어 요약 결과를 평가하는 지표로 ‘일관성, 문법성, 관련성, 유창성’을 설정하였다. 본 연구의 정성적 평가지표는 표 3과 같다.

표 3 정성적 평가 지표

(1) 일관성: 전체적인 요약의 서론-본론-결론의 내용이 주제에 대해 일관적으로 이어지는지 평가한다.
(2) 관련성: 요약문에 포함된 정보가 원문 기사문과 관련이 있는지에 대해 판단한다. 환각효과 여부도 포함하여 평가한다.
(3) 문법성: 어미, 조사 등을 포함한 문장의 문법에 대한 오류를 판단한다.
(4) 유창성: 요약문을 읽는 과정에서 문장의 흐름이 자연스러운지 판단한다. 지나치게 문장의 호흡이 길거나, 앞의 문장을 다시 읽어야 할 만큼 직관적인 이해가 어려운지 등을 평가한다.

‘일관성’은 주제에 대한 전체 내용이 일관적인지 평가

하고자 설정하였다. ‘관련성’은 환각효과로 인한 거짓 정보, 과장, 과소 표현된 정보를 제외한 원문과 관련된 정보만 포함하고 있는지에 대한 평가가 필요하다고 판단하여 설정하였다. [12]의 ‘유창성’에서 포함한 대소문자, 문장 구성에 대한 평가의 적용은 한국어에 적합하지 않다고 생각하여 한국어 평가에 적합한 ‘문법성’ 지표를 설정하였다[4, 13]. [13]에서 외국인 학습자들이 습득하기 어려워하는 항목으로 조사, 어미, 높임 표현, 사동/피동, 부정, 활용 등을 제시하고 있어 ChatGPT가 외국어로 구축된 모델임을 고려해 ‘문법성’ 지표를 구성하였다. [14]에서 ‘유창성’은 글의 흐름이 자연스럽고, 원어민들이 사용하는 언어에 가까운 표현을 다양하게 사용하는 것으로 정의한다. 이에 본 연구에서도 요약문의 자연스러움을 평가하는 지표로 ‘유창성’을 선정하였다. 이렇게 설정한 평가 지표를 원문, 요약문과 함께 프롬프트로 제공했다. 이때 ChatGPT3.5가 생성한 요약문을 ChatGPT3.5가 평가하고, ChatGPT4가 생성한 요약문을 ChatGPT4가 평가하도록 하였다. ChatGPT3.5와 ChatGPT4가 정성적으로 평가한 결과와 국어학을 전공한 대학원생 2인이 평가한 인간 평가 점수를 평균하여 비교하였다. 평가 지표에 따른 평가 점수의 평균은 다음과 같다.

표 4 ChatGPT3.5 요약문 정성적 평가 결과

	일관성	관련성	문법성	유창성
ChatGPT3.5평가	4.7	5.0	4.9	4.4
인간 평가	4.2	4.3	4.2	4.2

표 5 ChatGPT4 요약문 정성적 평가 결과

	일관성	관련성	문법성	유창성
ChatGPT4평가	4.5	4.8	4.9	4.5
인간 평가	4.5	4.5	4.7	4.5

표 4에서 ChatGPT3.5의 정성 평가는 ‘관련성’이 고득점, ‘유창성’이 저득점으로 평가되었다. 인간 평가는 ChatGPT3.5와 동일하게 ‘관련성’이 높았으나 나머지 항목은 4.2점으로 ChatGPT3.5와 평가에 차이가 있음을 확인할 수 있었다. ChatGPT4의 정성 평가는 ‘문법성’이 높았고 ‘일관성’과 ‘유창성’이 낮았으며 인간 평가도 ‘문법성’에 높은 점수를 준 점이 일치했다. 또한 인간이 부여한 평가 점수보다 ChatGPT3.5와 ChatGPT4가 부여한 점수가 공통적으로 높은 편이고 특히 ChatGPT3.5와 인간의 평가 점수에 격차가 있음을 함께 확인할 수 있었다. 표 4와 표 5의 ChatGPT3.5, ChatGPT4와 인간 평가의 상관관계를 보고자 스피어만 상관계수로 분석하였다. 스피어만 상관계수는 두 변수의 순위 사이 통계적 의존성을 측정에 사용되는 비모수적인 상관계수로 -1에서 1까지의 값으로 표현된다.

