JDFR Journal of Data Forensics Research
데이터포렌식연구, 2권 2호, 2025년 12월, 143-159면
Journal of Data Forensics Research, Volume 2, Issue 2, December 2025, Pages 143-159.
https://doi.org/10.12972/JDFR.2025.2.2.2
Print ISSN 3058-9460
Online ISSN 3058-9479

원저

Self-Verification 기반 검증 절차를 통한 비정형 텍스트 핵심 정보 추출: 사이버범죄 수사 지원의 관점에서

윤병휘[1], 김지온[2]

[1]한림대학교 글로벌학부 정보법과학전공 학사
[2]한림대학교 융합과학수사학과 교수

교신저자: 김지온, jion972@hallym.ac.kr

요약

온라인상에서 발생하는 사이버범죄의 종류와 발생 건수가 증가하면서 수사를 위해 분석이 필요한 비정형 텍스트 데이터양이 역시 비례하여 증가하고 있다. 이러한 막대한 양의 데이터를 다루는 수사관들의 업무적 부담을 줄이기 위해 수사에 필요한 핵심 정보를 신속하게 추출할 수 있는 자동화 기법이 필요하다. 그러나 파인튜닝을 통해 도메인에 특화된 대규모 언어 모델(LLM)을 구축하는 방법은 비용 부담이 크며, 클라우드 기반 모델은 민감 정보가 유출될 위험으로 실무 적용에 제약이 따른다. 또한 대규모 언어 모델을 별도의 조치 없이 사용할 시 환각과도 같은 문제로 인해 결과의 신뢰성 확보가 어렵다. 이에 본 연구는 보안이 보장된 온프레미스 대규모 언어모델을 활용해 한글 비정형 텍스트로부터 10가지 핵심 정보를 추출하고, 정확도 향상을 위해 '셀프베리피케이션(Self-Verification)' 기반 검증 절차를 제안한다. 네 종류의 온프레미스 모델을 대상으로 검증 절차 적용 전·후 F1 점수를 비교한 결과, 대부분의 모델에서 성능 향상이 확인되었다. 본 연구는 보안 제약이 있는 수사 환경에서도 온프레미스 대규모 언어 모델과 체계적인 검증 프로세스를 결합하면 더 높은 핵심 정보 추출 정확도를 확보할 수 있음을 시사한다.

주제어

데이터포렌식, 정보 추출, 사이버범죄, 대규모 언어 모델, 온프레미스

Open Access

Received: December 11, 2025
Revised: December 27, 2025
Accepted: December 27, 2025
Published: December 31, 2025

Original Article

Self-verification-based framework for key information extraction from unstructured text: A cybercrime investigation support perspective

ByungHui Yoon[1], Jion Kim[2]

[1]Undergraduate Student, Hallym University, Republic of Korea
[2]Professor, Hallym University, Republic of Korea

Corresponding Author: Jion Kim, jion972@hallym.ac.kr

ABSTRACT

With the continuous increase in the type and number of online cybercrimes, the volume of unstructured text data that must be analyzed for investigations is increasing. To reduce the workload of investigators handling this massive amount of data, automated methods are urgently required to rapidly extract the key information required for criminal investigations. However, building domain-specialized large language models (LLMs) through fine-tuning entails substantial costs, and cloud-based models face practical limitations in real-world investigative settings owing to the risk of leaking sensitive information. Additionally, when LLMs are used as-is without additional safeguards, errors such as hallucinations hinder ensuring the reliability of their outputs. To address these challenges, this study employs security-preserving on-premise LLMs to extract ten categories of key information from Korean unstructured text messages and proposes a self-verification-based validation procedure to improve the extraction accuracy. We compared the F1 scores before and after applying the verification stage across the four types of on-premise models and observed performance improvements for most of them. These findings suggest that, even under stringent security constraints in investigative environments, combining on-premise LLMs with a systematic verification pipeline can achieve higher accuracy in extracting the core information necessary to support cybercrime investigations.

KEYWORDS

data forensics, information extraction, cybercrime, large language model, on-premise

I. 서론

경찰청 공개 자료에 따르면 사이버범죄 발생 건수는 2022년 217,807건에서 2023년 약 241,842건으로 증가하였다. 2023년 사이버범죄 중 사이버사기는 총 155,715건으로 전체의 64.39%를 차지하였으며, 사이버사기의 피해액은 2022년 1조 8천억 원을 기록한 것으로 보고되었다[1]. 또한, 2025년 5월 27일 분당경찰서에 따르면, 최근 디지털 금융기술, 인공지능(AI), 메신저 등을 악용한 복합형 사이버범죄가 빠르게 확산 중이며, 특히 SNS를 통해 접근하는 신종 사기 유형이 잇따라 발생하고 있다[2]. 여러 명이 한 팀을 이루어 어떤 미션(과제)을 수행할 시 금전을 제공한다고 피해자들을 기망하여 팀가입비 명목 등으로 금전을 편취하는 수법의 '팀미션 사기'[3], 주식투자를 미끼로 투자자를 속이는 '주식리딩방 사기'[4], SNS 또는 데이팅 앱 등을 통해 피해자와 연애 관계를 구축한 뒤 가짜 투자기회를 제공하는 '피그 부처링 사기'[5] 등, 다양한 유형의 복합형 사이버사기가 등장하면서 수사관이 전처리와 분석을 해야 하는 데이터의 양과 복잡성 역시 증가한다. 사이버사기는 SNS 채팅, 문자메시지 등의 비정형 텍스트 데이터, 즉 구조나 규칙이 없는 텍스트를 통해 계좌번호 등과 같은 핵심정보가 교환된다. 이러한 비정형 텍스트로부터 핵심정보를 추출하는 과정은 데이터의 양이 많을수록 수작업으로 처리하기에는 상당한 시간이 소요되며 수사관 개개인의 부담은 더욱 커진다. 실제로 2024년에는 사이버 수사 경력 채용 전형으로 임용된 수사관 중 약 10%가 수사 과중과 적응 실패로 공직에서 떠나기도 하였다[6].

수사관들의 업무 부담을 줄이기 위해 자연어 처리 (Natural Language Processing)와 파인튜닝(fine-tuning)을 통해 사이버 사기 수사 지원에 특화된 모델을 구축하는 방법을 고려해 볼 수 있지만, 개발 과정에서 대규모 연산 자원, 인력, 비용, 광범위한 훈련 데이터 등이 요구된다[7]. 한편 ChatGPT와 같이 사전 학습이 되어있는 클라우드 서비스 기반 대규모 언어 모델(Large Language Model)은 구축 과정을 줄일 수 있다. 하지만 입력되는 정보가 모두 외부 사업자의 서버로 전송·저장되는 클라우드 서비스 특성상 수사 관련 데이터를 입력할 시 민감정보가 유출되는 2차 피해의 위험성이 존재한다. 즉, 피해자·피의자 인적 사항 등의 개인정보가 포함된 수사 데이터를 클라우드 서비스 기반 대규모 언어 모델을 통해 처리하는 행위는 부적절하며, 이에 보안 요구사항을 충족하는 대체 방안을 모색할 필요가 있다. 하지만 대규모 언어 모델은 실제로 존재하지 않는 출처와 사실이 마치 있는 것처럼 생성해내는 "환각"의 문제가 발생할 수 있어[8] 수사 환경에서 전적으로 신뢰하여 활용하기에는 정확도가 낮다. 따라서 본 연구는 온프레미스 대규모 언어 모델을 활용하여 사이버 사기 관련 비정형 텍스트로부터 핵심 정보를 추출한 후, 모델이 스스로 검증하는 셀프베리피케이션 절차를 적용함으로써 그 정확도를 높이고자 한다.

II. 이론적 배경

2.1. 범죄예방을 위한 빅데이터와 공개출처정보의 활용

세계가 빅데이터 시대에 접어들면서 사람들은 일상적인 활동을 통해 막대한 양의 데이터를 생산하고 있다. 빅데이터의 정의는 명확하게 하나로 정해져 있지 않으나[9], McKinsey & Company[10]는 빅데이터를 "일반적인 데이터베이스 소프트웨어 도구로는 수집·저장·관리·분석하기 어려울 정도의 규모를 가진 데이터셋"이라고 정의하며, 어느 정도 규모의 데이터셋을 빅데이터로 볼 것인지는 고정되어 있지 않고 유동적이라고 설명한다. 빅데이터의 범위에는 대한민국 국민의 일상생활에서 생성되는 데이터뿐만 아니라, 범죄 행위 과정에서 공개적으로 남

겨지는 각종 데이터도 포함된다. 이러한 공개 데이터는 수집·분석 과정을 거쳐 공개출처정보(Open Source Intelligence)로 활용될 수 있다.

