한국언론학보
2025년 69권 1호 5~38쪽
https://doi.org/10.20879/kjjcs.2025.69.1.001
http://www.comm.or.kr

대규모 언어 모델은 분석 도구가 될 수 있는가?
GPT를 활용한 내용 분석의 신뢰도와 타당도를 중심으로

이상혁 서울대학교 기초교육원 강의 전담 교수
김은미 서울대학교 언론정보학과 교수

Can LLMs Be Analytical Tools?
A Content Analysis Study Using GPT Focusing on Reliability and Validity

Sanghyuk Lee*
(Assistant Teaching Professor, Faculty of Liberal Education, Seoul National University)
Eunmee Kim**
(Professor, Department of Communication, Seoul National University)

This study explores the applicability of Large Language Models (LLMs) to computational text analysis methods (CTAM) in communication studies. Despite the increasing potential of LLMs as text analysis tools due to advancements in LLMs technology, addressing internal reliability and external validity remains crucial, as these are inherent limitations of LLMs when used as measurement tools. Unlike traditional coding methods or rule-based text categorization systems, LLMs do not always guarantee consistent or reproducible outputs. This raises a critical question regarding whether LLMs can function as appropriate text analysis tools in academic research.

To assess the potential of LLMs for CTAM, this study evaluates their text analysis outputs in terms of internal reliability and external validity. Specifically, it examines whether LLMs produce consistent results for the same prompts upon repeated analyses (internal reliability) and whether their outputs align with human coding results (external validity). For a large dataset of fact-checked news articles, human coders performed information extraction, and the same dataset was analyzed using a LLMs. A multi-step validation process was conducted, assessing internal reliability through repeated analyses and external validity through comparison with human coding results.

The findings indicate that text analysis using LLMs exhibits an acceptable level of internal reliability and external validity. Iterative analyses demonstrated that LLMs provide consistent analytical results, while comparisons with human coders confirmed sufficient level of external validity. However, reliability and validity significantly decreased depending on the genre of the news article and the type

* odhinn84@snu.ac.kr, corresponding author
** eunmee@snu.ac.kr

of information analyzed. The study found that LLMs exhibited lower reliability and validity when processing news articles relying heavily on numerical data or economic news, suggesting that the usability of LLMs in text analysis may vary depending on specific conditions.
Therefore, using LLMs as a measurement tool in research requires careful consideration of procedural frameworks and the nature of the data being analyzed. To ensure internal reliability, it is recommended to implement elaborate prompt engineering and repeated measurement for the same prompts. Additionally, external validity should be reinforced through comparisons with human coding results. Furthermore, LLMs need to be tested across various conditions and contexts to determine the specific circumstances under which they perform optimally.
The findings of this study, verified through these processes, are expected to lay the foundation for the utilization of LLMs as meaningful analytical tools in communication studies and to contribute to the advancement of text analysis methodologies. While LLMs have inherent limitations, these can be mitigated through the systematic procedures established in this study. The analytical procedure proposed here requires further discussion and refinement to develop into a standardized framework for future research.

Keywords: large language models, computational text analysis method, reliabilty, validity, computational method

국문초록

본 연구는 커뮤니케이션 연구 영역에서 대규모 언어 모델(LLMs)을 컴퓨터 기반 텍스트 분석 방법(CTAM)에 적용할 수 있는가의 여부를 탐색한다. LLM 기술 발전에 따른 텍스트 분석 도구로서의 가능성이 높아지고 있음에도 불구하고 측정 도구로서의 LLMs의 한계라고 할 수 있는 내적 신뢰도와 외적 타당도에 대한 문제를 해결할 필요가 있다. 기존의 전통적인 코딩 방식이나 규칙 기반 텍스트 분류 시스템과 달리 LLMs의 경우 항상 일관되거나 재현 가능한 출력이 보장되지 않기 때문이다. 이는 LLMs이 학술 연구에서 적절한 텍스트 분석 도구로 기능할 수 있는지에 대한 커다란 의문을 제기한다.
본 연구에서는 LLMs이 CTAM에 적용될 수 있는 그 가능성을 평가하기 위해 내적 신뢰도와 외적 타당도의 관점에서 LLMs의 텍스트 분석 결과물을 평가한다. 특히 LLMs의 결과물이 반복된 동일 프롬프트에 대해 일관된 결과를 출력하는지(내적 신뢰도), 인간 코딩 결과와 비교하여 동일한 결과를 출력하는지(외적 타당도)를 평가한다. 다수의 팩트체크 뉴스 기사에 대해 인간 코더들이 정보 추출(information extraction)을 수행한 데이터를 동일하게 LLMs으로 분석하게 하였다. 이를 통해 반복 분석을 통한 내적 신뢰도 평가, 인간 코더의 결과와의 비교를 통한 외적 타당도 평가라는 다단계 검증 절차를 수행했다.
연구 결과, LLMs을 통한 텍스트 분석의 결과물은 신뢰할만한 수준의 내적 신뢰도와 외적 타당도를 확보하고 있음을 보여주고 있다. 반복 분석을 통해 LLMs이 일정한 분석 결과를 제공할 수 있으며, 인간 코더와의 비교를 통해 충분한 외적 타당도를 확보할 수 있음을 보였다. 그러나 뉴스 기사의 장르나 정보의 종류에 따라 신뢰도와 타당도가 크게 하락하는 현상이 확인되었다. 수치 데이터를 주로 사용하거나, 경제 뉴

스의 경우 신뢰도와 타당도가 낮게 나타났으며, 이는 LLMs의 텍스트 분석 결과가 조건에 따라 사용 가능성이 달라질 수 있음을 의미한다.
따라서 LLMs을 연구의 측정 도구로 사용하기 위해서는 절차와 분석 대상에 대해 주의가 필요하다. 정교한 프롬프트의 제작, 동일 프롬프트의 반복 측정을 통한 내적 신뢰도 확인, 그리고 인간 코딩과의 비교를 통한 외적 타당도 확보라는 절차가 제안된다. 또한 다양한 조건과 맥락에 대한 테스트를 통해 LLMs이 어떤 조건에서 어느 정도 수준의 성능이 나타나는지 검증할 필요가 있다.
이 과정을 통해 확인된 본 연구의 결과는 LLMs이 커뮤니케이션 연구 분야에서 의미 있는 분석 도구로서 활용될 수 있는 기반을 마련하고 미래의 텍스트 분석 방법론에 기여할 것으로 기대한다. LLMs의 한계점을 연구 수행 절차를 통해 극복할 가능성을 확인할 수 있었다. 본 연구에서 제안된 분석 절차의 경우 앞으로 지속적인 논의를 통해 표준화된 절차로 발전해야 할 필요가 있다.

핵심어 : 대규모 언어 모델(LLMs), 컴퓨터 기반 텍스트 분석 방법(CTAM), 신뢰도, 타당도, 컴퓨터 기반 방법론

1. 서론

지난 20년 동안 컴퓨터를 활용한 텍스트 분석(Computational Text Analysis Methods, 〔CTAM〕)은 지속적으로 발전하여 현재는 커뮤니케이션 연구 분야에서도 대중적인 방법론으로 자리잡고 있다(Birkenmaier et al., 2024). 뉴스, 채팅, 온라인 게시글 등 다양한 텍스트를 분석 대상으로 하는 커뮤니케이션 연구 분야에서 텍스트는 중요한 데이터로 활용될 수 있는데, CTAM는 텍스트를 수량화하여 평가하고 분류하는 등의 작업에서 큰 역할을 하고 있다.

CTAM의 핵심적인 요소는 물론 텍스트를 분석해서 결과를 반환하는 언어 모델이다. BERT(Bidirectional Encoder Representations from Transformers)와 같은 언어 모델은 텍스트를 입력받아 내부에서 이를 인식, 분석하고 이에 따라 특정한 결과를 내어놓는다. 이러한 언어 모델을 구축하는 과정에서 딥 러닝 등의 고급 컴퓨팅 기술이 활용되며, 사전에 학습된 언어 모델은 용도에 따라 미세 조정(fine-tuning) 과정을 거쳐 실제 텍스트 분석에 활용된다. 컴퓨터 공학이나 전산언어학 등의 연구 분야에서는 더 효율적이고 정확한 언어 모델을 개발하는 것을 목표로 다양한 연구를 진행하고 있다.

커뮤니케이션 분야에서는 언어 모델 자체의 성능보다는 이렇게 제작된 언어 모델을 활용하여 실제 텍스트를 분석하여 현상을 확인하는 다양한 연구를 진행 중이다. 뉴스 내용 분석에서부터 댓글의 속성을 분류하는데 이르기까지 다양한 텍스트를 대상으로 연구 분야에서 언어 모델을 텍스트 분석에 활용하고 있다. 그러나 커뮤니케이션 분야에서는 언어 모델에 대해 단순히 이용만 하는 것이 아니라 언어 모델을 활용하는 방법 자체에 대한 논의도 이루어져왔다(Birkenmaier et al., 2024; Song et al., 2020; Van Atteveldt et al., 2021). 특히 언어 모델이 생성해낸 결과를 검증하는 방법에 대한 논의가 핵심적이다. 언어 모델을 활용하여 대규모 텍스트 분석을 수행하는 연구는 일반적으로 언어 모델이 생성한 결과를 인간이 수행한 분석 결과와 비교하는 방식으로 그 타당성을 평가한다. 즉, 인간이 평가한 결과를 정답(gold standard)으로 간주하고 언어 모델이 생성한 결과가 인간이 만들어낸 결과와 얼마나 일치하는지를 확인하는 것으로 외적 타당도를 확보하는 것이다. 이를 위해 인간 코더(human coder)를 확보하여 전체 데이터를 일부분을 코딩하는 작업이 선행되어야 하는데, 커뮤니케이션 분야에서는 이 부분에 주로 초점을 맞추어 적절한 연구 방법을 탐색하고 있다. 어떤 코더가 적절한지, 몇 명의 코더를 사용하는 것이 가장 효율적인지, 전체 데이터에서 몇 %의 샘플을 추출하여 인간 코딩을 수행하야 하는지, 이 모든 것을 포함한 적절한 연구 절차를 제시하는 등의 논의가 진행되고 있다. 이는 언어 모델 그 자체를 다루는 법 보다는 언어 모델을 활용하는 방법에 대한 논의이며, 언어 모델이 만들어낸 결

과를 신뢰하기 위한 외부 타당성을 확보하는 방법에 대한 문제로 귀결된다(Birkenmaier et al., 2024).

기술의 발전으로 인해 다양한 언어 모델이 빠르게 발전하고 등장하면서 최근 생성형 AI라고 불리는 GPT를 위시한 여러 대규모 언어 모델(Large Language Models, 〔LLMs〕)가 나타나면서 사회에 큰 충격을 주고 있다. 기존의 제한적인 기능을 가진 언어 모델에 비해 LLMs는 자연어 처리 및 생성에 있어 압도적인 성능을 보여주고 있다. LLMs는 기존의 BERT 기반 모델에 비해 긴 텍스트 처리와 문서 수준의 맥락을 고려한 분석, 그리고 특정 전문 영역 특화 작업에서 뛰어난 성능을 보이는 것으로 평가되고 있다. 특히 GPT와 같은 LLMs는 BERT가 주로 문장 수준의 문맥 분석에 한정되는 반면, 긴 텍스트나 문서 전체를 포괄적으로 이해하고 분석할 수 있는 능력을 보여준다(Kim et al., 2023; Miah et al., 2024). GPT-3.5를 활용하여 금융 분석 보고서와 같은 전문적인 영역의 보고서를 분석한 연구에서도 높은 성능을 보였으며, 투자 전략 수립에서 BERT 기반 모델보다 높은 연간 수익률을 기록했다(Kim et al., 2023). 또한, GPT는 수천억 개에 달하는 매개변수로 훈련되어 BERT보다 광범위한 데이터에 대한 일반화 성능이 뛰어나 문서 수준의 텍스트를 분석함에 있어서도 맥락을 충분히 고려할 수 있어 높은 성능을 보여주고 있다(Javaji et al., 2024; Miah et al., 2024). 이러한 특성은 LLMs가 복잡하고 다면적인 텍스트를 분석하는데 있어 기존의 컴퓨터 기반 텍스트 분석 도구보다 더 효율적임을 보여준다.

이로 인해 CTAM을 활용하여 진행되고 있던 영역에서도 LLMs를 기존의 언어 모델 대신 사용할 수 있는지에 대한 검토와 시도가 각종 연구 분야에서 이어지고 있다(Lee et al., 2024; Parker et al., 2024; Pelaez et al., 2024; Tai et al., 2024). 그러나 이런 연구들의 경우 LLMs의 연구 도구로서의 측면을 면밀히 검토한 뒤 연구에 적용하지 않았다는 한계가 있다. 겉으로 드러난 LLMs의 텍스트 분석 능력이라는 기술적 특성에만 주목하여 LLMs가 연구 도구로 어떤 특성을 가지고 있는지에 대한 검증 작업이 진행되지 않은 상태로 실제 분석에 투입한 것이다. 새로운 도구를 도입해서 다량의 데이터 분석을 시도하는 것은 바람직하지만 이것이 도구로서 가진 타당성이나 한계점, 상대적 장단점을 검증하는 연구가 병행되지 않은 채 내용분석에 LLMs를 그대로 적용할 경우 도구만능주의의 오류로 인해 이를 통해 밝혀진 연구 결과의 신뢰성을 담보할 수 없다. 특히 BERT와 같은 기존의 언어 모델과 새로 등장한 LLMs는 기능적으로 동일한 부분과 그렇지 않은 부분이 있는 바, 기존의 언어 모델을 활용한 CTAM 과정에 기존 언어 모델의 위치를 LLMs로 교체하여 연구에 도입해도 되는지에 대한 검토가 선행될 필요가 있다.