표 6 정성적 평가 결과 스피어만 상관계수

	인간평가 & ChatGPT3.5	인간평가 & ChatGPT4
일관성	-0.030	0.723
관련성	-0.040	0.508
문법성	-0.101	0.406

유창성	0.038	0.565
-----	-------	-------

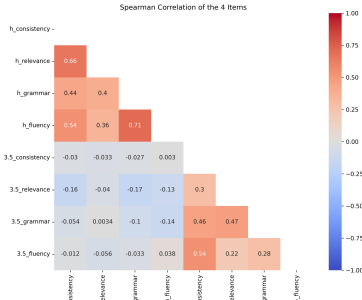


그림 1 인간&ChatGPT3.5 상관관계 시각화

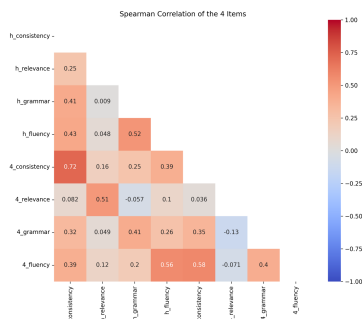


그림 2 인간&ChatGPT4 상관관계 시각화

표 6에서 ChatGPT3.5와 인간평가 결과는 모든 평가 지표의 상관관계수가 0에 가까운 수치로, 유사성이 거의 없다고 볼 수 있다. 표 4에서 두 평가 주체(ChatGPT3.5, 인간) 간의 점수 차이가 작지 않은 것도 이를 뒷받침한다. 반면 표 6의 ChatGPT4와 인간평가 결과에 따라 ChatGPT4가 인간과 유사한 경향이 있는 것으로 보인다. 특히 ‘일관성’에 대해 가장 인간과 유사하게 평가했으며 나머지 평가 지표들의 상관관계수도 ChatGPT3.5와 비교하여 높은 편이다. 표 6을 시각화한 그림 1과 2로 인간 평가와 ChatGPT3.5, ChatGPT4의 차이를 확인할 수 있다.

정량적 평가 결과는 ChatGPT3.5와 ChatGPT4 모두 정답 요약문과 유사성을 가진다고 평가했다. 그러나 평가 지표를 활용한 정성적 평가 결과는 다소 차이가 있었다. 정성적 평가 결과 인간, ChatGPT3.5, ChatGPT4 평가 결과 간의 차이가 있었다. ChatGPT 3.5의 평가 결과는 인간 평가와 유사하다고 보기 어려운 반면 ChatGPT4는 비교적 인간 평가와 유사하게 평가할 수 있음을 확인했다.

5.2 오류 요약문 평가

앞서 진행한 요약문 평가에서 ChatGPT3.5와 ChatGPT4의 정성적 평가 성능에 차이가 있었다. 이에 ChatGPT가 정문으로 작성된 요약문뿐만 아니라 오류가 있는 요약문도 평가가 가능한지 추가로 평가를 진행하였다. 평가 지표에 따라 한국어 오류 요약문을 평가하고 적절한 점수를 부여하는 것이 가능한지 살펴봄으로써 오류 요약문 평가 성능을 확인하는 것이 목적이다. 정답 요약문 데이터에 오타, 조사 및 어미 오류, 일관성을 해치거나 요약문과 관련이 없는 정보를 첨가하는 등의 방식으로 오류가 있는 요약문을 100건 생성하였다. 평가 프롬프트는

정문 요약문 평가와 동일한 프롬프트를 제공하여 오류 요약문을 정성적으로 평가하도록 했다. 다음은 오류 요약문에 대한 인간 평가와 ChatGPT3.5와 ChatGPT4의 정성적 평가 결과를 스피어만 상관계수로 분석한 결과이다.