김지온[11]은 사이버범죄 발생 건수가 증가하고 온라인 소통을 매개로 한 사이버범죄가 확산됨에 따라 수사기관에서는 다양한 형식의 데이터 간 연관 관계를 분석하여 사이버상에서 이루어지는 범죄징후를 신속하고 효과적으로 탐지 및 차단해야 할 필요성이 높아졌다는 점을 지적했다. 김성준[12]은 범죄 예방에는 예측 가능한 데이터와 이를 분석하는 기술이 결합하여 구축된 대한민국 범죄 예방 시스템의 주요 특징 중 하나로 범죄첩보분석 시스템과 과학적 범죄 분석 시스템에 축적된 데이터베이스를 제시하였다. 해당 데이터베이스는 개별 범죄 사건이나 잠재적 범죄자를 중심으로 서로 연동되며, 향후 인공지능 알고리즘을 통해 국내 범죄 발생 양상과 치안 환경에 적합한 범죄 발생 예측 시스템을 구축하기 위한 핵심 자원으로 활용될 수 있다. 이에 대한 사례로 현재 수사 단서와 핵심 정보로 이루어진 데이터를 수집하고, 이들 간의 연관 관계를 분석·추론하여 용의자를 특정하고 검거함으로써 추가 피해를 방지하는 사이버범죄 수사 단서 통합분석 및 추론시스템 개발 중이다[13]. 이러한 시스템을 구축하기 위해서는 데이터 수집뿐만이 아닌, 수집된 데이터가 목적에 맞게 활용될 수 있도록 정제·구조화하는 전처리 과정이 필수적이다.

2.2. 자연어 처리 기반 정보 추출의 한계

Maciej[7]에 의하면 기존의 연구 논문에서 데이터를 수작업으로 추출하던 방식을 자연어 처리(NLP) 등을 활용한 자동화된 데이터 추출로 대체하려는 시도가 증가하고 있다. 자연어 처리와 같은 기법은 많은 양의 연구 논문에서 데이터를 효율적으로 추출할 수 있게 해주지만, 초기 구축 단계에서 상당한 노력과 전문성, 그리고 학습을 위한 데이터셋, 즉 데이터의 집합이 필요하다고 Maciej는 주장한다. 초기 구축 단계를 거쳐 어느 분야에 특화된 정보 추출 프로그램이 완성되더라도 이후에 새로운 데이터가 유입되는 등의 환경 변화에 따라 모델을 지속적으로 최신화해야 하는 유지보수의 부담이 크다. 특히 사이버범죄 수사에서 다루어지는 데이터는 새로운 범죄 유형이나 범죄조직이 등장하는 경우, 범행 과정에서 생성되는 텍스트 데이터에 사용되는 용어와 표현 방식, 대화 구조가 기존 학습 데이터셋에 포함되지 않았을 가능성이 존재한다. 이러한 새로운 데이터의 유형을 반영하기 위해서는 추가적인 데이터 수집과 모델 재학습이 필요하며, 그 과정에서 상당한 시간과 인력이 추가로 투입되어야 한다. 따라서 새로운 유형의 사이버 범죄로 인해 데이터의 유형이 변화하더라도 과도한 인력 및 시간 자원 투입 없이 이에 적응하여 작업을 수행할 수 있는 유연성이 필요하다.

2.3. 대규모 언어 모델의 한계

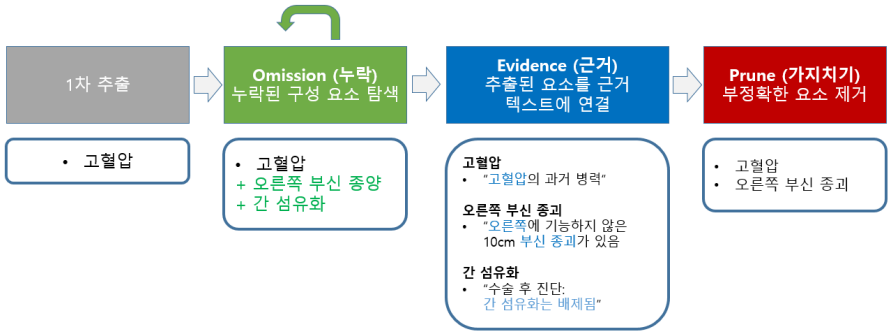
대규모 언어 모델은 이미 대량의 코퍼스(말뭉치)를 기반으로 사전 학습을 거친 모델이기 때문에, 별도의 추가 파인튜닝(fine-tuning)이나 도메인 특화 모델 학습 과정 없이도 다양한 자연어 처리 과제를 수행할 수 있다는 장점이 있다. 파인튜닝(fine-tuning)은 사전 학습된(pre-trained) 모델을 특정 작업이나 사용 목적에 맞게 재적응시키는 과정을 의미한다[14]. 하지만 대규모 언어 모델은 현재의 성능적 한계로 인해 자연어 처리를 대체할 수 있는 완벽한 대안으로 보기는 어렵다. Michael[8]은 대규모 언어 모델의 핵심 원리를 언어적 패턴을 예측하는 것이라고 설명하였다. 즉, “주어진 단어들의 순서를 보았을 때 다음에 올 가능성이 가장 높은 단어는 무엇인가?”라는 질문에 대한 답을 예측하는 것이지 사람이 사고하는 것과 같이 답을 유추해내는 것이 아니다[8]. 대규모 언어 모델이 통계적 분석을 통해 요청된 답을 생성하는 과정에

서 비논리적인 내용을 생성하는 현상인 환각(hallucination)과 같은 문제가 발생한다[15]. 사이버범죄 수사에서 다루는 데이터의 상당수는 비정형 텍스트(unstructured text)의 형태를 띤다. 비정형 텍스트는 일정한 서식을 따르지 않고 자유로운 형식으로 작성된 글을 의미하며, 이메일과 SNS 게시물로 작성된 텍스트가 포함된다. 범죄자와 피해자들이 카카오톡과 같은 SNS 메신저를 통해 남긴 대화 속에는 줄임말, 이모티콘, 오타, 비표준어, 중의적 표현 등이 혼재되어 있다. 이처럼 여러 정보가 일정한 형식 없이 뒤섞여 나타나는 텍스트 속에서 계좌번호, 금액, 인물 정보와 같은 핵심 정보를 자동으로 식별·추출하는 것은 대규모 언어 모델에게도 높은 정확도가 보장되지 않는 작업이다.

또한, ChatGPT를 비롯한 클라우드 기반 대규모 언어 모델은 입력되는 정보가 전부 외부 사업자의 서버로 전송·저장되는 구조적 특성상 수사 과정에서 수집된 데이터를 그대로 활용할 경우 개인정보 유출과 같은 2차 피해가 발생할 수 있는 위험성이 존재한다. 특히 수사기관은 디지털 증거, 피해자·피의자 인적 사항 등 민감한 정보를 처리한다는 점에서, 일반적인 상업용 클라우드 대규모 언어 모델을 직접 사용하는 것은 적절하지 않으며, 보안 요구사항을 충족하는 대체 방안을 모색할 필요가 있다.