기존의 CTAM 방식과 달리 LLMs를 언어 모델로서 CTAM에 활용할 수 있는지 확인하기

위해 특히 핵심적으로 살펴보아야 할 부분은 내적 신뢰도(intra-coder reliability)라고 볼 수 있다. LLMs의 경우 프롬프트(prompt)라고 불리는 지시 사항을 자연어로 입력하고, 이 자연어의 내용을 분석한 언어 모델이 답변을 자연어로 출력하는 방식으로 작동한다. 이 과정은 BERT와 같은 기존의 CTAM에 사용된 언어 모델과 유사한데, 문제는 이 과정에서 LLMs가 동일한 입력에 대해 다른 결과를 출력하는 경향이 있다(Deng et al., 2023)는 것이다. 이는 기존의 CTAM에 사용되어왔던 다른 언어 모델과 달리 LLMs가 분석 도구로서의 신뢰도가 낮다는 것으로, LLMs를 CTAM의 일종으로 간주할 수 있는지의 여부를 결정하는 핵심적인 질문이라고 할 수 있다. 한 마디로, 일정한 결과가 출력될 것이라는 신뢰가 확보되지 않고서는 LLMs는 학술 연구의 도구로 사용되기 어렵다. 또한 과연 LLMs가 만들어낸 결과물이 목적에 적합한지에 대한 외적 타당도 역시 검증해야 할 대상이다. 이는 언어 모델을 활용한 텍스트 분석 과정에 일반적으로 적용되는 것으로, 언어 모델이 생성해낸 결과물이 과연 적절한가를 평가하기 위해 인간 평가 결과와 비교하는 과정을 거치게 된다. LLMs 역시 동일한 외적 타당도 평가 과정을 거칠 필요가 있다.

따라서 본 연구에서는 LLMs의 구조적 특성에도 불구하고 텍스트 분석 도구로서 활용이 가능한지를 검토할 것이다. 특히 LLMs가 생산해내는 텍스트 분석 결과물의 일관성, 즉 내적 신뢰도가 확보되는지에 주목한다. 또한 일반적으로 CTAM의 효과를 측정하는 방법인 외적 타당도, 즉 정답률이 어떻게 나타나는가 역시 함께 확인할 것이다. 이를 위해 본 연구에서는 인간 코딩을 통해 이루어진 연구의 데이터를 표준(gold standard)으로 삼아 LLMs를 통한 분석 결과가 얼마나 일관적인지, 그리고 얼마나 인간 코딩과 일치하는지 확인할 것이다. 마지막으로 LLMs가 생산한 분석 결과가 각 주제 분야에 따라 인간 코딩과의 일치도가 어떻게 달라지는지 역시 확인할 것이다. 본 연구를 통해 사회과학 분야에서 생성형 AI, 즉 LLMs를 텍스트 분석에 활용할 수 있는지의 여부를 확인할 수 있다. 이후 지속적으로 LLMs의 활용 가능성에 대한 테스트가 이어져야 하겠지만 그 시작점으로서의 의의가 있다.

본 연구의 문헌 연구는 다음과 같이 구성되어 있다. 우선 커뮤니케이션 분야에서 사용되고 있는 기존의 CTAM 연구 방법을 살펴보면서 LLMs를 어떤 방식으로 적용할 수 있을지에 대해 검토하고, 다음으로 LLMs를 텍스트 분석에 활용하려는 몇 가지 시도들과 이 시도들이 놓치고 있는 LLMs의 분석 도구로서의 한계점을 지적한다. 마지막으로 LLMs의 내적 일관성 문제와 분석 대상에 따른 결과의 차이가 나타날 수 있다는 점에 대해 논의한다.

2. 문헌 연구

1) 커뮤니케이션 연구 분야에서의 CTAM 활용

커뮤니케이션 연구 분야에서는 CTAM을 활용한 다양한 연구들을 진행해왔다. 언론 보도 성격 분석(DeButts & Pan, 2024), 소셜 미디어 내용 분석(Rains et al., 2023), 뉴스 기사 주제 분석(Jürgens & Stark, 2022), 스트리밍 채팅의 내용 분석(Asbury-Kimmel et al., 2021) 등과 같이 커뮤니케이션 과정에서 발생한 텍스트를 CTAM을 통해 분석하고 가설을 검증하는 방식으로 이루어진다.

언론 보도 성격 분석(DeButts & Pan, 2024)의 경우 중국에서 특정 언론사 소속의 외국 기자를 추방한 시점 전후로 해당 언론사들의 중국 관련 기사가 어떻게 변화했는지에 대한 텍스트 분석을 실시했다. 이 연구에서는 기사들에 대한 감정 분석을 수행하여 기사에 담긴 문장들을 긍정/부정/중립으로 분류하고 기사 전체 문장들의 평균 감정 점수를 계산하여 기사의 전반적인 감정 톤을 결정하였다. 그리고 외국 기사 추방 전후로 중국에 대한 기사의 감정 톤에 대한 변화를 확인하였다. 소셜 미디어 내용 분석(Rains et al., 2023)의 경우 미세 조정을 통해 텍스트에 욕설이 포함되었는지의 여부를 판단하는 모델을 구축하였고, 이를 기반으로 정치 트윗의 욕설 여부 등을 분석하여 러시아의 인터넷 기관(Russian government's Internet Research Agency, IRA)이 미 대선 기간 동안 소셜 미디어를 활용하여 정치 과정에 어떻게 영향을 미쳤는지 확인하였다. 뉴스 기사 주제 분석(Jürgens & Stark, 2022)에서는 뉴스 플랫폼의 알고리즘 큐레이션이 뉴스 노출의 다양성에 미치는 영향을 분석하고 단기적인 긍정적 효과와 장기적인 부정적 효과를 분리하여, 플랫폼별 영향을 평가하였다. 이 과정에서 일어 기반의 BERT 모형을 미세 조정하여 뉴스 기사의 주제를 분류하는데 CTAM을 사용하였다. 스트리밍 채팅의 내용 분석(Asbury-Kimmel et al., 2021)의 경우 정치 토론을 중계하는 스트리밍에서의 채팅의 속성이 시청자의 인식과 태도에 어떻게 영향을 미치는지 분석하였다. 이 연구에서는 정치 토론이 진행되는 과정에서 나타나는 댓글의 독성(toxicity) 정도를 CTAM을 통해 분석하여 독성이 높은 댓글에 노출되었을 경우 정치 토론에 대한 시청자의 인식에 대해 측정했다.

이렇게 커뮤니케이션 분야에서 CTAM을 적용하여 텍스트를 분석하고 이를 기반으로 가설을 검증하는 방식의 연구가 다수 등장하고 있지만, CTAM을 적용하는 과정에서 연구자가 어떤 절차를 거쳐야 하는지에 대한 심도 깊은 논의도 찾아볼 수 있다. 송 등(Song et al., 2020)은 CTAM을 사용하는 과정에서 표준(gold standard)으로 사용되고 있는 인간 생성 데이터에 초점을 맞추고 있다. CTAM을 진행하는 과정에서 인간이 생성한 데이터는 주로 분류 성능을 검증,

즉 기계가 생성한 데이터가 얼마나 정답인지 평가를 진행하는 과정에서 사용된다. 그런데 이 검증용 인간 데이터의 구축 과정에 대한 논의는 상대적으로 부족했으며, 심지어 정치 커뮤니케이션 분야의 73개의 연구를 리뷰한 결과 약 58%(n=42)만이 인간 데이터를 활용한 검증 방식에 대해 구체적으로 보고한 것으로 나타났다(Song et al., 2020). 이들은 적절한 표준(gold standard)으로 사용하기 위한 인간 코딩 데이터를 생성하려면 어떤 조건이 충족되어야 하는지를 중점적으로 탐색했고, 몬테 카를로 시뮬레이션(Monte-Carlo simulation)을 통해 0.7 이상의 코더 간 신뢰도, 전체 데이터의 5% 이상, 그리고 무작위로 추출된 인간 코딩 데이터를 구축하기를 추천한다. 특히 데이터의 편향성의 경우 단독 요소로 평가하면 편향된 자료가 더 높은 성능을 보였지만 코더 간 신뢰도 등 다른 요소들을 고려할 경우 무작위로 추출된 데이터가 더 효율적인 것으로 확인되었다. 구체적으로, 허용 가능한 신뢰도(0.7 이상)를 전제로 수동 코딩된 검증 데이터 세트의 크기를 가능한 한 크게, 가급적 N = 1,300개 이상(즉, 조사 대상 전체 데이터의 1% 이상)으로 늘리기 위해 노력할 것을 권장하고 있다. 또한 코더 간 신뢰도는 검증 데이터의 비율과의 상쇄 효과가 나타났으며, 이에 따라 만약 코더 간 신뢰도가 0.7보다 낮을 경우 검증 데이터의 크기를 전체 데이터의 5% 이상으로 증가시키는 방법으로 적절한 성능을 유지하는데 도움이 된다고 주장한다. 이러한 연구 결과는 결국 코더간의 내적 신뢰도의 수준이 미치는 영향이 기계적인 텍스트 분석의 타당성을 검증하는데 있어 핵심적이라는 점을 말하고 있다.

CTAM을 수행하는 방법에 대한 또 다른 논의는 반 아테벨트 등(Van Atteveldt et al., 2021)의 연구에서 찾아볼 수 있다. 이들은 감정 분석을 기반으로 CTAM을 수행하는 절차에 대해 논의했다. 우선 수집된 뉴스 데이터를 기반으로 감정 분석을 실시하는데 이 결과를 숙련된 코더가 수동 코딩한 데이터, 온라인으로 코더를 모집하여 크라우드 코딩된 데이터, 머신 러닝 기법을 활용하여 기계적으로 코딩된 데이터, 사전 기반으로 기계적으로 코딩된 데이터로 나누어 그 결과를 비교하였다. 그 결과, 숙련된 코더가 코딩한 데이터가 가장 정확도가 높았고, 크라우드 코딩된 데이터 역시 예상과는 다르게 높은 수준의 정확도를 보여 연구에 사용할 수 있는 수준의 방법론이라는 것을 확인하였다. 반면 머신 러닝 기법이나 사전 기반의 컴퓨터를 활용한 자동 코딩의 경우 모두 수준 이하의 성능을 보였다. 저자들은 단순히 각 방법들의 결과를 비교한 것으로 연구를 마무리 하지 않고 이 연구 결과를 통해 CTAM을 사용하는 과정에서의 몇 가지 제안을 하고 있다. 먼저 크라우드 코딩의 경우, 2명의 코더로 먼저 코딩을 진행한 뒤, 두 코더가 동의하지 않는 케이스에 대해서만 세 번째 코더를 사용할 것, 그리고 코더가 늘어나더라도 전반적인 성능이 크게 향상되지는 않지만 코더 간 일치도가 높을 경우에는 적절한 선택이 될 수 있다는 것이다. 또한 CTAM의 전반적인 절차에 대해서도 표준(gold standard) 제작 과정에서 2명 이상의

코더를 사용하고 100~300개의 검증 데이터를 사용할 것, 기계 학습을 사용할 경우 1000개 이상의 문서를 코딩할 것, 진행하는 과정에서 지속적으로 유효성을 검사할 것 등의 제안을 남겼다. 비록 이 연구에서 코더의 숫자나 필요 신뢰도, 요구 샘플 데이터의 숫자 등은 내용 분석에 대한 크리펜도르프(Krippendorf, 2012)의 논의를 기반으로 하고 있지만, 그럼에도 불구하고 커뮤니케이션 분야의 관점에서 CTAM의 전반적인 절차를 제시했다는 점에서 의의가 있다.

정리하자면, 커뮤니케이션 분야에서는 주로 CTAM에 사용하는 기술 자체에 대한 논의보다는 해당 기술을 통해 나온 결과를 어떻게 검증하고 사용할 것인가에 대해 주목하고 적절한 절차를 제안하는 것으로 목표로 논의가 진행되었다. 일부 언어 모델을 변형시키는 연구가 있었지만, 주로 인간이 코딩한 학습용 데이터를 기반으로 모델에 추가 학습을 진행하거나 미세 조정(fine-tuning)을 수행하는 방식으로 이루어졌으며 기술적으로 모델 자체를 개발하거나 성능을 개선하는 방식의 연구는 커뮤니케이션 분야의 주된 연구 대상으로는 논의되지 않고 있다. 이는 언어 모델의 개발과 개선이 커뮤니케이션 분야의 연구 대상이 아니기 때문이기도 하지만, 동시에 커뮤니케이션 분야에서는 언어 모델을 통해 도출되는 결과물과 이를 활용한 가설 검증에 주목하고 있음을 보여주는 결과라고 볼 수 있다. 따라서 커뮤니케이션 분야에서는 언어 모델 그 자체와 언어 모델이 내부에서 텍스트를 처리하는 절차는 일종의 블랙박스로 간주하고, 그 결과를 어떻게 평가하고 검증할 것인지에 주로 초점을 맞추고 있다고 해석할 수 있다.