표 7 오류 요약문 정성적 평가 결과 스피어만 상관계수

	인간평가& ChatGPT3.5	인간평가& ChatGPT4	ChatGPT3.5 & ChatGPT4
일관성	0.004	0.156	0.747
관련성	0.170	0.270	0.618
문법성	-0.007	0.042	0.674
유창성	-0.072	0.013	0.772
평균	0.024	0.120	0.703

표 7에서 ChatGPT3.5의 인간 평가 상관관계 값이 대부분 0에 가깝고, 전체 평가 지표의 평균이 0.024라는 것을 통해 모델 평가와 인간 평가 사이에 상관관계가 거의 없음을 알 수 있다. 즉 ChatGPT3.5는 인간과 유사한 평가 능력을 갖추지 못했음을 추론할 수 있다. ChatGPT4는 ‘관련성’의 값이 상대적으로 높아 인간과 약간의 유사한 상관관계를 가지는 편이나 나머지 항목들은 아직 인간과 비슷한 평가를 한다고 보기에는 부족함이 있다. 다만 ChatGPT3.5와 ChatGPT4 모두 오류 요약문 평가에서 인간 평가와 비교하여 상관관계가 낮았지만 평가 점수와 함께 제시한 설명에서 ChatGPT3.5는 오류를 발견하지 못했고 ChatGPT4는 오류를 발견함으로써 평가의 질이 변별되었다.

표 8 오류 요약문 평가 설명 예시(일관성)

오류 요약문	한나라당 홍준표 민주당 원혜영 선진과 창조국의 모임 문국현 원내대표는 6일 임시 국회 종료 이틀을 남기고 푸른 바다에서 카누를 출항하는 우리의 모습을 발견했다 . 여야는 여야 간 이견이 없는 언론중재법 등 2건만 이번에 협의 처리하고 신문·방송 검열과 대기업 방송 진출 허용 등과 관련한 6개 쟁점 법안은 가능한 한 빨리 합의 처리하기로 했다. 이로써 지난해 12월 10일 개회한 후 4주 동안 야당의 본회의장 불법 점거와 국회의장의 질서유지권 발동 등으로 대치했던 정국이 정상화됐다.
ChatGPT3.5	요약문은 전반적으로 일관성을 유지하고 있습니다. 처음에 여야가 주요 쟁점법안 처리 문제를 일괄 타결했다는 내용을 언급하고, 그 후에 언론중재법 등 2건의 법안을 협의 처리하고 다른 6개 쟁점 법안에 대한 노력을 언급하며 요약이 이어지고 있습니다.
ChatGPT4	요약문의 내용이 일관적으로 이어지지 않으며, "푸른 바다에서 카누를 출항하는 우리의 모습을 발견했다."라는 내용이 원문과 관련이 없어 보입니다.

표 8의 정국 관련 기사에서 ChatGPT3.5는 “푸른 바다에서 카누를 출항하는 무리의 모습을 발견했다”라는 문장을 찾아내지 못해 내용의 일관성을 올바르게 평가하지 못했다. 하지만 ChatGPT4는 오류를 찾아 내용 일관성에 대해 평가가 가능함을 알 수 있었다.