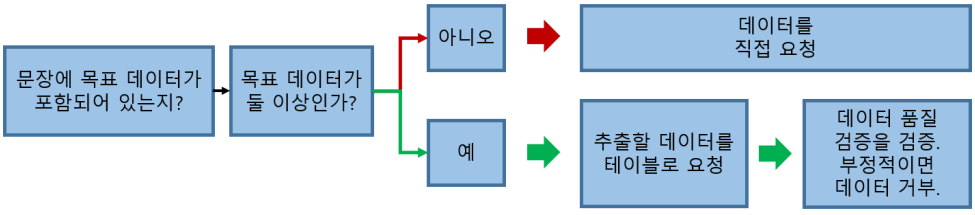
2.4. 선행 연구 분석

Gero et al.[16]은 프롬프트에서 퓨샷 학습에 기반을 둔 대규모 언어 모델을 활용하여 비정형 임상 텍스트로부터 임상 정보를 추출하는 방법을 제안하였다. 이 연구는 임상 노트와 임상시험 초록 등에서 진단 및 시술에 해당하는 ICD 코드, 임상시험 그룹, 약물명과 약물 상태 등의 정보를 리스트 형태로 추출하는 여러 임상 정보 추출을 목표로 삼았다. 기존 퓨샷 학습 기반 임상 정보 추출 기법은 전문가가 수작업을 정답 레이블을 부여한 라벨링 데이터(labled data)가 부족한 상황에서 유리하다는 장점이 있으나, 의료 영역에서 요구되는 수준의 정확도와 해석 가능성을 충분히 확보하지 못한다는 한계가 있다. 이에 Gero et al.[16]은 셀프베리피케이션이라고 부르는 단계적 검증 구조를 도입하여, 핵심 정보를 추출하는 과정에서 동일한 대규모 언어 모델에 서로 다른 프롬프트로 여러 차례 질의를 수행하여 스스로 자신의 1차 추출을 점검·보완하도록 하는 방법을 제안하였다. 셀프베리피케이션 검증 절차는 (1) 1차 핵심 정보 추출을 수행하는 단계, (2) 누락된 요소를 추가로 탐색하는 오미션(Omission) 단계, (3) 각 1차 추출값에 대응하는 근거 텍스트 구간을 찾는 에비던스(Evidence) 단계, (4) 제공된 근거를 바탕으로 잘못된 요소를 제거하는 프룬(Prune) 단계로 구성된다. 이 과정을 통해 대규모 언어 모델은 각 출력에 대한 근거를 명시하면서 스스로가 추출한 핵심 정보를 검증·수정할 수 있으며, 다양한 대규모 언어 모델과 여러 임상 정보 추출 작업에서 정확도가 전반적으로 일관되게 향상되는 것으로 확인되었다.



<Figure 1> 임상 정보 추출을 위한 셀프베리피케이션 절차의 예시 [16]

Polak & Morgan(2024)[7]은 재료과학 논문에서 벌크 탄성률(bulk modulus), 금속유리 임계 냉각속도(critical cooling rate), 고엔트로피 합금 항복강도(yield strength of HEAs) 값을 대규모 언어 모델과 제로샷(zero-shot) 프롬프트 설계를 통해 재료(Material), 값(Value), 단위(Unit)로 이루어진 삼중항(triplet)의 형태로 추출하고 검증하는 챗익스트랙트(ChatExtract) 방법을 제안하였다. 제로샷 프롬프트는 대규모 언어 모델이 스스로가 출력해야 하는 결과에 대한 예시 없이 오직 수행해야 할 작업에 대해 직접 설명하는 지시만을 제공하는 프롬프트 설계 방법이다[17]. 챗익스트랙트 방법은 문장 관련성 판별, 단일·다중 값 분기 및 세부 값 추출, 그리고 이진형 후속 질문을 통한 검증 단계로 구성된다. 대규모 언어 모델이 필요할 경우 '알 수 없음'으로 응답하도록 허용하는 형태의 질문을 제시하고, 추출된 데이터가 실제로 존재하지 않을 수 있다는 가능성을 명시함으로써, 존재하지 않는 데이터를 생성해내는 환각을 크게 억제하였다. 실제로 벌크 탄성률 데이터셋 실험에서 GPT-4는 전체 기준 정밀도 90.8%, 재현율 87.7%를 기록하였으며, GPT-3.5는 정밀도 70.1%, 재현율 65.4%를 기록하였다. 한편 현재 제안된 챗익스트랙트 방법은 특정 물성에 대한 재료-값-단위 단일 삼중항 추출을 기본 단위로 설계·평가하고 있어, 본 연구에서처럼 하나의 텍스트에서 서로 다른 유형의 다수 필드를 동시에 추출하는 문제를 직접적으로 다루지는 않는다. 또한, 후속 질문에서 추출된 데이터가 잘못되었을 가능성이 있다고 판단되면 그 텍스트에서의 추출된 데이터를 통째로 폐기하도록 설계되어 있어 정확도를 크게 향상시키지만, 이 과정에서 일부 유효한 데이터까지 함께 버려지는 트레이드오프(trade-off)가 존재한다.



<Figure 2> 간략화된 챗익스트랙트 방법의 흐름도[7]

앞서 진행된 본 연구의 선행 연구에서는, 사이버범죄 수사를 위한 조건을 충족하는 효율적인 전처리를 목표로 온프레미스(On-Premise) 환경에서 구동되는 EXAONE-3.5-7.8B-Instruct 대규모 언어 모델을 활용하여 카카오톡 텍스트 메시지로부터 핵심 정보를 추출하는 프로그램을 개발하였다. 온프레미스 환경이란 시스템의 데이터와 연산 자원을 외부 클라우드가 아닌 조직이 직접 관리하는 물리적 서버 내에서 운영하는 방식이다[18]. 본 연구는 Gero et al.[16]의 셀프베리피케이션 검증 절차와 Polak & Morgan[7]의 챗익스트랙트 방법에서 제안된 대규모 언어 모델 기반 검증 절차를 참고하여, 사이버사기 수사 도메인에 적합한 온프레미스 기반 핵심 정보 추출 프로그램을 설계하고, 이를 통해 범죄일람표 작성의 정확도를 제고하고자 한다.

III. 데이터 구조 및 연구설계

3.1. 데이터 구조

추출 대상 데이터의 맥락과 구조는 다음과 같다. 먼저 사이버범죄 전문가와의 자문을 통해, 범죄조직이 SNS로 대출 희망자를 모집한 뒤 허위 사업자 등록과 위조 서류를 이용하여 제2금융권 대출을 실행시키고 약 수백억 원을 편취한 실제 작업대출 사기 사건을 토대로 시나리오를

작성하였다. 이후 해당 시나리오를 바탕으로, 실제 수사 과정에서 수집되는 자료의 형식을 반영한 가상의 카카오톡 텍스트 메시지 데이터가 <Table 1>과 같은 형식의 엑셀 파일로 작성되었다.

<Table 1> "datetime", "sender" 그리고 "message"로 구성된 가상 사이버 사기 데이터셋의 일부

Datetime	Sender	Message
2021-12-31 09:03:00	박*훈 이사님	㈜다* (강**) ***도 **시 **동 0000 ****아파트 제***동 제**층 제****호 면적 67.86 KB시세 하한가 - 11,500 일반가 - 12,000 상한가 - 12,500 매매가 - 10,500 총7개동 490세대 담보대출대부 모집건 모집금액 7300만 이자 월1% 자금필요시기 6월30일 담보대출대부 7300만 월1% 모집건 1. 2시 40분뭇비해주세요~

주소와 주민번호와 같은 개인정보는 비식별화 처리되었다. 엑셀 파일의 칼럼은 datetime(메시지가 보내진 시간), sender(메시지를 보낸 사람의 이름), message(카카오톡 메시지 본문), 그리고 범죄일람표가 작성되기 위해 필요한 10개의 타깃 필드의 정답셋인 Ground-Truth 값으로 이루어졌다. Ground-Truth 값은 전문가가 준비한 가상 데이터셋과 함께 제공한 정답을 참고로 하여 작성하였다.

<Table 2> 대규모 언어 모델의 정확도 평가를 위해 작성된 정답셋의 일부 예시

대출일시_GT	대출자_GT	담보물건지주소_GT
12-31	홍길동(빅스모크)	경기도 **시 ***구 **동 0000 **아파트 제**동 제 **층 제*****호

3.2. 연구설계 및 가설

기존에 진행하였던 온프레미스 기반 대규모 언어 모델로 핵심 정보를 추출한 연구에서 도출한 84.33%의 정확도는 사이버 수사 영역에서 사용되기에는 부족했다. 본 연구에서는 정보 추출 정확도 개선을 위해 다음과 같은 차별적 시도를 진행하였다.

첫째, 대규모 언어 모델이 핵심 정보를 추출하기 위해 제공하는 지시와 요구사항을 더욱 이해하기 쉽게 작성하였다. 이전 연구에서 작성된 프롬프트는 대규모 언어 모델이 추출해야 하는 핵심 정보에 대한 설명이 부족하다고 판단하여 데이터를 추출하는 과정에서 대규모 언어 모델의 혼동을 줄일 수 있도록 자세한 설명을 추가하였다.