커뮤니케이션 분야의 기존 CTAM 관련 연구들은 인간 코딩 결과(gold standard)를 활용하여 결과를 검증하고, 또한 그 검증 과정이 어떠해야 하는지에 대한 논의가 핵심이다. 또한 이는 주로 기계적 텍스트 분석 결과의 타당성을 검증하는데 초점을 맞추고 있다고 볼 수 있다. BERT와 같은 언어 모델의 출력 결과를 그대로 받아들이기에는 아직 텍스트 분석 성능이 충분하지 않으며 이로 인해 분석 결과를 보정하기 위한 미세 조정(fine-tuning), 그리고 인간 코딩을 통한 결과물의 검증 등을 통해 언어 모델의 성능을 적절한 수준으로 확보하기 위해 노력하고 있는 것이다. 이를 통해 CTAM에서는 타당성이라는, 분석 도구의 핵심적인 조건 중 하나는 검증 과정을 구축하고 있다고 볼 수 있다.

문제는 기존 CTAM에서는 기계적 텍스트 분석 도구의 신뢰성을 확인하는 절차는 보여주지 않고 있다. 이는 오히려 기계적 작동이라는 측면에서 기존의 언어 모델이 안정성, 즉 신뢰도를 보여주고 있기 때문이다. BERT와 같은 기존의 언어 모델은 내부의 데이터에 기반하여 특정 텍스트가 입력될 경우 특정한 답을 일정하게 내놓으며, 이 과정은 기계적이기 때문에 높은 신뢰성을 담보할 수 밖에 없다. 즉, 동일한 입력에 무조건 동일한 결과를 생성하는 것이다. 이로 인해 CTAM에서는 기계적 텍스트 분석 도구의 결과물의 타당성은 검증할 필요가 있었지만 신뢰성은

검증할 필요는 없었다. 만약 LLMs를 적용한 텍스트 분석을 수행한다면 그 결과의 타당성 검증은 기존의 CTAM에서 사용된 방법을 그대로 적용할 수 있을 것으로 보인다. 반면 기존 CTAM 방식에서는 필요하지 않았던 언어 모델의 신뢰성을 검증할 필요가 추가적으로 발생한다.

2) LLMs를 텍스트 분석 연구에 접목하려는 시도들과 한계

강력한 텍스트 분석 능력을 가진 것으로 알려진 LLMs를 실제 텍스트 분석 연구에 적용하고자 하는 시도는 여러 영역에서 다양하게 이루어져 왔다. 가장 대표적인 방법은 바로 프롬프트를 정교하게 작성하고자 하는 시도였다. 파커 등(Parker et al., 2024)은 교육적 서베이의 결과를 LLMs를 통해 분석하고자 하는 시도에서 순환적인 프롬프트 사용과 이를 기반으로 한 정보 분류를 수행했다. 이들은 프롬프트를 입력할 때 단순히 지시사항을 입력한 것이 아니라, LLMs가 수행하여야 할 절차를 단계별로 구체적으로 지시하여 LLMs가 프롬프트를 기반으로 내적으로 추론(thought)을 수행할 때 지시받은 순서대로 단계적으로 진행(chain)하게끔 하는 방식(chain of thought, 〔CoT〕)을 통해 프롬프트의 정확성을 극대화하였다. CoT는 복잡한 추론 과제를 풀 때 인간의 사고 과정을 참고하여, 문제를 중간 단계로 분해하고 각각을 해결하는 과정을 거쳐 최종적인 답안을 도출하는 절차를 프롬프트에 도입한 방법이다(Wei et al., 2022). CoT는 프롬프트를 입력하는 과정에서 LLMs가 수행하여야 하는 추론의 절차를 구체적으로 입력한다. LLMs는 실제 이용자가 입력한 문제에 대한 답변을 생성하는 과정에서 프롬프트에 입력된 절차를 참고하기 때문에 단순 프롬프트보다 더 성능이 뛰어난다는 것이 저자들의 주장이다.

또 다른 시도는 LLMs가 생성한 결과물을 그대로 인간이 생성한 표준(gold standard) 데이터와 비교하여 정답률을 확인하는 방식이다. 이 방법은 기존의 CTAM 방법론의 연구 절차와 같이 언어 모델의 결과물을 대상으로 외부 타당성을 검증하던 방식으로, LLMs의 결과물 생성 과정이 무엇이든 그 결과물이 정확하다면 충분하다는 관점이라고 볼 수 있다. LLMs를 활용한 최근의 몇 가지 연구들(Lee et al., 2024; Suter & Meckel, 2024)는 LLMs가 생성해 낸 결과물을 인간이 동일하게 작업한 결과물과 비교함으로써 LLMs가 각 연구 영역에서 사용될 수 있는 가능성을 확인했다. 수터와 멕켈(Suter & Meckel, 2024)은 GPT를 이용하여 미국과 독일의 신문에서 AI 관련 헤드라인과 리드를 수집하고 이를 기반으로 코딩 지침을 작성하였고, 이후 이 코딩 지침을 바탕으로 GPT가 뉴스 기사를 주제별로 분류하고 감정을 평가하게 하였다. 전체 데이터 중 10%를 추출하여 두 명의 인간 코더로 하여금 GPT와 동일하게 코딩하게 한 뒤 그 결과를 GPT가 생성한 주제 분류와 감정 평가 결과와 비교하여 일치도를 확인했다. 또 다른 연구(Lee et al., 2024) 역시 의료 맥락에서 텍스트 분석을 수행하는 과정에서 LLMs를 사용하였

다. 이들은 인터뷰 기록을 기반으로 텍스트에서 주제를 파악하는데 LLMs를 도입하였으며, 기계적으로 생성된 결과를 인간 코더가 생성한 결과와 비교하여 그 정확도를 확인하였다.

마지막으로 LLMs에게 동일한 결과물을 다수 생성하게 한 뒤, 그 결과의 빈도를 확인하는 방식을 고려해볼 수 있다. 이는 LLMs의 내적 일관성을 해결하기 위한 직접적인 방법 중 하나로 볼 수 있다. 기본적으로 LLMs가 항상 동일한 결과를 반환하지 않는다는 것을 가정한 상태에서, 동일한 분석을 여러 번 수행하게끔 함으로써 내적 일관성을 검증하고 유지하고자 하는 방법이다. 타이 등(Tai et al., 2024)은 LLMs를 활용하여 텍스트 분석을 수행했는데, 3개의 텍스트에 대해 5개의 속성(Autonomy, Persistence, Perception of Identity, Novelty, STEM Interests)을 정의하고 해당 속성이 샘플 텍스트에 존재하는지를 분석하게 하였다. 이 과정에서 LLMs가 해당 텍스트를 총 160번 반복하여 분석하게 한 뒤 이 중 몇 %가 5개의 속성이 있다고 응답했는지를 확인하였다.

Table 1. The Result of LLM's Presence Analysis of the Five Attributes (Tai et al., 2024, p. 6)

Iterations	Autonomy	Persistence	Perception of Identity	Novelty	STEM Interests
5	0.200	0.600	0.000	0.000	0.000
10	0.300	0.600	0.000	0.000	0.100
20	0.350	0.550	0.050	0.000	0.100
40	0.325	0.625	0.075	0.000	0.100
60	0.328	0.534	0.086	0.000	0.103
80	0.275	0.575	0.075	0.013	0.138
100	0.300	0.600	0.070	0.010	0.120
120	0.308	0.625	0.075	0.008	0.133
140	0.316	0.632	0.075	0.015	0.128
160	0.325	0.594	0.088	0.013	0.131

Note. Iterations(반복 분석 회수), Autonomy(자기 주도성), Persistence(지속성), Perception of Identity(자기 정체성 인식), Novelty(창의성), STEM Interests(STEM 분야에 대한 관심)

〈Table 1〉에서 확인할 수 있듯이, 첫 번째 샘플 텍스트에 대해 총 160번 반복(iterations) 분석을 수행한 결과 자기 주도성(autonomy)에 대한 내용이 존재한다고 응답한 비율이 32.5%, 약 52회로 나타났다. 저자들은 이렇게 다수의 분석을 반복하여 분석 결과가 다양하게 나오더라도 결과의 비율을 확인하여 분석에 활용하였다. 다만 이 연구에서는 일관되지 않은 결과가 나왔을 때 이를 오류나 프롬프트의 한계로 판단하지 않고 실제 분석 결과로서 받아들

였다. 이들은 총 3명의 인간 코더 중 몇 %가 해당 요소가 텍스트에 존재한다고 응답했는지의 비율을 비교하는 방식으로 분석을 수행했다. 예를 들어 첫 번째 텍스트의 자기 주도성의 경우 LLMs는 전체 응답 중 약 30%가 존재한다고 응답했는데, 인간 코더 3명 중 1명이 해당 텍스트에 자기 주도성이 있다고 응답했다. 이 비율이 약 1/3이기 때문에 LLMs의 분석이 적절했다고 해당 연구에서는 판단하고 있다.

문제는 이런 연구들의 경우 LLMs가 가지는 분석, 측정 도구로서의 근본적인 한계를 고려하지 않고 있다는 것이다. 특히 신뢰도(reliability)와 타당도(validity)라는 두 가지 기준은 연구에서 사용되는 측정 도구로서는 필수적으로 갖춰야 할 조건이다. 그러나 BERT와 같은 기존의 CTAM에 사용되던 언어 모델과 달리 ChatGPT와 같은 최근의 LLMs는 그대로 CTAM에 적용될 수 있는지의 여부가 아직 명확하지 않다. 이는 현재 LLMs에서 처리되고 반환되는 결과물이 기존의 언어 모델과의 차이가 있기 때문인데, 가장 핵심적인 문제점은 분석 후 결과가 일정하게 나타나지 않는다는 점을 들 수 있다(Deng et al., 2023; Li et al., 2024). 연구에 사용되는 분석 도구로서의 신뢰도라는 측면에서 동일한 텍스트가 입력되었을 때 결과가 일정하게 나오지 않는다면 이는 신뢰할 수 있는 분석 도구가 될 수 없다. 현재 일반적으로 사용되는 인간 코더 역시 이러한 문제에서 자유로울 수 없으며, 코더 간 신뢰도(inter-coder reliability)뿐만 아니라 코더 내 신뢰도(intra-coder reliability)를 확보하기 위해 교육 등의 다양한 노력을 하고 있지만 그럼에도 불구하고 코더 내 신뢰도 확보에 대한 문제제기는 계속되고 있다(Lamprianou, 2023). 특히 언어 모델 자체를 개발하던 초기 컴퓨터 공학 영역에서 인간이 생성한 결과를 정답으로 간주했던 것에 반해, 사회과학 영역에서는 인간 코딩 결과에 대한 불신을 가지고 있었다(DiMaggio, 2015). 인간이라면 당연히 내적 일관성을 가지고 대상을 평가할 것이라는 가정을 통해 인간 코더를 사용하여 텍스트와 같은 비정형 데이터를 측정하고 있지만 이에 대한 의문과 검증의 필요성이 지속적으로 제기되고 있는 것이다. 일관성을 유지해야 한다는 명제를 가지고 코딩을 진행하는 인간 코더 역시 일관성의 문제를 가질 수 있는데, LLMs의 출력 결과가 일관적이지 않다는 문제가 제기되었으며 이는 측정 도구로서 LLMs는 일종의 코더 내 신뢰도의 문제가 있다는 것으로 볼 수 있다. 특히 내용 분석 과정에서 코더에게 요구되는 능력이 일정한 기준을 꾸준히 유지하고(reliability), 주어진 기준에 맞춰 평가(validation)해야 한다(Su et al., 2016)는 면에서 내적 일관성의 부족은 측정 도구로서의 치명적인 문제라고 볼 수 있다.

정리하자면, LLMs의 강력한 텍스트 분석 및 생성 기능으로 인해 텍스트 분석이라는 연구 분야에서 LLMs를 사용하고자 하는 시도가 지속되고 있다. 그럼에도 LLMs는 아직 안정적인 연

구 분석 도구로서의 지위를 확보하기 있지 못하고 있는데 가장 큰 이유는 LLMs의 내적 일관성, 즉 신뢰도의 문제가 있기 때문이라고 볼 수 있다. 물론 그 외에도 기계적인 텍스트 분석이라는 측면에서 타당성을 검증해야 하는 문제도 남아있다고 할 수 있다. 따라서 본 연구에서는 이런 문제를 해소할 수 있는 방법을 탐색하기 위해 우선 기존의 커뮤니케이션 연구 분야에서 활용되어 왔던 CTAM에 대해 살펴볼 것이다.