표 9 오류 요약문 평가 설명 예시(문법성)

오류 요약문	10일 끝난 피겨스케이팅 전국선수권대회 주니어 남자 싱글에서 우승한 이동원은 종합 144.62점을 얻어 2위 이준형(117.56점)을 제쳤는데 이는 시니어 남자 싱글 우승자 김민석(132.55점)보다 높다. 김연아와 김나영을 가르쳤던 신혜숙 코치의 지도로 실력이 부쩍 는 이동원은 김연아가 여자 피겨의 인기를 살렸듯이 자신도 남자 피겨 붐을 일으키고 싶다고 밝혔다. 또래 선수들가 하지 못하는 트리플 점프 5개를 모두 소화해 내는 이동원은 한 아이스쇼에서 복싱 선수 복장으로 연기를 펼치고 시상식 직후엔 점프 세리머니를 하는 등 쇼맨십도 뛰어나다.
ChatGPT3.5	문법적 오류는 발견되지 않습니다. 문장 구조와 표현이 명확하며 자연스럽게 흘러갑니다.
ChatGPT4	"또래 선수들가 하지 못하는" 부분에서 "선수들가"가 "선수들이"로 수정되어야 합니다.

표 9의 ‘문법성’에 대한 조사 사용 오류를 ChatGPT4는 발견하여 평가에 반영하였고 ChatGPT3.5는 발견하지 못했다. 표 8의 ‘일관성’과 마찬가지로 ChatGPT3.5는 오류를 발견하지 못하나 ChatGPT4는 가능한 것으로 보인다. 제시되는 설명의 유형도 ChatGPT3.5와 ChatGPT4가 차이가 있다. ChatGPT3.5는 대다수의 설명 형식이 아래 표 10과 동일했다. 이는 오류를 발견하지 못해 같은 설명문을 반복한 것으로 보인다.

표 10 ChatGPT3.5 평가 설명 출력 유형 예시

<ul style="list-style-type: none"> ● 일관성: 요약문은 전반적으로 일관성을 유지하고 있으며, 원문의 주요 내용과 흐름을 잘 따라가고 있습니다. ● 관련성: 요약문은 원문의 내용과 관련이 높으며, 주요 발언과 사건에 대한 내용을 충실히 다루고 있습니다. ● 문법성: 문법적인 오류나 오타가 없으며, 문장 구조와 표현이 명확하고 자연스럽게 흘러갑니다. ● 유창성: 요약문은 읽는 과정에서 흐름이 끊기지 않고 이어지며, 대체로 이해하기 쉽습니다.
--

반면에 ChatGPT4는 설명에 표 11과 같이 다양한 평가 근거를 제시하고 있다.

표 11 ChatGPT4 평가 설명 출력 유형 예시

● 일관성: 요약문은 국제 금융시장의 불안감과 주요 금융사들의 경영실적 악화, 그리고 기업들의 경기침체

<p>여파를 다루는 내용으로 전체적으로 일관된 흐름을 보이고 있다.</p> <ul style="list-style-type: none"> ● 관련성: 요약문은 원문의 주요 내용들을 잘 반영하고 있다. 국제 금융시장의 불안감, 주요 금융사들의 경영실적 악화, 그리고 기업들의 경기침체 여파 등의 내용이 요약문에서도 명확하게 나타나 있다. ● 문법성: 요약문은 문법적으로 올바르게 작성되었다. 어미, 조사, 오타 등의 문제점을 찾기 어렵다. ● 유창성: 요약문은 유창하게 읽히며, 되돌아가서 읽어야 이해하는 부분이 없다. 한번에 내용을 이해할 수 있도록 직관적으로 작성되었다.

ChatGPT4가 오류를 찾을 수 있음에도 불구하고 인간 평가와 상관관계가 낮은 이유는 오류를 안정적으로 찾아내는 능력의 부족으로 보인다. ChatGPT3.5와 비교하여 상대적으로 오류를 잘 찾아내고 평가하지만 모든 오류를 찾아내지는 못한다.