둘째, 잘못된 그라운드 트루스(Ground-Truth)값을 수정하였다. 기존에 제공된 그라운드 트루스 값은 실제 정답과는 다르게 작성되어 대규모 언어 모델이 정보를 올바르게 추출했음에도 불구하고 오답처리가 된 것을 확인하였다. 따라서 대규모 언어 모델의 핵심 정보 추출 정확도를 실제 성능보다 낮게 평가하는 일이 발생하지 않도록 원본 텍스트 메시지와 그라운드 트루스값을 비교하면서 검토를 진행하여 수정하였다.

셋째, 이전 연구에서는 Exaone 3.5-7.8 모델만을 사용하여 대규모 언어 모델 간의 한글 비정형 텍스트를 대상으로 한 핵심 정보 추출 정확도를 비교하지 않았다. 본 후속 연구에서는 Exaone와 Llama 계열 모델의 성능을 비교해보고자 한다.

넷째, 셀프베리피케이션 검증 절차를 도입하면서 오직 프롬프트만 주어진 대규모 언어 모델로부터 정확도가 얼마나 상승하였는지 확인하여 셀프베리피케이션 검증 절차의 효율성을 증명해보고자 하였다.

연구 가설로서는 프롬프트만을 사용한 대규모 언어 모델 대비 셀프베리피케이션 검증 절차를 통한 단계가 각 대규모 언어 모델 모델의 전체 및 필드별 정확도를 유의미하게 향상시키는지와 셀프베리피케이션 기반 검증이 어느 필드의 정확도 향상에 큰 기여를 하는지에 대한 가설을 설정하였다.

이 가설을 바탕으로 ① 각 대규모 언어 모델 별로 오직 핵심 정보 추출을 위한 규칙과 지시가 포함된 프롬프트만으로 이루어진 프로그램을 구동시켜 대규모 언어 모델의 기본적인 핵심 정보 추출 정확도를 확인하였고 ② 1차 추출값을 대규모 언어 모델의 추가 호출을 통해 셀프베리피케이션 기반 검증 절차를 거쳐 1차 단계에서 대규모 언어 모델이 잘못 추출한 값들을 수정하게 하여 정확도가 얼마나 올라가는지 확인하였다.

Ⅳ. 셀프베리피케이션 기반 핵심 정보 추출 및 검증 절차

4.1. 퓨샷 학습과 LangExtract 스키마 활용

본 연구에서는 Google에서 공개한 Gemini 기반 오픈 소스 라이브러리인 랭익스트랙트(LangExtract)가 제공하는 구조·퓨샷 기반 구조화 추출 방식을 참조하여 대규모 언어 모델의 1차 핵심 정보 추출 단계를 설계하였다. 랭익스트랙트는 대규모 언어 모델이 사용자가 정의한 출력 구조를 사용하도록 강제시키며, 대규모 언어 모델이 사용자의 의도대로 데이터를 추출할 수 있도록 퓨샷 학습을 통해 일관된 추출 결과를 생성하도록 유도하는 등의 핵심 기능을 제공한다. 이러한 랭익스트랙트의 기능들을 참조하여 본 연구의 1차 대규모 언어 모델 추출 단계를 설계하였다.

```
14) 아래 예시를 few-shot 가이드로 활용한다. Output Primer의 형식을 정확히 그대로 따른다.
###Example###
<ExampleInput>
매수자 :주식회사 놀란드 서울시 00 **동 0000 ****타워 제***동 제**층 제****호 면적 5월 6일 상환예정  333.68 XP시세 하한가 -  80,500 일반가 -
85,500 상한가 -  90,000 매매가 -  64,500 총 2개동  141세대 담보대출대부 모집건 법무비고객부담 5/2일  모집금액 5.8억 이자율1% 투자자모집할시
예치금 1% 자자금필요시기 잔금일 4월 11일 담보대출대부 5.8억 월 1% 모집건 1.홍길동/별빛토끼님 6100만 입금완료 2.임꺽정/치즈만두님 1500만입금완료 3.
박둘쇠/블랙고양이님 2천만 입금완료 4.최만수/초코송이님 1.6억만 입금완료 5.김봉길/소주한잔님 5천 입금완료 6.이춘복/한칠성님 1천만 입금완료 7.정두석/
무한도넛님 1천만(강호동님) 입금완료 8.장복동/스타크래프트테란고수님 3천만(입금완료) 9.이만두리/저그1황님 1천만(마일리지는 정복자님으로) 입금완료
10.오달수님/매치/뽀뽀님 200 입금완료 11.송학수/얼탱(이춘복) 2천만 입금완료 12.허사또리
</ExampleInput>
<ExampleOutput>
{"대출일시":"04-11","사용기간":"5월6일","대출자":"주식회사 놀란드 대표자 박상혁","담보물건지 주소":"서울시 00 **동 0000 ****타워 제***동 제**층
제****호","대출금액":"5.8억","대출금액(숫자)":"580000000","대출수수료":"월 1%","투자자":"홍길동/별빛토끼, 임꺽정/치즈만두, 박둘쇠/블랙고양이,
최만수/초코송이, 김봉길/소주한잔, 이춘복/한칠성, 정두석/무한도넛, 장복동/스타크래프트테란고수, 이만두리/저그1황, 오달수/매치/뽀뽀, 송학수/얼탱
(이춘복), 허사또리","금액":"6100만, 1500만, 2천만, 1.6억, 5천, 1천만, 1천만, 3천만, 1천만, 200, 2천만","금액(숫자)":"61000000, 15000000,
20000000, 160000000, 50000000, 10000000, 10000000, 30000000, 10000000, 2000000, 20000000"}
</ExampleOutput>
```

<Figure 3> 랭익스트랙트 구조에서 활용된 퓨샷 프롬프트의 일부

다음은 범죄일람표 작성에 필요한 대출일시, 사용기간, 대출자, 담보물건지 주소, 대출금액, 대출금액(숫자), 대출수수료, 투자자, 금액, 금액(숫자)의 10개 필드를 포함시킨 정보 추출 구조를 제시하였다. 정보를 추출하는 데 있어 대규모 언어 모델이 참고할 수 있도록 각 필드에 대한 의미적 정의, 허용 표현, 정규화 규칙을 프롬프트 상에서 상세히 명시하였다. 예를 들어 “금액”은 텍스트에서 확인한 금액 표현과 금액의 순서를 그대로 사용하며, “대출자”는 이름 뒤에 별명이 확인되면 함께 추출하도록 지시하는 등의 규칙을 제공하였다.

프롬프트 상에서 정의된 정보 추출 구조에 대한 소수의 대표적인 예시를 제공하기 위해 퓨샷 학습을 채택하였다. 전문가가 제공한 정답셋을 기반으로 입력된 비정형 텍스트의 예시와 그에 따라 올바르게 추출된 핵심 정보의 목록에 대한 예시를 각각 제공하였다. 또한, 각 예시는 10개 핵심 정보 필드가 모두 채워진 단일 JSON 객체로 제공하여 대규모 언어 모델이 올바른 출력 형

식을 내재화 할 수 있도록 유도하였다. 마지막으로 대규모 언어 모델이 자신의 출력에 대한 보조 설명을 배제하고 정해진 구조만 제공하는 JSON 블록을 제공하도록 하였고, 대규모 언어 모델이 추출할 대상인 데이터를 찾지 못하는 경우 “null”로 기입할 수 있도록 하였다.

```
def call_exaone_return_json(msg: str) -> Optional[dict]:
    """
    메시지를 LLM에 전달하여 JSON(dict) 반환. 실패 시 None.
    - 스키마+few-shot 예시 기반 구조화.
    (여기서는 1차 초안 + 내부 Self-check & Fix까지 포함)
    """
    system = """###Instruction###

역할: 당신은 비정형 부동산 대출 텍스트에 특화된 엔터프라이즈급 한국어 정보 추출 엔진 "EXAONE"이다.
대상: 전문 수사관/분석가.

당신의 임무는 정확히 10개의 필드를 추출하고 정확히 하나의 JSON 객체만을 반환하는 것이다. 아래 모든 요구사항을 지키며, 추가 설명 없이 JSON만 출력하라.

스키마 (키 이름과 순서는 반드시 정확히 일치해야 함):
["대출일시","사용기간","대출자","담보물건지 주소","대출금액","대출금액(숫자)","대출수수료","투자자","금액","금액(숫자)"]
```