3) LLMs의 신뢰성 문제의 원인과 텍스트 종류에 따른 성능 차이

LLMs의 내적 일관성, 즉 신뢰성의 부족 현상을 한마디로 말하자면 동일한 프롬프트를 입력했을 때도 다른 내용의 결과가 출력될 수 있다는 것이다. 이는 GPT와 같은 LLMs가 기능적으로 주어진 텍스트의 다음 부분을 예측하는 데 초점을 맞추고 있기 때문으로 볼 수 있다. LLMs는 조건부 확률을 기반으로 다음 단어를 선택하는데 이 확률은 주어진 텍스트, 즉 이전에 나왔던 단어와 그 순서 등에 의해 결정된다. 따라서 GPT와 같은 LLMs의 경우 주어진 텍스트, 즉 인간이 입력한 프롬프트(prompt)의 역할이 중요한데 이 프롬프트가 모호하거나 여러 해석이 가능할 때 결과의 다양성이 증가할 수 있다. LLMs가 출력 결과의 다양성을 조정할 수 있도록 하는 하이퍼 매개 변수(hyper parameter, GPT의 경우 'temperature')를 제공할 수 있는 이유 역시 이러한 구조적 특징에 기인한다. 그런데 내용 분석과 같은 텍스트 분석에 투입되는 기사, 트윗, 블로그 등의 글은 대체적으로 인간이 교육을 받고 판단해야 할 정도로 복잡하며 모호한 경우가 다수이다. 이는 LLMs의 입장에서는 여러 관점으로 해석될 수 있는 모호한 프롬프트에 해당될 수 있으며 결과적으로 LLMs는 구조적인 측면에서 출력 결과의 일관성을 항상 담보할 수 없는 불안정한 분석 도구라고 볼 수 있다.

프롬프트를 정교하게 정의하여 LLMs가 다양성을 나타낼 수 없도록 한다고 하더라도 여러 가지 요소들이 LLMs의 결과물의 일관성에 영향을 미칠 수 있다. 이에 대한 다양한 논의 중 흥미로운 결과는 바로 LLMs가 작동하는 과정에서의 계산 효율성이 출력 다양성에 영향을 미칠 수 있다는 것이다(Feng et al., 2024). 이들은 "LLMEffiChecker"라는 시스템을 개발하면서 LLMs에 많은 부하를 가할 경우 효율성이 저하하는 현상을 확인했다. 특히 단어가 단 하나 추가되었을 뿐임에도 이를 해독(decoding)하기 위해 복잡한 연산이 진행됨으로 인해 결과물을 생성하는데 상대적으로 오랜 시간이 요구되는 현상이 확인되었다. 이는 단순히 연산에 오랜 시간이 걸리는 것을 떠나 LLMs에 부하를 가함으로써 기존에 생성된 토큰 시퀀스, 즉 원래 단어를 생성하는 맥락에서 벗어나 다른 후보 토큰, 즉 대체 단어에 집중하게 할 수 있다는 점을 보여준다. 이로 인해 출력 결과물의 일관성이 유지되지 않을 수 있다는 것이다. GPT나 Claude와 같은 상용

LLMs의 경우 이용자의 컴퓨터의 성능에 의존하는 것이 아니라 기업이 보유한 컴퓨팅 파워에 의해 연산되는데, 이 때 어떤 이유에서건 LLMs가 연산되는 컴퓨터에 부하가 과다하게 가해질 경우 연산 성능이 저하되어 결과의 일관성이 유지되지 않을 수 있다는 것이다. 그 외에도 모델 자체의 편향성(bias toward yes-response, Dentella et al., 2023)이나 모델의 크기(Peng et al., 2024) 등에 의해서도 결과의 일관성이 유지되지 않을 수 있다. 이는 철저하게 외부 요인으로 인해 결정되는 현상이며 이런 외부적인 요소로 인해 결과물의 일관성이 유지되지 못할 수 있다는 것은 분석 도구로서의 불안정성을 보여주는 특징이라고 할 수 있다.

그러나 이런 문제로 인해 LLMs를 아예 측정 도구에서 배제하기 보다는 부족한 신뢰도 문제를 해결하는 방법을 고려하는 것이 더 효과적이다. 실제로 기존의 CTAM의 경우 언어 모델의 부족한 타당성 문제를 인간 코더와 미세 조정을 통해 해결하고 있다. 따라서 LLMs의 신뢰도 문제 역시 분석 방법의 절차적 방법을 통해 해결할 수 있다고 볼 수 있다.

본 연구에서는 이전에 LLMs를 시범적으로 활용했던 기존 연구들 중 타이 등(Tai et al., 2024)의 연구에서 사용된 다수의 동일 분석 결과를 생성하는 방법을 차용할 것이다. 측정 도구의 신뢰도를 측정하는데 있어 일반적으로 사용되는 방법은 측정-재측정(test-retest), 코더 간(inter-coder), 코더 내(intra-coder) 신뢰도를 측정하는 것이다. 이 중 다수의 동일 분석 결과를 생성하는 것은 서로 다른 코더들이 동일 문항에 대해 동시에 측정하는 것으로 볼 수도 있고, 반대로 동일한 코더가 동일한 문항을 반복적으로 측정하는 것으로 볼 수도 있다. 이는 LLMs의 개별 행위를 독립적인 것으로 볼 것인지에 따라 달라질 수 있는데, 본 연구에서는 동일 입력에 대해 동일하지 않은 결과가 나올 수 있다는 LLMs의 특성에 주목하여 개별 분석 결과를 서로 다른 주체(agent)가 수행한 결과로 간주하고 코더 간 신뢰도를 파악할 것이다. 이런 절차를 추가함으로써 LLMs의 분석 도구로서의 활용 가능성을 확인할 수 있을 것이다.

LLMs 자체의 활용 가능성을 검증하는 것에 더해, 어떤 종류의 텍스트에 대해 LLMs의 분석 성능이 달라질 수 있는지 역시 확인할 필요가 있다. LLMs의 경우 내부의 작동 원리가 블랙박스와 같이 명확하지 않기 때문에(Feng et al., 2024) 작동 결과가 어떻게 나타날지 사전에 예측하기 어렵고 실제 분석에 투입한 뒤 결과를 확인하는 방식으로 테스트를 진행할 필요가 있다. 또한 일률적으로 텍스트를 분석하는 것이 아니라 투입되는 데이터, 분석 내용 등 다양한 형태로 유목을 구분하여 각각의 조건에 따라 분석 성능이 어떻게 달라지는지를 검증할 필요도 있다. BERT와 같은 컴퓨터 기반 텍스트 분석 도구의 경우 매우 강력하지만 동시에 인간과 같은 연역적 판단을 한다고 보기에는 어렵다. 따라서 투입되는 텍스트에서 사용되는 단어의 종류나 배치가 달라질 경우 유사한 텍스트라 하더라도 분석 결과가 다르게 나타날 수 있다. LLMs 역시

언어 모델이며 연역적 판단은 미숙하다고 알려져 있다(Han et al., 2024). 특히 GPT와 같은 LLMs가 구조적으로 수학적 연산을 처리하는데 취약하다는 보고(Gandolfi, 2024)를 볼 때 모든 텍스트 자료에 대해 동등한 수준의 처리 능력을 보이는 것은 아니라는 추정이 가능하다. 또한 본 연구에서 초점을 맞추고 있는 정보 추출(information extraction) 영역에서도 숫자로 구성된 정보의 경우 텍스트에 비해 추출 및 인식이 잘 이루어지지 않는다는 보고도 존재한다(Rasool et al., 2024). 따라서 분석 대상인 텍스트가 어떤 내용을 다루고 있느냐에 따라 그 분석의 정확도가 달라질 가능성이 있다.

이는 단순히 분석 재료가 되는 텍스트 종류에 따른 분석 성능의 차이를 비교하는 것이 아니다. 생성형 AI의 등장과 함께 다학제적 관심이 커지고 있는 기계 행동 차원에서 LLMs에 대한 다양한 테스트의 필요성에 의거한 것이다. 기계 행동(machine behavior)은 지능형 기계, 특히 인공 지능으로 구동되는 기계가 나타내는 동작과 반응을 말한다(Rahwan et al., 2019). 인공지능과 같은 기계가 인간과 환경에 대해 사회적, 문화적, 경제적, 정치적 영역을 포함하여 여러 맥락에서 상호 작용하는 방식을 다학제 측면에서 연구해야 한다는 관점이다. 기계 행동에 대한 연구는 본질적으로 컴퓨터 과학, 심리학, 사회학 및 윤리와 같은 분야에서 파생된 학제 간 연구이며 따라서 광범위한 접근 방식은 기계가 다양한 대상과 상호 작용하는 과정의 복잡성과 의미를 이해하는 데 필수적이다.

AI와 인간의 협업은 더 이상 미래에 도래할 개념적인 존재가 아니라 현재 우리 삶에 영향을 미치는 시스템이 되었다. 그럼에도 불구하고 생성형 AI의 작동원리의 불투명성은 AI의 작동과 그 결과에 대한 인간의 불신을 불러온다. 부이스만(Buijsman, 2024)은 투명성(transparency, 시스템 전체에 대한 정보 제공)과 설명 가능성(explainability, 개별 결정의 이유를 설명)을 구분하고, 투명성은 정확성과 공정성 같은 특징을 다루지만 대중에게 AI 시스템이 어떻게 작동하는지에 대한 충분한 통찰을 제공하지 못하고 이로 인해 불확실성과 불신이 생긴다고 주장한다. 또한 AI의 행동이 예측하기 어려운 점이 있기 때문에 이를 이해하지 못하거나 AI가 어떻게 결정을 내리는지 예측할 수 없는 경우 인간은 이에 대한 불신을 가지게 되며, 단순히 AI의 작동 알고리즘을 이해하는 것만으로는 신뢰를 얻을 수 없다고 주장한다.

따라서 AI의 기계 행동, 즉 LLMs가 특정 조건에서 어떤 방식으로 행동하는지에 대한 구체적인 테스트 결과가 요구된다. 기계 행동을 이해한다는 것은 결국 기계가 다양한 사회적 맥락과 대상에 대해 어떻게 반응하고 상호작용하는가에 대한 탐색이다. 특히 LLMs를 일종의 기계 또는 인공지능으로 간주한다면 LLMs가 어떤 맥락에 어떻게 행동하는지를 파악하는 것이라고 볼 수 있다.

본 연구에서는 LLMs라는 기계(machine)가 텍스트 분석이라는 행동(action)을 수행하는 과정에서 그 대상 텍스트가 어떤 내용이냐에 따라(context) 결과가 달라질 수 있다는 점에서 기계 행동을 탐색하고자 한다. 특히 본 연구에서 분석 대상으로 삼는 뉴스의 경우, 일반적으로 정치, 사회, 경제 등과 같이 여러 장르로 분류되며 각 장르에 따라 텍스트의 형식과 내용에 큰 차이가 있다. 또한 동일 장르의 기사라 하더라도 그 안에서 다루는 정보의 내용에 따라 내용이 달라질 가능성이 크다. 따라서 본 연구에서는 뉴스의 장르, 그리고 정보의 종류에 따라 LLMs를 통한 분석의 결과, 즉 신뢰도와 타당도의 차이에 대해서도 검증할 것이다. 특히 연역적 판단과 수학적 처리에의 한계가 있는 LLMs의 특성을 고려할 때 경제 장르의 기사가 상대적으로 신뢰도와 타당도가 낮게 나타날 것이라고 예상할 수 있다. 역시 마찬가지로 수치를 많이 다루게 되는 통계 정보 역시 신뢰도와 타당도가 낮게 나타날 수 있다.

3. 연구 방법

1) 연구 문제

연구 문제 1. 반복 생성된 LLMs의 생성물은 높은 내적 신뢰도를 보이는가?

첫 번째 연구 문제는 LLMs가 만들어내는 결과가 일정하지 않다는 우려에서 시작한다. 내적, 외적 이유로 인해 LLMs는 그 성능에 제약을 받을 수 있으며, 예측하기 어려운 순간에 나타나는 이러한 성능 제약으로 인해 동일한 프롬프트에도 일정하지 않은 결과가 나타날 수 있다. 본 연구에서는 이에 대한 대안으로 타이 등(Tai et al., 2024)의 연구에서 사용한 동일 프롬프트에 대한 다수의 응답 결과를 생성하는 방법을 차용할 것이다. 다수의 응답 결과를 생성하고 이 중 특정한 응답이 다수를 차지하는지의 여부를 통해 LLMs가 동일한 프롬프트에 대해 일정한 결과를 도출하는지, 즉 내적 신뢰도가 있는 분석 도구인지를 확인할 것이다.

연구 문제 2. LLMs가 생성한 결과는 인간 코딩 결과와 비교하여 높은 타당도를 보이는가?