표 12 ChatGPT4 평가 설명 오류 예시

오류 요약문	인천 강화군의 외딴 섬인 제주도 의 소녀 가장 장효선 양이 온갖 역경을 이겨 내고 고려대, 인하대, 명지대, 건국대 등 4곳에 합격해 화제다. 장 양은 기초생활수급자인 할머니를 돌보며 고1 쌍둥이 여동생들을 뒷바라지하는 소녀 가장으로, 지난 달에는 자신의 처지와 가족의 생활을 글로 써 시 교육청 주최 ‘효행실천사레발포대회’에서 최우수상을 받기도 했다. 장 양은 할머니와 동생들 뒷바라지 때문에 가까운 인하대 진학을 결정했지만 등록금과 수업료, 책값, 생활비 등을 마련할 길이 막막하기만 하다
ChatGPT4	요약문은 원문의 주요 내용을 정확하게 반영하고 있으며, 장효선 양의 학업과 가정 상황, 그리고 대학 진학 결정에 관련된 정보를 제공하고 있습니다. 장효선 양의 역경과 대학 합격 사실이 잘 나타나 있습니다.

표 12의 오류 요약문 ‘인천 강화군의 외딴 섬인 제주도’와 같이 잘못된 정보에 대해 평가해야 하는 ‘관련성’ 항목에 만점을 부여하고 있다. 이 외에도 오류를 찾아내지 못한 편이다. 표 7에서 ChatgPT3.5와 ChatgPT4의 상관관계는 높은 편인데 이는 ChatgPT3.5와 ChatgPT4가 인간 평가와 유사하지 않다는 것을 뒷받침한다. ChatGPT4가 ChatGPT3.5보다 향상된 모델이지만 한국어 중심으로 학습된 모델이 아니기 때문에 한국어 요약문에서 오류 발견이 어려운 편으로 생각된다. 앞의 논의와 같이 ChatGPT3.5는 대부분의 오류를 발견하지 못하고, ChatGPT4는 오류를 발견할 수는 있으나 불안정함이 원인으로 추측된다. 위의 오류 평가 결과 상관계수와 평가에 함께 제시된 설명을 고려하여 종합하면 ‘ChatGPT3.5와 ChatGPT4’ > ‘인간과 ChatGPT4’ > ‘인간과 ChatGPT3.5’ 순으로 평가가 유사하다고 볼 수 있고 ChatGPT3.5와 ChatGPT4 모두 오류를 발견하여 평가하는 성능이 높지

않다는 것을 알 수 있었다.

5. 결론

대화 시에 맥락을 고려하고 긴 문장 처리에 능한 ChatGPT가 인간만큼 요약문 평가가 가능한 성능을 가졌는지 연구하고자 ChatGPT를 활용하여 정문 요약문, 오류가 있는 요약문 평가를 진행하였다. ChatGPT가 생성한 요약문과 정답 요약문의 유사성 관계에 대해 정량적으로 BERTScore를 통해 평가한 결과, ChatGPT3.5와 ChatGPT4의 요약문에 대한 점수는 유사성을 가진 수준임을 확인하였다. 정성적 평가는 ‘일관성, 관련성, 문법성, 유창성’의 4개 평가 지표를 설정하여 인간에게 제공한 것과 동일한 프롬프트를 ChatGPT3.5와 ChatGPT4에게 제공하여 평가를 하도록 하였다. 실험 결과 요약문 평가에 대한 인간 평가 결과와의 상관관계에서 ChatGPT3.5와 ChatGPT4간의 성능 차이를 발견하였다. ChatGPT3.5는 인간 평가와 상관관계가 거의 없었고 ChatGPT4는 평가 항목에 따라 편차는 있지만 상대적으로 ChatGPT3.5보다 인간 평가와 상관관계가 있어서 인간을 보조하는 ChatGPT4의 활용 가능성을 확인할 수 있었다.