<Figure 4> 대규모 언어 모델이 일관한 구조의 결과를 강제하는 프롬프트

4.2. 셀프베리피케이션 검증 단계

프롬프트에서의 자세한 목표에 대한 설명과 규칙을 제공한다는 전제하에도 대규모 언어 모델의 근본적인 한계인 환각에 의해 완벽한 결과를 기대하기는 어렵다. 따라서 본 연구에서는 대규모 언어 모델이 자신이 1차적으로 추출한 핵심 정보를 스스로 점검·보완하도록 하는 셀프베리피케이션 기반 검증 절차를 구축하였다. 전체적인 셀프베리피케이션 파이프라인은 오미션(Omission), 에비덴스(Evidence), 룰-프룬(Rule-Prune), 그리고 대규모 언어 모델 기반 SV 프룬(SV-Prune) 단계로 나뉜다. 각 단계는 셀프베리피케이션_pipeline(msg, data) 함수 내부에서 순차적으로 수행되며, 랭익스트랙트 출력 구조 기반 1차 추출 결과를 입력으로 받아 점진적으로 정제된 예측값을 생성한다.

4.2.1. 오미션 단계를 통한 데이터 누락 보완

오미션 단계는 셀프베리피케이션 파이프라인의 첫 번째 절차로서, “투자자”와 “금액”처럼 여러 데이터가 쉼표로 구분되어 나열되는 목록형 항목을 대상으로 한다. 본 연구의 추출 대상인 카카오톡 메시지에서는 다수의 투자자와 다양한 금액 표현이 섞여 있는 경우가 많아 1차 추출 단계에서 일부 투자자 이름 또는 괄호 속 별명·수취인 정보를 누락하는 문제가 자주 발생하였다. 오미션 단계의 목적은 대규모 언어 모델이 누락한 핵심 정보가 없는지 한번 더 확인하도록 하여, 데이터 추출의 정확도를 보다 정확하도록 보완하는 데에 있다.

오미션 단계는 1차 LangExtract 스키마 기반 추출 결과를 입력으로 받아, 투자자 필드와 금액 필드에 대해 각각 Omission 함수를 호출한다. 이때 각 핵심 정보 항목의 현재까지의 추출값은 하나의 문자열(예: “홍길동/별빛토끼, 임꺽정/치즈만두”)로 정리되어 있으며, 이는 원문 메시지와 추출된 각 핵심 정보에 대한 설명과 함께 대규모 언어 모델에 전달된다. 이 때 대규모 언어 모델에게는 메시지 출력 형식이 엄격하게 제한된다. 현재 리스트가 충분히 정확하다고 판단되는 경우에는 “SAME”이라는 단일 문자열만을 반환하도록 하고, 누락이나 오류가 존재한다고 판단될 때에만 전체 리스트를 새로 구성한 셀프베리피케이션 문자열 한 줄로 출력하도록 설계하였다. 예시로서 원문 메시지에 “홍길동/별빛토끼, 임꺽정/치즈만두, 순자/분리(이방원)”라는 투자자 목록이 등장하지만 1차 추출 결과에 “홍길동/별빛토끼, 임꺽정/치즈만두”만 포함

되어 있는 경우를 가정할 수 있다. 추출된 투자자 목록이 오미션 단계로 전달되면, 대규모 언어 모델은 “원문에 등장하는 모든 투자자를 누락 없이 포함해야 한다”는 프롬프트 규칙에 따라 기존 목록을 재검토하고, 세 번째 투자자를 포함한 “홍길동/별빛토끼, 임꺽정/치즈만두, 박순자/분리(원순재)”와 같은 새로운 셀프베리피케이션 문자열을 반환한다. 반대로 누락이 없다고 판단될 경우에는 “SAME”만을 반환함으로써, 불필요한 수정을 피한다.

오미션 단계에서는 대규모 언어 모델이 리스트를 과도하게 수정하는 것을 방지하기 위하여, 동일 필드에 대한 검토를 최대 2회까지만 반복하도록 제한하였다. 한편, 대규모 언어 모델이 “* 수정 필요:”나 “Candidate:”와 같이 불필요한 텍스트를 함께 생성하는 경우도 발생할 수 있는데, 이러한 텍스트는 별도의 필터를 통해 “정상적인 C셀프베리피케이션 문자열이 아닌 출력”으로 판정되어 무시되며 이때는 기존 리스트가 그대로 유지된다.

마지막으로, 오미션 단계에서 확정된 투자자 및 금액 리스트는 간단한 규칙 기반 후처리를 거친다. 투자자 리스트의 경우 이름 뒤에 “님”과 같은 존칭을 제거하고, 불필요한 공백을 정리하여 일관된 문자열로 정규화한다. 금액 리스트의 경우에는 괄호 안 설명, “입금완료”와 같은 보조 설명, 시간 표현 등 금전과 직접 관련이 없는 요소를 제거하고, 숫자와 단위만 남도록 정제한다. 이와 같은 설계를 통해 오미션 단계는 1차 추출 결과를 최대한 유지하면서도, 실제로 누락된 항목을 국소적으로 보완하는 기능을 수행한다.

4.2.2. 에비던스 단계에서의 근거 기반 검증

에비던스 단계는 셀프베리피케이션 검증 절차의 두 번째 단계로, 추출된 각 필드 값이 원문 메시지의 어느 구간에 근거하는지를 수집하는 역할을 한다. 이 단계의 목적은 현재 추출 결과가 실제 텍스트의 어떤 부분에서 유도되었는지에 대한 근거를 기록하는 것이다. 이를 통해 후속 단계에서, 근거가 없는 값과 근거가 있는 값을 구분하여 정제할 수 있다.

Evidence 단계에서 입력되는 정보는 두 가지이다. 첫째는 오미션 단계까지 반영된 현재 추출 결과인 10개의 데이터 목록으로 구성된 JSON 객체이며, 둘째는 해당 추출이 기반한 카카오톡 원문 메시지 전체 텍스트이다. 근거 추출을 위한 프롬프트는 대규모 언어 모델이 정확히 하나의 JSON 객체만을 출력하도록 요구하며, 그 구조를 “대출일시_근거”, “사용기간_근거”, “대출자_근거”, “담보물건지 주소_근거”, “대출금액_근거”, “대출수수료_근거”, “투자자_근거”, “금액_근거”의 8개 필드로 고정한다. 각 근거 목록의 값은 원문 메시지에서 그대로 복사 가능한 문자열이어야 한다. 예를 들어, 추출된 대출일시가 “03-03 (3월 3일)”인 경우에는 “자금필요날짜 3월3일”과 같이 해당 날짜가 언급된 구절이 “대출일시_근거”로 선택될 수 있다. 이와 비슷하게 추출된 대출자가 “홍길동”이라면 “매수인 홍길동”과 같은 표현이 “대출자_근거”로 기록된다. 특정 항목에 대하여 원문 텍스트에서 근거를 찾을 수 없는 경우에는, 해당 근거 항목의 값을 “null”로 표기하여 신뢰할 만한 근거가 없음을 명시하였다. 여기서 null은 해당 값이 존재하지 않거나 확인되지 않았다는 의미의 결측 값을 나타낸다. 마지막으로 대규모 언어 모델 출력 중 마지막 JSON 객체만을 추출한 후, 이를 해석하여 위에서 정의한 8개 근거 항목에 대응시키고 문자열 형태로 정규화한다. 이 과정을 통해 각 항목의 추출값과 원문 텍스트 사이에 일종의 “연결”이 형성되며, 다음 단계인 룰-프룬와 대규모 언어 모델 기반 SV-Prune에서 이 연결 정보를 바탕으로 추출값의 신뢰성을 판단하게 된다.

4.2.3. 근거를 기반으로 한 룰-프룬(Rule-Prune) 단계

룰-프룬 단계는 에비던스 단계에서 제공된 정보를 활용하되, 추가적인 대규모 언어 모델 호

출 없이 순수 규칙 기반으로 수행되는 필터링 절차이다. 이 단계의 설계 원칙은 “값을 새로 만들거나 수정하지 않고, 근거가 전혀 없는 값만 제거한다”는 점에 있다.