두 번째 연구 문제는 일반적인 CTAM의 절차 중 하나인 인간 분석 결과와의 일치도를 확인하는 것이다. 이는 분석 결과의 외적 타당도를 검증하는 절차로 볼 수 있으며, LLMs가 생성

한 결과가 실제 활용할 수 있는 데이터인지를 확인하는 절차로 생각할 수 있다. BERT를 포함한 CTAM 기법들을 활용한 연구들은 대부분 이런 형태의 검증 방식을 차용하고 있으며(Song et al., 2020) 또한 언어 모델이 생성한 결과물의 외적 타당도를 검증하기 위한 가장 현실적이고 타당한 방법이라 할 수 있다. 따라서 본 연구에서도 LLMs가 생성한 결과물에 대해 인간 코딩 결과와의 비교를 통해 그 타당도를 검증할 것이다.

연구 문제 3. 숫자를 다루는 정보 근거와 뉴스 장르의 경우 LLMs 생성물의 신뢰도와 타당도는 낮게 나타날 것인가?

연구 가설 1-1. 경제 뉴스의 경우 타 장르 뉴스에 비해 신뢰도와 타당도가 낮게 나타날 것이다.

연구 가설 1-2. 통계 정보의 경우 타 정보에 비해 신뢰도와 타당도가 낮게 나타날 것이다.

전체 텍스트 데이터를 대상으로 외적 타당도에 대한 검사와 함께 분석 대상인 텍스트의 종류에 따른 차이 역시 살펴볼 필요가 있다. 특히 LLMs의 성능이 과연 텍스트의 종류에 따라 달라질 것인지 검증할 필요가 있는데, 이는 이해하기 어려운 내부 작동 원리를 가지고 있는 LLMs의 특성 상 다양한 조건에 대한 작동 결과를 테스트해볼 필요가 있다. 따라서 본 연구에서는 뉴스 장르에 따른 신뢰도와 타당도의 차이도 확인할 것이다. 특히 수치를 다루는데 있어 한계가 있다고 알려진 LLMs를 고려할 때 경제 뉴스, 그리고 통계를 다룰 때 다른 경우에 비해 신뢰도와 타당도가 낮게 나타날 것이라고 가정할 수 있다.

2) 분석 자료

본 연구의 분석 자료는 서울대 팩트체크 센터와 언론정보학과 중점연구소가 함께 제작한 뉴스 코딩 결과를 활용하였다. 이 데이터는 914개의 팩트체크 기사에 대해 각각의 기사가 어떤 근거를 활용하여 팩트체크를 진행하였는지에 대한 인간 코딩 결과를 담고 있다. 914개의 기사 중 두 개 이상의 주장을 하는 기사의 경우 여러 주장에 대해 각각 코딩되었으며 결과적으로 이 데이터는 총 1213개의 측정 값을 가지고 있다. 데이터를 생성하는 과정에는 총 3명의 코더가 투입되었다. 교육을 통해 각 코더 간 신뢰도(크롬바흐 알파)가 0.7 이상이 되는 것을 확인한 뒤, 전체 기사를 1/3으로 나누어 각 코더에게 배당하여 코딩을 진행하였다.

1213개의 팩트체크 기사 측정 값은 정치, 사회, 경제 등 여러 분야에 걸쳐 있는데, 본 연구에서는 이 중 측정 값의 개수 상위 4개 분야, 98개의 기사를 비율에 맞춰 무작위로 추출하여 분석 데이터로 활용하였다. 상위 4개 분야의 경우 전체 데이터에서 정치(517개), 사회(371개), 경

제(234개), 생활/문화(30개)이며, 무작위 추출 결과 정치(44개), 사회(32개), 경제(20개), 생활/문화(2개)의 측정 값들이 선택되었다. 최종적으로 이렇게 무작위 추출된 98개의 팩트체크 기사 측정 값에 대한 인간 코딩 결과가 분석용 데이터로 선정되었다.

코더들은 팩트체크 기사에 대해 어떤 근거를 사용했는지의 여부를 체크했다. 근거의 종류는 통계, 보고서, 법률, 제도/규정, 다른 뉴스 기사, 보도 자료, 기타 온라인 자료, 인터뷰의 8개 항목으로 구성되어 있으며, 각 항목은 그 유무를 총 5개까지 체크하였다. 즉 해당 기사에 특정한 종류의 근거가 있다면 총 5개까지 있다고 코딩된다. 예를 들어 하나의 팩트체크 기사에서 3개의 통계 자료를 근거로 사용하고 있다면 아래의 그림 1과 같이 코딩된다.

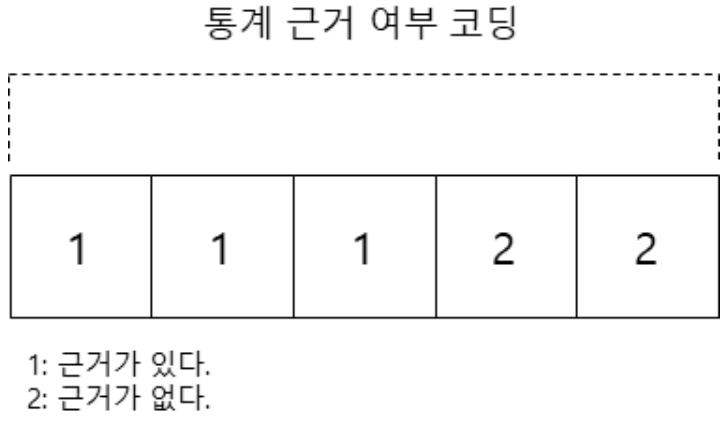

Figure 1. How to code for evidence in fact-checked articles

개별 근거에 대해 총 5개의 코딩 항목이 주어지며, 해당 근거 내용이 있다면 앞에서부터 1을 입력하는 방식이다. 이런 방식으로 코더들은 하나의 팩트체크 기사에 있는 검증 주장에 따라 8종류의 근거 항목들에 대한 40개의 체크 리스트를 작성하였다.

이와 같이 텍스트에 있는 특정한 정보의 유무를 체크하는 방식은 텍스트 처리 방법 중 정보 추출(information extraction)에 해당한다고 볼 수 있다. 정보 추출의 경우 기존의 CTAM을 활용할 경우 다른 종류의 텍스트 처리보다 어려운 영역이지만, LLMs에서는 능숙하게 수행할 수 있을 것이라 기대할 수 있다(Dagdelen et al., 2024). 따라서 LLMs의 성능과 활용 가능성을 평가하는 목표를 가지고 있는 본 연구에서는 상대적으로 난이도가 높은 정보 추출 방식을 사용한 이 데이터를 활용하여 LLMs으로 하여금 인간 코딩과 동일한 작업을 수행하게 한 뒤, 그 결과를 인간 코딩과 비교하여 정확도를 평가한다.

3) LLMs 분석

선정된 98개의 기사 텍스트에 대해 인간 코딩과 비교가 가능하도록 LLMs의 분석 결과를 생성하였다. 이를 위해 우선 개별 기사에 대해 코딩을 진행할 수 있게 해주는 프롬프트를 제작하였다. 최초에는 코딩 북 정보를 그대로 프롬프트로 전달하거나 텍스트로 변환하여 입력하는 방식을 사용했으나 이 방법으로는 답변의 안정성이 유지되지 않았다. 특히 8개 근거 종류 중 답변이 누락되거나 일부만 응답하는 등의 문제점이 지속적으로 나타나 프롬프트를 정비할 필요성이 제기되었다.

이 문제를 해결하기 위해 본 연구에서는 프롬프트를 정리해주는 서비스 "prompty"(https://chatgpt.com/g/g-aZLV4vji6-prompty)를 활용하였다. 해당 서비스는 "ChatGPT"를 기반으로 개인이 제작한 서비스로, 원하는 결과를 얻기 위한 프롬프트 정교화 작업, 즉 프롬프트 엔지니어링(prompt engineering) 작업을 간편하게 제공해주는 서비스이다. 이용자가 LLMs가 수행하기 원하는 내용을 입력하면 이 서비스는 적절한 프롬프트를 생성하여 제공해준다. 이 서비스를 통해 최종적으로 생성된 프롬프트는 〈Table 2〉와 같다.

Table 2. Prompts Used in Analysis

```
'''Analyze the provided article text based on the following criteria for content analysis. Ensure each criterion is addressed separately and clearly.

text:{이 위치에 기사 본문이 입력됨}

Criteria for content analysis:

**Source Usage**:
  - **Statistics** (up to 5): "1 = Yes, 2 = No"
    - Stat1, Stat2, Stat3, Stat4, Stat5
  - **Reports** (up to 5): "1 = Yes, 2 = No"
    - Report1, Report2, Report3, Report4, Report5
  - **Laws** (up to 5): "1 = Yes, 2 = No"
    - Law1, Law2, Law3, Law4, Law5
  - **Rules** (up to 5): "1 = Yes, 2 = No"
    - Rule1, Rule2, Rule3, Rule4, Rule5
  - **Other News Articles** (up to 5): "1 = Yes, 2 = No"
    - News1, News2, News3, News4, News5
  - **Press Releases** (up to 5): "1 = Yes, 2 = No"
    - Press1, Press2, Press3, Press4, Press5
  - **Online Sources** (up to 5): "1 = Yes, 2 = No"
    - Web1, Web2, Web3, Web4, Web5
  - **Interviews** (up to 5): "1 = Yes, 2 = No"
    - Interview1, Interview2, Interview3, Interview4, Interview5

Please provide the analysis in a structured format, ensuring all criteria are addressed accurately.'''
```

본 연구에서는 GPT-4o 모델에 프롬프트와 기사 본문을 투입하여 분석을 진행하였다. 당시에는 GPT-4o 모델이 초기 모델만 존재하여 따로 버전을 지정하지 않았다. 동일 프롬프트를 통해 98개의 기사를 총 10번씩 분석하게 하여 그 결과를 저장하였다. 또한 세션 내 학습효과를 제거하기 위해 각 분석은 모두 새로운 세션을 열어 진행하였다. 만약 실시간으로 이용자가 입력한 프롬프트와 이에 대한 답변이 학습된다면 이런 방법으로도 학습효과를 피하긴 어렵지만 일반적으로 LLMs 서비스 사업자들이 매우 정제된 데이터만 언어 모델에 학습한다는 점을 고려하면 즉각적인 학습이 일어나기 어려울 것으로 판단하였다. 그러나 총 10회의 분석 과정에서 후반부로 갈수록 정확하지 않은 내용이 출력되는 문제가 발생했다. 응답이 숫자로 출력되지 않고 텍스트로 출력되거나, 엉뚱한 답변이 나타나는 등의 문제가 나타나 7회차부터 정상적인 답변으로 사용할 수 없는 분석 결과가 나타나기 시작했으며, 10회차의 경우에는 절반에 가까운 분석이 정상적으로 출력되지 않아 에러로 입력되었다는 한계점이 발견되었다.

4) 신뢰도와 타당도 평가

본 연구에서는 LLMs의 결과에 대한 신뢰도와 타당도를 평가하게 된다. 신뢰도는 한 기사에 대한 10번의 분석을 LLMs를 통해 수행하면서 10개의 결과가 동일한지 평가하는 것이며 타당도는 LLMs가 생성한 측정 값이 인간 코더가 생성한 측정 값과 동일한지 평가하는 것이다. 일반적으로 평가 내용의 일치도, 즉 신뢰도의 경우 크리펜도르프 알파(Krippendorf's alpha)나 크론바흐 알파(Chronbach alpha), 코헨의 카파(Cohen's Kappa), 플레이스의 카파(Fleiss' Kappa) 등을 사용하여 평가할 수 있다. 그런데 본 연구의 경우 평가해야 할 5개 문항의 값이 단순한 개별 문항이 아니라는 것에 주목해야 할 필요가 있다. 본 연구에서의 8개 근거 분야에 대한 각 5개씩의 문항은 일반적인 설문이나 내용 분석 문항과는 달리 독립적인 영역을 측정하는 것이 아니라 모두 동일한 내용이 존재하는가의 여부에 대한 응답이며, 이 문항이 5개인 이유는 최대 5개까지 존재 여부를 체크할 수 있기 때문이다. 따라서 일반적인 설문이나 실험 등의 측정 문항과는 차이가 있는 이 측정값에 대해 일반적인 일치도 평가 방법을 사용하는 것이 적절하지 않을 수 있다. 따라서 본 연구에서는 신뢰도와 타당도를 가장 단순한 방법을 사용하여 측정할 것이다. 5개의 측정 값이 모두 일치할 경우에는 일치, 단 하나라도 다를 경우 일치하지 않는 것으로 평가한다.

정리하자면 신뢰도의 경우 개별 뉴스 기사에 대해 10번의 동일 분석을 반복하고 이 중 동일한 응답의 비중을 측정하고, 타당도의 경우 인간 코딩의 결과와 동일한 응답의 비중을 측정하는 방식으로 이루어진다. 단, 일부 기사에 대한 gpt의 응답이 총 10회를 채우지 못함에 따라 단

순 빈도 측정이 아닌 전체 응답 횟수 기반 동일 응답의 비율로 평가된다. 즉, 어떤 기사의 경우 8번, 어떤 기사의 경우에는 9번의 GPT 분석 결과가 나타났기 때문에 이 분석 횟수를 모수로 삼아 각 기사에 대한 GPT의 응답들의 일치율을 측정한다.