더불어 ChatGPT가 영어로 학습된 모델임을 고려하여 한국어 오류를 잘 찾아내어 올바르게 평가할 수 있는지, 적절한 점수를 부여하는지 확인하고자 한국어 오류 요약문에 대한 평가를 진행한 결과, 정문 요약문 평가와 달리 한계점을 발견하였다. 오류 요약문 인간 평가와의 상관관계에서 ChatGPT3.5의 평가 결과는 유의미한 관계를 확인하기 어려웠고, ChatGPT4 평가 결과는 인간 평가와 약한 상관관계가 있었다. 그 근거로 ChatGPT3.5는 오류를 거의 찾아내지 못해 제대로 평가를 하지 못하고 평가와 함께 제시된 설명도 반복되었지만 ChatGPT4는 요약문의 오류에 대한 내용을 반영한 설명을 제시하여 두 모델의 성능 차이가 있음을 확인할 수 있었다. 다만 ChatGPT4가 ChatGPT 3.5에 비해 오류를 잘 찾아내어 평가하는 것과 별개로 여전히 인간 평가보다 ChatGPT 3.5 평가와의 상관관계가 더 높아 이에 아직까지는 ChatGPT4를 활용하더라도 요약문의 오류를 안정적으로 평가하는 것에는 무리가 있어 보인다.

앞으로 생성 언어 모델의 요약 평가 능력을 키우기 위해서는 안정적으로 정성적 평가를 하는 모델도 필요하지만 평가를 잘 지시하는 것도 중요하다. 본 연구에서는 프롬프트를 zero-shot으로 제공했으나 few-shot으로 프롬프트를 제공하는 연구와, 언어에 따른 프롬프트 성능 차이 연구를 향후 진행하여 안정적인 성능의 가능성을 살펴보고자 한다. 또한 주관적인 평가로부터 발생하는 논쟁을 줄일 수 있는 명확한 평가 지표를 설정하는 연구와 오류가 있는 요약문의 성능 향상을 위해 모델에 미세 조정을 진행하는 방법으로 연구를 진행하는 것이 향후 과제이다.

참고문헌

[1] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2023). Benchmarking large language models for news

summarization. arXiv preprint arXiv:2301.13848.
 [2] Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). Exploring the limits of chatgpt for query or aspect-based text summarization. arXiv preprint arXiv:2302.08081.
 [3] Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). Human-like summarization evaluation with chatgpt. arXiv preprint arXiv:2304.02554.
 [4] Chiang, C. H., & Lee, H. Y. (2023). Can Large Language Models Be an Alternative to Human Evaluations?. arXiv preprint arXiv:2305.01937.
 [5] Wang, J., Liang, Y., Meng, F., Shi, H., Li, Z., Xu, J., ... & Zhou, J. (2023). Is chatgpt a good nlg evaluator? a preliminary study. arXiv preprint arXiv:2303.04048.
 [6] Chin-Yew Lin, "ROUGE: A Package for Automatic Evaluation of Summaries", In Proceedings of Association for Computational Linguistics 2021, 2021
 [7] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi, "ERTScore: Evaluating Text Generation with BERT". ICLR 2020. arXiv:1904.09675
 [8] Yulong Chen, Yang Liu, Liang Chen, Yue Zhang. "DialogSum: A real-life scenario dialogue summarization dataset". In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 5062-5074, 2021
 [9] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, Aleksander Wawer. "SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization". In Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 70-79, Hong Kong, China. Association for Computational Linguistics, 2019
 [10] 국립국어원, 국립국어원 문서 요약 말뭉치(버전 1.0), <https://corpus.korean.go.kr>, 2020
 [11] Koubaa, A. (2023). GPT-4 vs. GPT-3.5: A concise showdown.
 [12] Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, Dragomir Radev; SummEval: Re-evaluating Summarization Evaluation. Transactions of the Association for Computational Linguistics 2021; 9 391-409.
 [13] 김영주, 이선영, 이정민, 남주연, 이선희 and 최은정. (2012). 한국어 문법성 판단 테스트 개발 연구. 국어교육, 137, 353-380.
 [14] 양민철, 김민정, & 임해창. (2011). 영어 논술 자동 평가를 위한 언어 유창성 측정 방법. 한국정보과학회 언어공학연구회: 학술대회논문집(한글 및 한국어 정보처리), 25-29.