룰-프룬 단계는 추출 결과 JSON과 근거 JSON을 입력 받아 각 항목과 그에 대응하는 근거 항목을 매핑한다. 예를 들어, “대출일시”는 “대출일시_근거”와, “대출자”는 “대출자_근거”와, “금액”은 “금액_근거”와 연결된다. 이후 각 항목에 대해 다음과 같은 단순한 규칙을 적용한다. 해당 항목의 추출값이 “null”이 아니고 비어 있지 않은 반면, 대응하는 근거 항목이 “null”이거나 완전히 빈 문자열로 되어 있다면, 해당 항목은 원문 텍스트에 대한 근거가 전혀 없는 것으로 간주하고 그 값을 “null”로 덮어쓴다. 반대로 근거 문자열이 조금이라도 존재한다면, 값의 세부 내용이 다소 이상하더라도 룰-프룬 단계에서는 이를 수정하지 않고 그대로 유지한다. 투자자와 금액 항목에도 같은 규칙이 적용된다. 단, 이 단계에서는 쉼표로 구분된 개별 항목 하나하나를 조정하지 않고 전체 단위로만 판단한다. 즉, “투자자” 항목이 비어 있지 않지만 “투자자_근거”가 “null”인 경우에만 필드 전체를 “null”로 만들고, 그렇지 않다면 추가적인 수정은 가하지 않는다. 대규모 언어 모델이 생성한 값을 근본적으로 재구성하지 않으면서도, “어떤 근거에도 연결되지 않는 값”이 자동으로 제거되도록 함으로써, 후속 분석에서 무에서 기반한 추출 결과가 포함될 위험을 줄이는 효과를 갖는다.

4.2.4. 대규모 언어 모델 기반 셀프베리피케이션-프룬 단계

대규모 언어 모델 기반 셀프베리피케이션-프룬 단계는 셀프베리피케이션 절차의 마지막 단계이다. 앞서 설명한 룰-프룬 단계가 근거의 유무만을 바탕으로 추출된 항목의 삭제 여부를 결정하는 반면, 대규모 언어 모델 기반 셀프베리피케이션-프룬 단계는 추출 결과와 근거, 그리고 원문 메시지를 통합적으로 고려하여 값의 수정·삭제를 허용하는 더욱 적극적인 정제 기능을 수행한다. 이 단계에서 입력되는 정보는 오미션, 에비던스, 그리고 룰-프룬 단계를 거친 현재 추출 결과인 Prediction JSON 객체, 각 필드에 대한 근거 구문을 담고 있는 Evidence JSON 객체, 그리고 원문인 전체 카카오톡 메시지이다. 대규모 언어 모델은 이 세 가지 정보를 동시에 제공 받고, 각 필드가 정의에 부합하는지 근거와 모순되지 않는지를 재검토한다. 프롬프트에서는 각 필드에 대한 구체적인 정의와 판단 기준이 다시 한번 제시된다. 예를 들어, 프롬프트에서는 대출일시에 대해 “대출 필요 시점이나 자금 필요 날짜를 나타내는 표현이어야 하며, 소유권 이전일과 혼동되어서는 안 된다”고 명시하였으며, 대출자에 대해서는 “차주 또는 매수인과 같이 실제로 대출을 받는 당사자여야 하며, 투자자의 이름을 포함해서는 안 된다”고 규정하였다.

해당 단계에서 대규모 언어 모델은 Prediction JSON과 Evidence JSON을 비교하고, 원문 텍스트의 문맥을 함께 고려하여 다음과 같은 조치를 수행하도록 지시받는다. 정의에 명백히 어긋나거나 근거와 모순되는 필드 값은 “null”로 설정한다. 투자자 리스트에는 투자자가 아닌 이름이나 오타가 포함되어 있을 수 있는데, 이 경우 해당 항목을 제거하거나 수정한다. 금액 목록에서는 금전과 관련 없는 숫자나 시간 표현을 삭제하고, 원문에 근거가 분명히 존재하지만 1차 추출에서 누락된 항목이 있다면 목록에 추가할 수 있다. 잘못 추출된 날짜나 주소, 대출자 값에 대해서는, 근거와 원문을 참고하여 원문 문자열에 존재하는 올바른 값으로 교정하도록 유도한다. 이때 대규모 언어 모델의 출력은 1차 핵심 정보 추출 단계와 동일하게 10개 필드를 키로 갖는 단일 JSON 객체로 제한되며, “대출금액(숫자)”와 “금액(숫자)”처럼 숫자 기반 필드는 기존과 동일한 쉼표로 구분된 정수 문자열을 유지하는 방향으로 설계되었다.

V. 대규모 언어 모델 별 성능 비교

본 연구에서는 EXAONE-4.0 (32B), EXAONE-3.5 (32B), lama-3.1-Korean-8B-Instruct, Meta-Llama-3.8B 네 가지 대규모 언어 모델에 대한 셀프베리피케이션 기반 검증 절차 적용 전후의 핵심 정보 추출 정확도를 비교하였다. 셀프베리피케이션 검증 절차 적용 이전 전체 F1 점수를 기준으로 EXAONE-32B가 89.79%로 가장 높은 정확도를 보였고, EXAONE-4.0-32B가 82.62%로 그 뒤를 이었다. EXAONE은 이중 언어 모델으로서 영어뿐만 아니라 한국어에서도 긴 맥락 이해도 성능이 최고 수준[19]으로 홍보된 만큼 본 연구에서 사용된 한국어 프롬프트를 따르는데 큰 문제가 없었던 것으로 보인다. 반면 Llama-3.1-Korean-8B-Instruct와 Meta-Llama-3.8B의 전체 F1 점수는 각각 55.21%와 77.81% 수준으로 핵심 정보를 추출해내기 위해 제공된 프롬프트의 지시와 규칙을 이해하기 위한 어려움이 있었던 것으로 추측된다. 셀프베리피케이션 기반 검증 절차를 적용한 이후에는 EXAONE 계열 두 대규모 언어 모델에서는 정확도가 소폭 향상되었다. EXAONE-4.0 (32B)는 F1 점수가 3.62% 상승하였고, EXAONE-3.5 (32B)는 89.79%에서 91.25%로 약 1.46% 개선되었다.

<Table 3> Self Verification 검증 절차 적용 후의 F1 점수

핵심 정보 필드	EXAONE-4.0– 32B	EXAONE-3.5– 32B	Llama-3.1-Korean-8B-Instruct	Meta-Llama-3.8B-Instruct
대출일시	93.33%	92.78%	16.51%	42.86%
사용기간	96.48%	100.00%	57.93%	85.88%
대출자	90.82%	93.40%	74.59%	73.20%
담보물건지 주소	97.00%	96.00%	78,59%	85.86%
대출금액	98.00%	98.49%	90.43%	83.84%
대출금액(숫자)	98.00%	89.45%	67.38%	82.72%
대출수수료	99.50%	98.99%	90.32%	97.46%
투자자	97.98%	82.65%	51.76%	45.24%
금액	70.65%	86.29%	55.56%	47.06%
금액(숫자)	54.55%	74.11%	57.92%	45.65%
전체 정확도	86.24%	91.25%	66.67%%	70.45%

<Table 4> Self Verification 검증 절차 적용 전의 F1 점수

핵심 정보 필드	EXAONE–4.0-32B	EXAONE-3.5– 32B	Llama-3.1-Korean-8B-Instruct	Meta-Llama-3.8B-Instruct
대출일시	8.21%	96.91%	10.00%	9.36%
사용기간	96.45%	97.98%	89,01%	95.38%
대출자	88.78%	89.34%	84.42%	74.61%
담보물건지 주소	95.00%	91.00%	86.43%	84.42%
대출금액	98.00%	98.00%	97.00%	97.49%
대출금액(숫자)	99.50%	96.97%	82.83%	90.91%
대출수수료	98.49%	94.95%	97.98%	97.46%
투자자	76.53%	73.10%	67.35%	69.07%
금액	88.21%	85.86%	67.35%	87,76%
금액(숫자)	75.51%	73.74%	85.86%	62.24%
전체 정확도	82.62%	89.79%	55.21%	77.81%

필드 단위의 결과를 종합하면, 네 모델 모두 대출수수료와 같이 구조적 단서가 비교적 명확한 필드에서는 셀프베리피케이션 단계 적용 전후를 통틀어 90% 내외의 높은 정확도를 유지하였다. 반면 투자자, 금액, 금액(숫자)와 같이 메시지 안에 다수의 핵심 정보 값이 포함되거나 표기 방식이 다양한 핵심 정보의 추출에서는 Llama 계열 모델이 전반적으로 낮은 성능을 보였으며, EXAONE-4.0 (32B) 모델은 금액과 금액(숫자)의 핵심 정보를 추출하는데 있어 오히려 정확도가 낮아진 것으로 확인되었다. 이로써 네 종류의 대규모 언어 모델 중에서는 EXAONE-3.5 (32B) 모델이 셀프베리피케이션 검증 절차에 가장 적합한 모델로 확인되었다.