Table 3. Distribution of GPT Response Counts by News Genre

기사 장르	6회 응답 기사 수	7회 응답 기사 수	8회 응답 기사 수	9회 응답 기사 수	10회 응답 기사 수	총합계
경제		3		7	10	20
사회		3	3	7	19	32
생활/문화					2	2
정치	1	2	4	12	25	44
총합계	1	8	7	26	56	98

이후 뉴스 장르와 근거 분야에 따라 일치율들의 평균을 구해 보고하는 방식으로 일치도 분석이 이루어진다. 타당도의 경우 이전 연구들(DeButts & Pan, 2024; Jürgens & Stark, 2022; Rains et al., 2023)에서 약 60~80% 정도의 타당도를 보였던 것을 기준으로 본 연구에서는 70%의 타당도를 기준으로 평가를 진행할 것이다. 신뢰도의 경우 이전 연구들에서 주로 사용되었던 평가 도구를 사용할 수 없으므로, 본 연구에서는 임의로 타당도와 동일하게 70%의 답변이 일치해야 신뢰도가 확보된 것으로 판단할 것이다.

5) 기술 통계

Table 4. Frequency of Evidence Use by Evidence Area, News Genre

뉴스 장르	통계	보고서	법률	규정	타 뉴스	보도자료	온라인	인터뷰	전체 평균
경제	0.8	0.8	0.6	0.05	0.0	0.05	0.0	1.25	3.55
사회	0.406	0.438	0.75	0.0	0.125	0.0	0.063	0.844	2.625
생활/문화	0.5	1.0	0.0	0.5	0.0	0.5	0.0	1.0	3.5
정치	0.318	0.295	0.773	0.023	0.25	0.068	0.045	1.25	3.023
전체 평균	0.889	0.909	1.414	0.061	0.303	0.101	0.081	2.202	3.010

우선 팩트체크 뉴스에서 사용된 정보들의 근거 분야, 그리고 뉴스 장르별로 근거가 어느 정보의 빈도로 사용되었는지 확인하였다. 정보들의 근거 분야에서는 인터뷰(2.202)가 가장 많이

사용되었으며 그 뒤로는 법률(1.414), 보고서(0.909), 통계(0.889)가 뒤를 따랐다. 다른 언론사 뉴스(0.303), 보도자료(0.101), 온라인 게시글(0.081), 제도/규정(0.061)은 상대적으로 사용 빈도가 적었다. 이는 팩트체크의 근거로 사용 가능한 인터뷰나 보고서, 통계가 주로 사용되었으며 반대로 그 진위 여부를 확인하기 어려운 정보들의 경우 사용 빈도가 적었다고 해석할 수 있다. 뉴스 장르를 기반으로 보면 경제(3.55), 생활/문화(3.5), 정치(3.023), 사회(2.625) 순으로 근거가 사용되었다. 뉴스 장르, 정보 근거의 분포에 대한 카이제곱(χ^2) 분석 결과는 유의미하지 않은 것으로 나타나 특이한 분포는 확인되지 않았다($\chi^2 = 294.0$, $p = 0.218$).

4. 분석 결과

1) LLMs를 통한 반복 분석과 신뢰도

첫 번째 연구 문제인 LLMs의 분석 도구로서의 신뢰도를 평가하기 위해 총 10번의 반복 분석을 수행하고 그 결과가 얼마나 일정한지 확인하였다. 이 과정은 몇 가지 차원에서 이루어질 수 있는데, 본 연구에서는 우선 첫 번째 응답이 이후 나타나는 9개의 응답 중 몇 개와 동일한지를 평가하였다. 앞서 설정한 기준에 따라 첫 번째 응답을 포함한 최대 10개의 분석 결과 중 70% 이상이 동일해야 신뢰도가 확보되었다고 간주한다. 이는 본 연구에서 제기했던 LLMs의 분석 안정성을 확인하기 위한 것으로, LLMs를 통해 텍스트 분석을 한 번만 분석을 수행하더라도 이 응답이 신뢰할 수 있는가에 대한 검증이다.

Table 5. The Average Frequency of GPT Results that Matched the First Output Result

뉴스 장르	통계	보고서	법률	규정	타 뉴스	보도자료	온라인	인터뷰	전체 평균
경제	0.458	0.438	0.698	0.752	0.932	0.843	0.904	0.590	0.702
사회	0.552	0.564	0.628	0.723	0.865	0.882	0.739	0.603	0.694
생활/문화	0.222	0.389	0.889	1.000	0.889	0.833	0.833	0.889	0.743
정치	0.602	0.708	0.799	0.738	0.784	0.796	0.731	0.544	0.713
전체 평균	0.549	0.599	0.724	0.741	0.843	0.834	0.771	0.580	0.705

분석 결과, 전반적으로는 첫 번째로 생성된 텍스트 분석 이후에도 다수의 동일한 결과를 생성하는 것으로 나타났다. 전체 평균으로는 첫 분석 이후 9번의 반복된 분석 결과 중 약 70.5%의

응답이 첫 번째 분석과 동일한 결과를 생성했다. 뉴스 장르를 기준으로 평가할 때도 생활/문화(74.3%), 정치(71.3%), 경제(70.2%), 사회(69.4%) 분야에서 모두 70%에 가까운 일치 빈도를 보였다. 신뢰도의 판단 기준인 70%의 측면에서는 사회 장르 뉴스를 제외한 모든 장르가 기준을 넘은 것으로 나타났으며, 사회 분야 뉴스 역시 거의 70%에 가까운 수치로 나타났다. 따라서 첫 번째로 생성된 텍스트 분석 결과를 신뢰도 있는 평가 결과로서 판단할 수 있는 가능성을 보였다.

그러나 전체 평균과 달리 근거 분야와 뉴스 장르를 조합하면 일부 영역에서는 주의를 기울여야 할 필요성이 나타난다. 특히 근거 분야 중 통계(54.9%), 인터뷰(58.0%), 보고서(59.9%)의 경우 일치율이 상대적으로 낮은 수치를 보였으며, 특히 통계 근거의 경우 경제(45.8%), 생활/문화(22.2%) 뉴스 장르에서 나타났을 때 매우 낮은 일치율을 보였다. 이는 경제와 생활/문화 뉴스 분야에서 통계 수치가 근거로 활용되었는지 여부를 LLMs를 통해 분석할 경우 첫 번째 분석 결과를 신뢰하기 어려울 가능성이 있다는 것을 의미한다. 이는 보고서가 근거로 활용되었는지의 여부를 경제(43.8%)와 생활/문화(38.9%) 뉴스 분야에서 평가할 때도 동일하다.

Table 6. The Average Frequency of GPT Results that Matched Dominant Output Result

뉴스 장르	통계	보고서	법률	규정	타 뉴스	보도자료	온라인	인터뷰	전체 평균
경제	0.511	0.594	0.708	0.736	0.864	0.776	0.804	0.630	0.703
사회	0.657	0.609	0.705	0.706	0.791	0.819	0.766	0.674	0.716
생활/문화	0.550	0.400	0.800	0.900	0.800	0.750	0.750	0.800	0.719
정치	0.626	0.683	0.768	0.753	0.778	0.803	0.768	0.586	0.721
전체 평균	0.611	0.635	0.736	0.737	0.800	0.802	0.774	0.628	0.715

이번에는 첫 번째 출력 결과가 아닌, 10번의 분석 중 가장 다수의 분석 결과가 얼마만큼의 비중을 차지하는지 확인하였다. 이 값이 70% 이상일 경우 앞서 설정한 기준의 일치율을 보여 신뢰도를 확보했다고 볼 수 있다. 전반적으로 앞선 분석에 비해 수치가 상승했으며, 이는 첫 번째 분석 결과가 가장 다수의 분석 결과가 아닐 가능성이 높다는 것을 보여준다. 뉴스 장르별로는 모든 장르가 70%보다 높은 수치를 보였다. 근거 분야에서는 타 뉴스, 보도자료, 온라인 자료, 법률, 규정 부문에서 70% 이상의 일치율을 보여 대부분의 영역에서 일치도가 확보된 것으로 볼 수 있었다. 다만 통계(61.1%), 인터뷰(62.8%), 보고서(63.5%)에서 기준치보다 낮은 일치율이 나타났다.

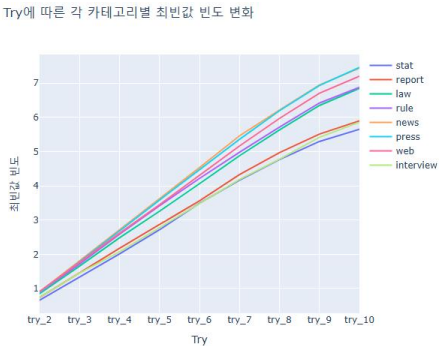

Figure 2. Change in frequency of the top output result by number of trials

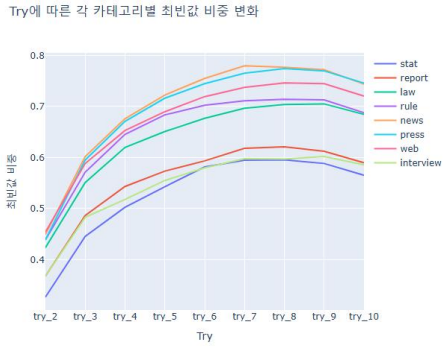

Figure 3. Change in weight of top output results by number of trials

Figure 4. Change in frequency of the top output result by number of trials (removing response errors)

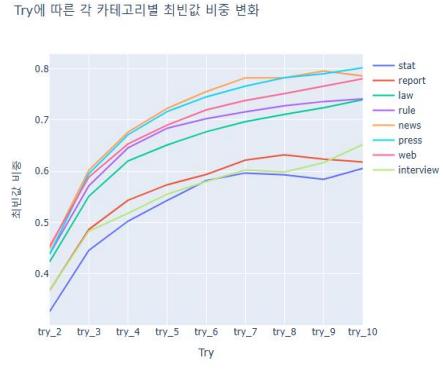

Figure 5. Change in weight of top output results by number of trials (removing response errors)

총 10회의 반복 분석을 수행하는 동안 최빈의 분석 결과가 얼마나 누적되었는지 확인한 결과, 빈도상으로는 시행 횟수가 증가할수록 최빈값이 지속적으로 증가하는 것으로 나타났다. 분석 결과의 최빈값을 시행 횟수로 나눠 시행 횟수가 증가할수록 정답이라고 할 수 있는 최빈값이 전체 중 얼마만큼의 비중을 차지하는지 역시 확인하였는데, 전반적으로 7~9회에서 최대 비중을 보이는 것으로 나타났다(그림 3). 전반적으로 8회차 또는 9회차부터 비중이 감소하는 것으로 나타났지만, 이는 분석 과정에서 GPT에 오류가 발생해 7~10회차 분석 결과가 나타나지 않은 일부 분석 결과에 의한 것으로 볼 수 있다. 실제로 응답 오류를 제거하고 최빈값의 비중이 반복 수행이 진행됨에 따라 얼마나 증가하는지를 확인한 결과, 7~9회에서 하락하는 경향이 나타나지 않았다(그림 5). 단, 최빈값의 비중이 감소하지는 않았지만 7회차 분석부터 최빈값이 비중의 증

가 속도가 감소하는 것으로 나타났으며, 몇몇 영역에서는 최빈값이 비중이 감소하기도 하였다. 이는 10회의 반복 안에서는 여전히 반복 측정을 통한 신뢰도가 안정적으로 나타나는 기준점을 파악하기 어렵다는 것을 나타낸다.

따라서 첫 번째 연구 문제는 부분적으로 긍정되었다고 볼 수 있다. 10번의 분석 중 첫 번째 응답을 기준으로 삼았을 때 상당수의 조건에서 첫 번째 응답이 가장 다수의 응답이 되는 것으로 확인되었으며, 최빈의 분석 결과를 사용할 경우에도 다수의 조건에서 기준 비율 이상의 동일 응답이 나타나는 것도 확인할 수 있었다. 이는 LLMs가 텍스트 분석에 사용될 때 신뢰성 있는 분석 도구로서 기능할 수 있는 가능성을 보여준다. 다시 말해, LLMs의 경우 지시를 내릴 때마다 출력되는 결과가 달라질 것이라는 우려가 있지만, 프롬프트를 정교화하여 텍스트 분석을 수행할 경우 그 결과가 일정하게 나타날 가능성이 높기 때문에 해당 분석 결과를 신뢰할 수 있다는 것이다. 그럼에도 불구하고 일부 영역에서는 최빈의 출력 결과가 다수를 차지하지 못하는 것으로 나타났다. 즉 10번의 반복 분석을 수행할 때 마다 일정하지 않은 결과가 나타났으며 동시에 가장 다수의 분석 결과가 70% 이상을 차지하지 못했다. 이런 경우에는 LLMs가 안정적인 텍스트 분석 도구로서 신뢰받기 어렵다. 따라서 이를 해결하기 위해서는 LLMs를 통한 텍스트 분석을 수행하는 과정에서 텍스트를 종류에 따라 적절하게 분류한 뒤, 각 조건에 따라 다수의 분석 결과를 중복 생성하고 최빈의 분석 결과를 활용할 수 있는지를 검증하는 단계를 추가하는 것이 LLMs를 더 신뢰도 있는 분석 도구로서 활용할 수 있는 방법이라고 볼 수 있다.

2) 인간 코딩 결과와의 일치도

LLMs를 활용한 텍스트 분석의 외적 타당도를 검증하기 위해 인간 코더가 분석한 결과와 비교하여 그 일치도를 확인하였다. 10번의 분석을 수행한 LLMs의 분석 결과 중 가장 다수의 출력값을 GPT의 응답으로 선정하고 이를 인간 코딩 결과와 일대일 비교를 수행하였다. 인간 코딩 결과와

Table 7. Results of Agreement between Multiple GPT Outputs and Human Coding Results

뉴스 장르	통계	보고서	법률	규정	타 뉴스	보도자료	온라인	인터뷰	전체 평균
경제	0.100	0.500	0.600	0.800	0.900	0.850	0.950	0.600	0.663
사회	0.469	0.594	0.688	0.813	0.906	0.969	0.813	0.438	0.711
생활/문화	0.500	0.500	0.500	1.000	1.000	1.000	1.000	1.000	0.813
정치	0.614	0.591	0.614	0.727	0.864	0.932	0.841	0.364	0.693
전체 평균	0.459	0.571	0.633	0.776	0.888	0.929	0.857	0.449	0.695

GPT의 코딩 결과가 일치할 경우 1, 일치하지 않을 경우 2로 코딩한 뒤 각 조건 별로 일치율의 비율을 확인하였다.

분석 결과, 인간 코딩과 LLMs의 분석 결과는 약 69.5%가 일치하는 것으로 나타났다. 이는 이전 연구(DeButts & Pan, 2024; Jürgens & Stark, 2022; Rains et al., 2023)에서 보통 60~80% 정도의 정확도를 보인 언어 모델의 결과물을 활용했다는 점에서 이 결과는 활용 가능한 수준의 일치도라고 볼 수 있다. 특히 내용 분석 등(Krippendorf, 2012)과 같은 통계 분석 방법에서도 0.7을 일반적인 기준점으로 사용한다는 점에서 LLMs를 활용한 텍스트 분석이 CTAM에 사용될 수 있는 가능성을 보였다고 할 수 있다. 경제 뉴스의 경우 70%의 약간 못 미치는 정확도를 보여 가장 낮은 것으로 나타났다. 그 외 장르들의 경우에는 모두 70%를 넘거나 거의 다다른 정확도를 보였다. 정보 근거 분류에서는 인터뷰(0.449), 통계(0.459), 보고서(0.571)에 대해서는 매우 낮은 정확도를 보였으며 법률(0.663)은 70%에 가까이, 나머지 근거들의 경우 정확도가 70%가 넘는 것으로 나타났다. 결국 전체 평균의 경우에는 상당히 높게 나타났지만 텍스트의 종류나 추출하고자 하는 정보의 종류에 따라 그 정확도, 즉 외적 타당도가 낮게 나타날 수 있다는 점에서 주의해야 할 필요가 있다. 비록 본 연구의 기준인 70%에는 약간 못 미쳤지만 연구 문제 2 역시 부분적으로 긍정할 수 있다. 즉, LLMs를 통해 생성한 텍스트 분석 결과는 인간 코딩 결과와 상당히 유사하며 적절한 분석을 위한 외적 타당도를 갖췄다고 볼 수 있다.

3) 뉴스 장르와 정보 종류에 따른 차이

연구 가설 1-1은 동일한 분석 결과를 기반으로 뉴스 장르에 따른 신뢰도와 타당도의 차이를 확인하는 것이다. 특히 수치를 주로 다루는 경제 뉴스가 타 장르 뉴스에 비해 신뢰도와 타당도가 낮게 나타날 것으로 예상하였다. 분석 결과, 신뢰도와 타당도에서 경제 뉴스가 가장 낮은 수치를 보이는 것으로 나타났다. 이를 통해 연구 가설 1-1은 채택되었다.

연구 가설 1-2의 경우 정보의 종류에 따른 신뢰도와 타당도의 차이를 확인하는 것이며, 역시 수치가 주로 나타나는 통계 정보에서 신뢰도와 타당도가 낮게 나타날 것으로 예상하였다. 분석 결과, 신뢰도에서 통계 정보가 가장 낮은 것으로 나타났으나 타당도에서는 인터뷰 정보가 가장 낮은 수치를 보였다. 이를 통해 연구 가설 1-2는 부분적으로 채택되었다.

비록 본 연구의 연구 가설에서는 경제 뉴스, 통계 정보만을 대상으로 가설을 검증했으나, 그 외에도 기준에 도달하지 못한 뉴스 장르와 정보의 종류가 나타났다. 특히 정보 종류에서 보고서와 인터뷰의 경우 통계 정보와 유사한 수준의 신뢰도와 타당도를 보였는데, 숫자를 잘 인식하지 못할 것이라는 기존 가정과는 다른 결과라고 할 수 있다.

또 한 가지 특이한 점은 뉴스 장르와 정보 종류의 조합으로, 경제 뉴스와 생활/문화 뉴스에서 통계와 보고서 정보를 다루는 경우 신뢰도와 타당도 양쪽에서 급격한 수치의 하락이 나타났다. 심지어 경제 뉴스에서 통계 정보를 다루는 경우 타당도가 0.1로 나타나 인간 코더와의 일치도가 10%에 불과했다. 이는 앞서 언급한 바와 같이 LLMs가 수치를 다루는데 있어 한계가 있고, 이로 인해 수치가 중요한 소재로 다뤄지는 텍스트에서는 성능이 크게 하락할 수 있다는 문제를 제기한다. 또한 인터뷰 정보 역시 신뢰도와 타당도 양쪽에서 낮은 수치를 보였으나 생활/문화 뉴스 장르에서만 높은 수치가 나타난 것도 특기할 점이다.

5. 논의

1) 분석 결과에 대한 논의

본 연구에서는 BERT 등의 분석 도구를 사용하고 있던 기존의 CTAM 과정에 LLMs를 도입할 수 있는지 그 가능성을 테스트하고자 하는 목적으로 시작했다. 이를 위해 LLMs를 통해 텍스트 분석을 수행하고, 그 결과에 대한 내적 신뢰도와 외적 타당도를 평가하는 방법으로 분석을 진행하였다. 내적 신뢰도의 경우 총 10번의 동일 분석을 수행하여 처음으로 출력된 분석 결과와 동일한 결과가 어떤 빈도로 등장하는지를 평가하였고, 외적 타당도의 경우 동일 텍스트에 대한 동일한 인간 코딩 결과와 비교하여 그 일치도를 확인하였다.

분석 결과, LLMs가 진행한 텍스트 분석은 상당한 수준의 내적 신뢰도를 확보하고 있는 것으로 나타났다. 특히 첫 번째로 출력된 분석 결과가 가장 다수의 분석 결과인 경우, 그리고 최빈의 분석 결과가 전체 응답의 과반 이상인 경우가 많았다. 일관적인 분석 결과를 출력한다는 측면에서 LLMs가 안정적이고 신뢰할 수 있는 텍스트 분석 도구로 활용할 수 있는 가능성을 보였다고 할 수 있다. 10번의 분석 결과 중 가장 다수의 결과를 활용한 외적 타당도 분석 결과 역시 LLMs가 의미 있는 텍스트 분석 도구로 활용될 수 있는 가능성을 보였다. 인간 코더의 분석 결과와 약 70%에 가까운 일치도를 보였으며, 이는 일반적인 CTAM을 활용한 연구에서 사용하는 기준에도 부합하는 수치이다. 따라서 LLMs으로 진행한 텍스트 분석의 결과 역시 BERT와 같이 텍스트 분석에 사용했을 때 텍스트를 지시한 대로 적절하게 분석해주는 기능을 가지고 있다고 볼 수 있다.

그럼에도 불구하고 LLMs의 내적 신뢰도와 외적 타당도를 평가할 때 주의해야 할 점이 있다. 분석해야 할 텍스트의 종류와 정보의 종류에 따라 내적 신뢰도와 외적 타당도가 상당히 낮아

지는 조건이 확인되었기 때문에 만약 현재 분석하고자 하는 대상이 이 조건에 해당될 경우 LLMs는 의미 있는 텍스트 분석 도구로서 기능하기 어렵다. 문제는 내부 작동 구조가 블랙박스에 가까운 LLMs의 특성 상 어떤 조건에서 신뢰도와 타당도가 낮아지는지 확인하기가 쉽지 않다는 것이다. 따라서 전체적인 분석 결과로는 신뢰도와 타당도가 확인되었다 하더라도, 개별 분석 과정에서는 여전히 LLMs를 사용하기 위한 여러 가지 검증이 뒤따를 수 밖에 없다.

본 연구에서는 이런 분석 결과와 기존의 CTAM 분석 과정을 함께 고려하여 LLMs를 텍스트 분석에 활용하는 다음과 같은 절차를 제안한다. 우선 **1) 텍스트 분석을 위한 프롬프트를 적절하게 제작**한다. 이 과정은 일반적으로 프롬프트 엔지니어링(prompt engineering)이라고 불리며 원하는 결과를 얻어내기 위한 적절한 프롬프트를 찾아내는 과정을 말한다. 또한 CoT와 같은 프롬프팅 기법, 그리고 프롬프트를 제작해주는 다양한 서비스 등을 이용하는 것을 통해 원하는 결과가 출력될 때까지 테스트를 진행한다. 다음 단계는 **2) 동일한 분석을 다수 수행하여 내적 신뢰도를 확인**하는 것이다. 전반적으로 LLMs의 텍스트 분석은 어느 정도 수준의 내적 신뢰도를 보여주고 있지만, 특정한 조건의 경우 내적 신뢰도가 확보되지 않을 가능성이 있다. 따라서 현재 진행하고 있는 분석 대상이 과연 그런 조건에 해당하는지 아닌지를 확인하여 현재 LLMs가 유용한 분석 도구로 활용될 수 있을지 아닐지를 검증해야 한다. 다음 단계로는 **3) 인간 코더 결과물을 통해 외적 타당도를 확인**하는 것으로, 일부 샘플을 통해 인간 코더 분석을 수행하고 이 결과와 LLMs 분석 결과를 비교함으로써 검증을 진행한다. 이러한 전반적인 검증 과정은 일반적인 CTAM에서도 사용되는 것으로, CTAM에 대한 기존 연구에서 구체적인 검증 방법(예. 필요 샘플의 수, 필요 일치도 수치 등)을 참고할 수 있다. 기존 CTAM 방법론과 다른 점은 내적 신뢰도 검증 부분이 추가된 것으로, 다수의 LLMs의 분석 결과 중 어떤 결과를 무엇으로 결정할 것인가에 대한 의사 결정에 대한 부분이며 이는 추후 더 심도 깊은 논의가 필요하다.

정리하자면, 본 연구의 결과를 통해 LLMs를 텍스트 분석 도구로 사용함에 있어 몇 가지 주의사항을 확인할 수 있었다. 1) 신뢰도 확보를 위한 다수의 반복 분석, 2) 타당도 확보를 위한 인간 코딩 결과와의 비교, 3) 텍스트의 종류에 따른 분석 결과의 편차 발생이 바로 그것이다. 이 사항들은 LLMs를 텍스트 분석 도구로 사용하기 위해 해결해야 하는 장애물이며 동시에 연구 방법과 절차의 개발을 통해 앞으로도 지속적으로 탐구되어야 할 필요가 있다. 특히 LLMs의 구조적 한계를 인식하고 4) 수치를 다루는 텍스트, 5) 연역적인 추론이 필요한 텍스트를 다룰 경우 더 세심한 주의와 관찰이 필요하다.

2) 한계점 및 추후 연구를 위한 제언

본 연구에서는 기존의 CTAM 분석 방법을 기반으로 LLMs를 분석 도구로서 활용할 수 있는 가능성을 확인했지만, 그럼에도 불구하고 LLMs의 활용 가능성은 추후 더 많은 영역의 검증이 필요하다. 무엇보다 본 연구에서 LLMs의 활용 가능성을 보였다 하더라도 이러한 결과가 앞으로 다른 연구, 다른 데이터를 대상으로도 계속해서 동일하게 나타날 것이라는 증거라고 할 수는 없다. 특히 정보 추출(information extarction)이라는 방식으로 분석을 진행해서 LLMs가 효과적으로 분석을 수행한 본 연구의 결과는 어디까지나 하나의 사례일 뿐이다. 따라서 앞으로 여타 다른 방식의 내용 분석에 대해서도 LLMs의 분석 성능에 대한 평가가 지속적으로 수행되어야 할 필요가 있다. LLMs가 어디까지나 블랙박스로서 내부 작동 원리에 대한 이해의 부재, 그리고 작동 원리에 대한 이해가 충분하더라도 생성된 결과물에 대한 불신을 완전히 불식하기 어려운 이상 LLMs를 활용한 텍스트 분석은 항상 검증의 대상이 되어야 하며, 앞으로도 더 다양한 데이터와 텍스트, 분석 방법을 대상으로 지속적인 테스트가 진행되어야 할 필요가 있다.