한편, 셀프베리피케이션 절차가 핵심 정보 추출 과정에 추가됨에 따라 후처리를 위한 대규모 언어 모델 호출 횟수가 늘어나 전체 처리 시간은 다소 증가하는 것으로 확인되었다. 그럼에도 동일 분량의 비정형 텍스트 메시지를 수사관이 수작업으로 열람·분류하고 범죄일람표를 작성하기 위해 소요되는 시간을 고려하면, 이러한 절차는 높은 수준의 효율성을 제공한다. 나아가 전처리가 필요한 비정형 텍스트 데이터의 규모가 커질수록 동일 인력과 시간으로 처리할 수 있는 메시지의 양이 수작업에 비해 크게 증가하므로 상대적인 효율성은 더욱 두드러진다. 특히 셀프베리피케이션 절차를 통해 오탐(False Positive)과 누락을 사전에 줄임으로써 잘못된 핵심 정보를 바탕으로 범죄일람표를 수정해야 하는 상황을 예방할 수 있다. 따라서 핵심 정보 추출에 소요되는 처리 시간의 증가는 추출 결과의 신뢰도를 높일 수 있다는 이익이 더 크다는 점에서 실무 적용 측면의 트레이드오프를 긍정적으로 평가할 수 있다.

Ⅵ. 결론

6.1. 결론

본 연구는 사이버사기 수사 과정에서 확보된 비정형 텍스트 메시지로부터 범죄일람표 작성에 필수적인 핵심 정보를 온프레미스 대규모 언어 모델을 통해 자동으로 추출하고 검증하는 단계를 제안하고 그 성능을 검증하였다. 먼저 랭익스트랙트의 스키마 및 퓨샷(Few-shot) 기반 구조화 추출 방식을 참조하여 대출일시, 대출자, 담보물건지 주소, 대출금액 등 10개 항목을 선정하고, 각 항목의 의미와 허용 표현, 정규화 규칙을 프롬프트상에 명확히 정의하였다. 이를 기반으로 EXAONE-4.0-32B, EXAONE-3.5-32B, Llama-3.1-K, Llama-3.8B 등 다양한 대규모 언어 모델에 동일한 작업을 수행하게 함으로써, 한글로 작성된 비정형 텍스트로부터의 모델별 핵심 정보 추출 정확도를 분석하였다. 나아가 기존 Gero et al.(2023)과 Polak & Morgan(2024)의 연구를 사이버사기 수사 도메인에 최적화하여 오미션, 에비던스, 룰-프룬, 셀프베리피케이션-프룬으로 이어지는 다단계 검증 절차를 설계하였다. 이 과정에서 목록형 항목의 누락을 방지하고, 추출된 값의 근거가 되는 원문 텍스트를 추적하며, 규칙 기반의 정제와 원문, 예측값 그리고 근거를 종합적으로 고려한 최종 교정을 통해 추출 결과의 신뢰성을 높였다.

실험 결과, 셀프베리피케이션 절차를 적용했을 때 대부분의 모델에서 F1 점수가 상승했음을 확인하였다. EXAONE-4.0-32B는 82.62%에서 86.24%로, EXAONE-3.5-32B는 89.79%에서 91.25%로 성능이 향상되었으며, Llama-3.1-Korean-8B-Instruct 모델 또한 뚜렷한 성능 개선을 보였다. 특히 대출일시, 투자자, 금액(숫자)과 같이 문맥 의존도가 높고 오탐지가 잦았던 핵심 정보 탐지가 어려운 항목에서 셀프베리피케이션 적용 후 성능이 개선된 점은 주목할 만하다. 본 연구에서 제안된 프로그램은 작업대출 사기라는 특정 시나리오에서의 범죄일람표 작성을 자동화하는 것을 목표로 설계되었으나, 특정 유형의 범죄에 한정되지는 않는다. 프롬프트에 포함된 지시사항과 규칙을 다른 범죄 종류의 데이터 환경에 맞게 수정하여 사용할 수 있는 유연

성을 지닌다는 점이 큰 장점이다. 다음은 셀프베리피케이션 절차 적용 전·후 정보 추출 결과의 차이를 예시적으로 보여주기 위해 작성된 도표이다.

<Table 5> 셀프베리피케이션 검증 절차 도입 전 추출 결과의 예시

추출 대상 텍스트	대출자	대출금액	대출금액(숫자)	투자자
매수인 김천재(에아) 전북 00광역시 00구 00동 **** 아파트. 소유권이전일 2022. 2. 8 담보대출대부 모집건 모집금액 10억 금리:월1.4% 연 14% 10/14일 전액 상환 예정입니다. 1.kim dong dong/김동동님 1.4억, 2. 달빛천사님 3000 ,2천만 라즐로님 2000 취소입니다~	김천자	1.4억, 2000	14000000, 3000	김동동, 달빛천사님, 라즐로

<Table 6> 셀프베리피케이션 검증 절차 도입 후 추출 결과의 예시

추출 대상 텍스트	대출자	대출금액	대출금액(숫자)	투자자
매수인 김천재(에아) 전북 00광역시 00구 00동 **** 아파트. 소유권이전일 2022. 2. 8 담보대출대부 모집건 모집금액 10억 금리:월1.4% 연 14% 10/14일 전액 상환 예정입니다. 1.kim dong dong/김동동님 1.4억, 2. 달빛천사님 3000 ,2천만 라즐로님 2000 취소입니다~	김천자 (에아)	1.4억, 3000, 2000	140000000, 30000000, 20000000	kim dong dong/김동동, 달빛천사, 라즐로

<Table 5>와 <Table 6>을 비교하면 대규모 언어 모델이 최초 추출 단계에서 누락한 대출자 성명 뒤의 별칭, 대출금액, 투자자 항목이 보완되었고, 숫자로 표기된 대출금액의 단위가 일관되게 교정되는 양상을 확인할 수 있다. 이는 셀프베리피케이션 절차가 대규모 언어 모델의 1차 추출 과정에서 발생하는 오류를 일정 부분 보완할 수 있음을 보여주며, 향후 성능이 개선된 온프레미스 대규모 언어 모델이 출시될수록 그 효과가 더욱 커질 것으로 예상된다.

6.2. 연구의 한계 및 문제점

본 연구는 데이터의 규모와 다양성 측면에서 일정 부분 한계를 지닌다. 실험에 활용된 데이터셋은 실제 사건을 기반으로 한 가상 데이터로, 특정 유형의 작업대출 사기와 유사한 맥락에 편중되어 있어 팀미션 사기나 주식리딩방 등 다양한 신종 범죄 유형이나 SMS, 이메일 등 타 플랫폼 데이터에 대한 일반화 성능을 충분히 검증하지 못하였다.

실험 설계 측면에서는 EXAONE 및 Llama 계열의 특정 모델과 고정된 프롬프트 조합만을 대상으로 분석이 수행되었다는 제약이 있다. 최신 모델이나 다양한 규모의 한국어 특화 모델들과의 광범위한 비교가 부재하여, 제안된 셀프베리피케이션 단계가 상이한 모델 환경에서도 일관된 성능 향상을 보장하는지에 대한 추가 검증이 필요하다. 아울러 본 연구는 F1 점수 중심의 정확도 평가에 집중한 나머지, 실제 수사 현장에서 필수적인 처리 속도, GPU 자원 효율성, 병렬 처리 능력 등 시스템 차원의 성능 요구사항을 구체적으로 다루지 못한 아쉬움이 있다.