구체적으로는 내적 신뢰도를 검증하는 과정에 대한 정밀한 검증이 추가로 필요하다. 과연 충분한 내적 신뢰도를 얻기 위해서는 일관된 결과가 복수의 분석 중 어느 정도 수준의 비중을 차지해야 하는지에 대한 기준이 제시될 필요가 있다. 본 연구에서는 단순하게 가장 최빈의 결과를 활용했지만 이 기준 자체도 논의의 대상이 될 수 있다. 또한 분석 텍스트와 분석 방법의 종류에 따라 내적 신뢰도가 낮은 조건이 나타날 수 있는데, 이 때 어떤 방법을 통해 내적 신뢰도를 확보해야 하는지 역시 추가적으로 논의될 필요가 있다. 특히 본 연구에서는 기존의 다양한 코더 간, 코더 내 신뢰도 측정 방식을 사용하지 못하고 단순 빈도 측정 방식으로 신뢰도를 확인했다는 점에서 큰 한계가 있다. 이를 위해 후속 연구에서는 기존의 신뢰도 측정 방식을 사용할 수 있는 데이터를 활용하여 LLMs의 신뢰도를 평가할 필요가 있다. 또한 타당도가 낮을 경우 BERT와 같은 기존의 언어 모델에서는 학습용 데이터를 구축하여 모델을 미세 조정하거나 추가 학습을 진행하는 방식을 사용한다. 그렇다면 LLMs에서의 텍스트 분석도 동일한 방법을 사용할 것인지, 아니면 다른 방법을 동원할 수 있는지에 대한 검토가 진행되어야 한다. 두 번째로는 적절한 프롬프트를 만들어내는 과정에 대한 정규화 작업이 필요하다. 본 연구에서는 상용화된 서비스를 이용했지만 어떤 방법을 사용하는 것이 적절한 텍스트 분석 결과를 만들어 낼 것인가에 대한 추가적인 검증이 요구된다. 특히 프롬프트의 경우 분석 성능에 직접적인 영향을 미치는 요소인 만큼 어떤 절차를 거쳐야 하는지에 대한 명확한 방법론이 구축될 필요가 있다.

또한 반복적인 텍스트 분석을 수행하면서 LLMs가 이 내용에 대해 학습을 진행하는지에 대해서도 명확하게 밝힐 필요가 있다. 일반적으로는 언어 모델은 사전 학습된 상태로 공개되며

이를 조정하는 방법은 미세 조정을 수행하는 방법 외에는 없다고 알려져 있다. 자연어를 활용하는 언어 모델을 제작하는 기업들은 실시간으로 이용자와의 상호작용을 학습할 경우 잘못된 정보를 학습하여 인종 차별적 대화를 하게 된 마이크로소프트의 챗봇 "Tay"의 사례를 반면교사로 삼는 경우가 많기 때문이다. 즉, 실시간 학습은 제작자들이 통제할 수 없는 언어 모델이 만들어질 가능성이 높기 때문에 이에 대해서는 허용하지 않는 것으로 알려져 있다. 그러나 이는 언어 모델이 이용자의 프롬프트를 학습용 데이터로 사용하지 않는다는 것은 아니며, 아직까지 이에 대해 언어 모델 제작 기업들이 명확하게 밝힌 바는 없다. GPT를 제공하고 있는 openai 사의 개인정보 처리 정책(https://openai.com/policies/privacy-policy/)에 따르면 이용자의 데이터는 필요하다면 모델 학습에 사용될 수 있다. 따라서 단순히 실시간 학습이 이루어지지 않는다고 판단할 것이 아니라, 과연 어느 정도의 빈도와 시점으로 이용자의 프롬프트가 모델 학습에 사용되는지 명확하게 밝힐 필요가 있다.

LLMs의 하나의 세션을 한 명의 코더로 판단할 것인가 아니면 한 명의 코더가 반복적으로 입력한 작업으로 판단할 것인지에 대한 명확한 개념화 역시 필요하다. 즉, 코더 간 신뢰도를 측정할 것인가 코더 내 신뢰도를 측정할 것인가의 문제가 된다. 본 연구에서는 이에 대해 더 심도 깊게 논의하지 않고 단순하게 코더 간 신뢰도로 간주하였으나, 장기적으로 LLMs를 사용한 텍스트 분석을 수행하기 위해서는 반드시 개념화가 필요한 영역이라고 할 수 있다.

1000개가 넘는 전체 기사 데이터가 있음에도 일부 기사를 표본 추출하여 분석에 사용한 것도 본 연구의 한계점으로 볼 수 있다. 이는 컴퓨터 기반 텍스트 분석 기법의 장점인, 대량의 데이터를 빠르게 분석할 수 있다는 장점을 활용하지 못한 것이다. 다만 프로그래밍 기법을 사용하여 GPT 서비스를 이용할 경우 사용량에 따른 비용을 지불하게 되는데, 이 사용량은 입력 토큰의 수(≒입력한 텍스트의 양)와 출력 토큰의 수(≒출력한 텍스트의 양)에 따라 결정된다. 내용 분석을 수행할 경우 일반적으로 입력해야 할 텍스트의 양이 상당히 많아지며 이는 서비스를 이용하는 과정에서의 비용 증가로 이어지는 문제가 있다. 추후 연구에서는 이를 고려하여 사전에 연구를 계획할 필요가 있다.

마지막으로 외적 타당도의 경우 일반적인 CTAM에서도 주로 사용되는 만큼 여러 이전 연구를 통해 방법론이 자리잡고 있지만, 여전히 필요 샘플 데이터의 숫자나 최소 요구 일치도 등과 같이 통일되지 않은 영역이 남아 있다. 따라서 이 역시 추가적인 검증이 필요하다. 다만 이는 LLMs에만 국한되지 않고 CTAM 전체 영역의 숙제라고 볼 수 있다.

LLMs의 등장은 우리 사회 전체를 급속도로 인공 지능의 활용에 대해 고민하도록 인도하고 있다. 현재도 LLMs는 빠르게 발전하고 있으며, 추론 등의 기능 역시 추가되어 앞으로 어떤

능력을 보여줄지 확신하기 어려울 정도이다. 학술 영역 역시 이러한 영향에서 벗어나기 어려우며, 따라서 사회과학 영역 역시 이러한 LLMs를 어떤 방향으로 활용할 것인지에 대해 고민할 필요가 있다. LLMs가 단순히 텍스트를 인간과 유사하게 생성해내는 신기하기만 한 기술이었다면 안정성에 대한 위험을 무릅쓰고 연구에 활용할 필요성이 낮을 것이다. 그러나 최소한 텍스트 분석에 있어 LLMs는 놀라운 수준의 능력을 가지고 있으며 그 적용 영역은 날이 갈수록 확장되고 있다. 또한 지속적인 발전을 통해 앞으로 어느 정도 수준에 도달할 수 있을지 가늠하기 어려운 수준이다. 따라서 LLMs를 어떤 방법으로 활용할 것인지에 대한 지속적인 논의가 필요하다. LLMs는 이미 우리 앞에 놓여진 현실이며 이를 피해갈 수 없기에 어떻게 활용할 것인지에 대해 직시해야 한다.

References

Asbury-Kimmel, V., Chang, K. C., McCabe, K. T., Munger, K., & Ventura, T. (2021). The effect of streaming chat on perceptions of political debates. *Journal of Communication, 71*(6), 947-974.

Birkenmaier, L., Lechner, C. M., & Wagner, C. (2024). The search for solid ground in text as data: A systematic review of validation practices and practical recommendations for validation. *Communication Methods and Measures, 18*(3), 249-277.

Buijsman, S. (2024). Transparency for AI systems: A value-based approach. *Ethics and Information Technology, 26*(2), 34.

Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., ... & Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications, 15*(1), 1418.

DeButts, M., & Pan, J. (2024). Reporting after removal: The effects of journalist expulsion on foreign news coverage. *Journal of Communication, 74*(4), 273-286.

Deng, X., Bashlovkina, V., Han, F., Baumgartner, S., & Bendersky, M. (2023, April). *LLMs to the moon? Reddit market sentiment analysis with large language models*. Paper presented at the ACM Web Conference 2023, Austin, TX.

Dentella, V., Günther, F., & Leivada, E. (2023). Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences, 120*(51), e2309583120.

DiMaggio, P. (2015). Adapting computational text analysis to social science (and vice versa). *Big Data & Society, 2*(2). https://doi.org/10.1177/2053951715602908

Feng, X., Han, X., Chen, S., & Yang, W. (2024). LLMEffiChecker: Understanding and testing efficiency degradation of large language models. A*CM Transactions on Software Engineering and Methodology, 33*(7), 186.

Gandolfi, A. (2024). GPT-4 in education: Evaluating aptness, reliability, and loss of coherence in solving calculus problems and grading submissions. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00403-3

Han, S. J., Ransom, K. J., Perfors, A., & Kemp, C. (2024). Inductive reasoning in humans and large language models. *Cognitive Systems Research, 83*, 101155.

Javaji, P., Sreeya, P. S., & Rajesh, S. (2024, July). *Detection of AI generated text with BERT model.* Paper presented at the 2nd World Conference on Communication & Computing (WCONF), Raipur, India.

Jürgens, P., & Stark, B. (2022). Mapping exposure diversity: The divergent effects of algorithmic curation on news consumption. *Journal of Communication, 72*(3), 322-344.

Kim, S., Kim, S., Kim, Y., Park, J., Kim, S., ... & Lee, Y. (2023, November). *LLMs analyzing the analysts: Do BERT and GPT extract more value from financial analyst reports?* Paper presented at the 4th ACM International Conference on AI in Finance (ICAIF '23), Brooklyn, NY. https://doi.org/10.1145/3604237.3627721

Krippendorff, K. (2012). *Content analysis: An introduction to its methodology* (3rd ed.). Sage Publications.

Lamprianou, I. (2023). Measuring and visualizing coders' reliability: New approaches and guidelines from experimental data. *Sociological Methods & Research, 52*(1), 525-553.

Lee, V. V., van der Lubbe, S. C., Goh, L. H., & Valderas, J. M. (2024). Harnessing ChatGPT for thematic analysis: Are we ready? *Journal of Medical Internet Research, 26*, e54974.

Li, L., Ma, Z., Fan, L., Lee, S., Yu, H., & Hemphill, L. (2024). ChatGPT in education: A discourse analysis of worries and concerns on social media. *Education and Information Technologies, 29*(9), 10729-10762.

Miah, M. S. U., Kabir, M. M., Sarwar, T. B., Safran, M., Alfarhood, S., & Mridha, M. F. (2024). A multimodal approach to cross-lingual sentiment analysis with ensemble of transformer and LLM. *Scientific Reports, 14*(1), 9603. https://doi.org/10.1038/s41598-024-60210-7

Parker, M. J., Anderson, C., Stone, C., & Oh, Y. (2024). A large language model approach to educational survey feedback analysis. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-024-00414-0

Pelaez, S., Verma, G., Ribeiro, B., & Shapira, P. (2024). Large-scale text analysis using generative language models: A case study in discovering public value expressions in AI patents. *Quantitative Science Studies, 5*(1), 153-169.

Peng, C., Yang, X. I., Smith, K. E., Yu, Z., Chen, A., Bian, J., & Wu, Y. (2024). Model tuning or prompt tuning? A study of large language models for clinical concept and relation extraction. *Journal of Biomedical Informatics, 153*, 104630.

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., ... & Wellman, M. (2019). Machine behaviour. *Nature, 568*(7753), 477-486.

Rains, S. A., Harwood, J., Shmargad, Y., Kenski, K., Coe, K., & Bethard, S. (2023). Engagement with partisan Russian troll tweets during the 2016 US presidential election: A social identity perspective. *Journal of Communication, 73*(1), 38-48.

Rasool, Z., Kurniawan, S., Balugo, S., Barnett, S., Vasa, R., Chesser, C., ... & Bahar-Fuchs, A. (2024). Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using CogTale dataset. *Natural Language Processing Journal*, *8*, 100083.

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., ... & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication, 37*(4), 550–572.

Su, L. Y. F., Cacciatore, M. A., Liang, X., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2016). Analyzing public sentiments online: Combining human- and computer-based content analysis. *Information, Communication & Society, 20*(3), 406–427.

Suter, V., & Meckel, M. (2024, June). *Using GPT-4 for text analysis: Insights from English and German language news classification tasks*. Paper presented at the 1st Workshop on Reliable Evaluation of LLMs for Factual Information (REAL-Info 2024), Buffalo, NY. https://doi.org/10.36190/2024.31

Tai, R. H., Bentley, L. R., Xia, X., Sitt, J. M., Fankhauser, S. C., Chicas-Mosier, A. M., & Monteith, B. G. (2024). An examination of the use of large language models to aid analysis of textual data. *International Journal of Qualitative Methods, 23*. https://doi.org/10.1177/16094069241231168

Van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures, 15*(2), 121-140.

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

최초 투고일 2024년 12월 06일
게재 확정일 2025년 01월 28일
논문 수정일 2025년 01월 30일