이와 더불어, 실제 시스템을 운용할 수사관과의 상호작용 및 사용성 평가가 부재했다는 점도 한계로 볼 수 있다. 수사관이 선호하는 인터페이스 형태나 적절한 자동화 수준, 결과에 대한 신뢰도 등을 정성적·정량적으로 분석하지 못해 실무적 효용성을 명확히 평가하기 어렵다.

6.3. 후속 연구 제언

앞서 언급한 한계를 극복하고 본 연구의 성과를 발전시키기 위해, 향후 연구에서는 우선적으로 데이터의 다양성과 규모를 확장하여 시스템의 일반화 성능을 검증해야 한다. 작업대출 사기 외에 메신저 피싱, 팀미션 사기 등 다양한 사이버범죄 유형을 포괄하는 멀티데이터셋을 구축하고, 이를 통해 제안 기법이 범죄 유형이나 표현 방식의 변화에 대응하여 추출 정확도를 유지할 수 있는지 체계적으로 분석할 필요가 있다.

이와 함께, 텍스트에 국한된 분석 대상을 멀티모달 데이터로 확장하는 시도가 요구된다. 실제 수사 과정에서 수집되는 계좌 캡처 이미지, 통화 녹음, 계약서 파일 등 다양한 형태의 증거를 텍스트 데이터와 결합하여, 이미지 내 정보와 대화 내용을 교차 검증하거나 데이터 간의 정합성을 확인하는 고도화된 통합 분석 모델로 발전시켜야 한다. 또한, 실제 수사 환경에서의 효용성을 높이기 위해 인간 수사관과 대규모 언어 모델 간의 협업 구조를 구체화하는 연구도 필수적이다. 추출 결과와 근거를 직관적으로 시각화하고 수사관이 이를 손쉽게 검토·수정할 수 있는 그래픽 상호작용적 그래픽 사용자 인터페이스 (Graphical User Interface)를 설계함으로써, AI의 자동화 능력과 전문가의 판단이 상호 보완되는 시스템을 구축해야 한다.

기술적인 측면에서는 셀프베리피케이션 단계의 구조 자체를 고도화하는 방향으로 연구가 진행되어야 한다. 단순한 순차적 검증을 넘어 이종 모델 간의 교차 검증이나 신뢰도 기반 가중치 적용, 규칙 기반 점수와 대규모 언어 모델의 예측 신뢰도를 결합하여 하나의 최종 판단 점수를 산출하는 하이브리드 스코어링 등을 도입할 수 있다. 마지막으로, 다양한 모델 크기와 학습 데이터, 프롬프트 설정에 대한 체계적인 비교 분석을 수행하여, 자원과 보안 제약이 있는 수사 환경에 가장 적합한 실용적인 온프레미스 대규모 언어 모델 조합과 운영 기준을 도출하는 후속 연구가 이어져야 할 것이다. 아울러 자동화된 추출 절차의 최적화, 병렬 처리 전략, 하드웨어 자원 활용 방안 등을 검토함으로써 개별 사건 단위의 정보 추출 시간을 단축하는 방법을 모색할 필요가 있다. 또한 현재보다 훨씬 더 많은 양의 메시지 데이터를 한 번에 처리할 수 있도록 배치 구성과 메모리 관리 전략을 체계적으로 실험·비교하는 연구도 추가로 수행될 필요가 있다.

참고문헌 (References)

[1] Yoon BH, Park H, Kim J. 2025. On-premise llm-based key information extraction from unstructured text for cybercrime investigation support. 2025 Korea Artificial Intelligence Conference, Jeju, pp. 170-171.

[2] Lee CJ. 2025. "Evolving cyber fraud, you are also a target": Bundang Police Station urges prevention of new types of crimes ["진화하는 사이버사기, 당신도 표적입니다." 분당경찰서, 신종 범죄 피해 예방 당부]. Daehan Today. Available at: https://www.dhtoday.com/news/articleView.html?idxno=160473 accessed on 2025. 12. 10.

[3] Kim BS. 2025. Are you aware of 'team mission scams'? A new type of emerging fraud [신종사기 '팀미션 사기'를 아시나요?]. Jangheung Today. Available at: http://www.jhtoday.net/news/articleView.html?idxno=16536 accessed on 2025. 12. 10.

[4] SJKP Law Firm LLP. [n.d.]. Stock investment fraud: 'Stock reading room' tactics, victim relief, and reporting methods [주식투자사기 '주식리딩방' 수법과 피해 구제 및 신고 방법]. SJKP Law Firm LLP. Available at: https://www.daeryunlaw-finance.com/lawInfo_new/2371 accessed on 2025. 12. 10.

[5] Bang SH. 2025. Virtual asset fraud losses hit 18 trillion won last year: "Romance scams surging" [가상자산 사기 피해액 작년 18조원…"로맨스 스캠 급증"]. Market in. Available at: https://marketin.edaily.co.kr/News/ReadE?newsId=02810966642070520 accessed on 2025. 12. 10.

[6] Park MJ. 2024. Telegram crimes rampant, while cyber investigation units struggle with personnel shortages ['텔레그램 범죄' 난리치는데…사이버수사 인력난 '허덕']. Seoul Economic Daily. Available at: https://v.daum.net/v/20241114204615293 accessed on 2025. 12. 10.

[7] Polak MP, Morgan D. 2024. Extracting accurate materials data from research papers with conversational language models and prompt engineering. Nature Communications, 15, 1569. https://doi.org/10.1038/s41467-024-45914-8

[8] Hicks MT, Humphries J, Slater J. 2024. ChatGPT is bullshit. Ethics and Information Technology, 26, 38. https://doi.org/10.1007/s10676-024-09775-5

[9] Huh SC. 2025. Criminal investigation using big data and the protection of privacy. Public Law, 54(1), 347-376. https://doi.org/10.38176/PublicLaw.2025.10.54.1.347

[10] Manyika J, Chui M, Brown B, et al. 2011. Big data: The next frontier for innovation, competition, and productivity. McKinsey & Company. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/big-data-the-next-frontier-for-innovation.pdf

[11] Kim J, Woo BK. 2022. A study on cybercrime detection and analysis technology. Journal of Korean Public Police and Security Studies, 19(3), 57-76. http://doi.org/10.25023/kapsa.19.3.202208.57

[12] Kim SJ. 2017. A study on construction of crime prevention system using big data in Korea, The Journal of The Institute of Internet, Broadcasting and Communication, 17(5), 217-221. http://doi.org/10.7236/JIIBC.2017.17.5.217

[13] Kim J. 2023. Development of an integrated analysis and inference system for cybercrime investigative leads [사이버범죄 수사단서 통합분석 및 추론시스템 개발]. Korea Institute of Science and Technology Evaluation and Planning, Chungcheongbuk-do, Korea. TRKO202400000786.

[14] Bergmann D. [n.d.]. What is fine-tuning? IBM. Available at: https://www.ibm.com/think/topics/fine-tuning accessed on 2025. 12. 10.

[15] Farquhar S, Kossen J, Kuhn L, et al. 2024. Detecting hallucinations in large language models using semantic entropy. Nature, 630, 625-630. https://doi.org/10.1038/s41586-024-07421-0

[16] Gero Z, Singh C, Cheng H, et al. 2023. Self-verification improves few-shot clinical information extraction. arXiv:2306.00024. https://doi.org/10.48550/arXiv.2306.00024

[17] Prompt Engineering Guide. [n.d.]. Zero-shot prompting. Prompt Engineering Guide. Available at: https://www.promptingguide.ai/techniques/zeroshot accessed on 2025. 12. 10.

[18] Park JH, Ahn MI, Kang WS, et al. 2019. Comparative analysis on cloud and on-premises environments for high-resolution agricultural climate data processing. Korean Journal of Agricultural and Forest Meteorology, 21(4), 347-357. https://doi.org/10.5532/KJAFM.2019.21.4.347

[19] LG AI Research. 2024. Release of three open-source EXAONE 3.5 models: Frontier AI-class models achieving state-of-the-art performance in instruction following and long context [EXAONE 3.5 3개 모델 오픈소스로 공개 - Frontier AI급의 모델, Instruction Following 및 Long Context 최고 수준 성능 달성]. Available at: https://www.lgresearch.ai/blog/view?seq=506 accessed on 2025. 12. 10.