사이버범죄 수사지원 특화 대규모언어모델 개발을 위한 벤치마크 구축 및 보상체계 설계

계효선, 이미라, 김지형*

한국전자기술연구원

hskye@keti.re.kr, mlzr@keti.re.kr, *jkim8@keti.re.kr

Designing a Benchmark and Reward Framework for Developing Specialized Large Language Models for Cybercrime Investigation

Hyoseon Kye, Mira Lee, Jeehyeong Kim*

Korea Electronics Technology Institute

요 약

본 논문은 사이버범죄 수사 추론에 특화된 대규모언어모델(Large Language Model, LLM) 학습을 위한 보상 시뮬레이션 프레임워크를 제안한다. LLM 강화 학습 기법인 Group Relative Policy Optimization(GRPO)과 같은 방법을 활용하기 위해서는, 특수 목적 모델 튜닝에 적합한 보상함수 설계가 필수적이다. 이를 위해 본 연구에서는 보상함수의 효과를 실험적으로 검증할 수 있는 시뮬레이션 프레임워크를 설계·구현하고, 사이버범죄 수사 추론 관련 벤치마크를 구축하였다. 본 프레임워크를 통해 다양한 LLM이 해당 벤치마크와 보상체계 하에서 어떠한 성능과 경향성을 보이는지를 분석하였다.

Ⅰ. 서 론

최근 주식 리딩방, 인터넷 물품 사기, 몸캠피싱 등으로 대표되는 사이버범죄는 빠르게 고도화·조직화되고 있으며, 이에 따라 수사기관의 대응 난이도도 지속적으로 상승하고 있다. 이러한 환경에서는 범용 언어모델만으로는 수사 현장에서 요구되는 맥락 이해와 추론을 충분히 수행하기 어려우므로, 사이버범죄 수사 추론에 특화된 대규모 언어모델의 개발 필요성이 커지고 있다 [1]. 최신 API 기반 초거대 모델의 활용은 개발 난도를 낮추고 경우에 따라 파인튜닝을 생략할 수 있다는 장점이 있으나, 실제 수사 환경에서는 보안상의 이유로 외부 API 사용이 제한되는 경우가 많고, 제한된 GPU 자원에서 구축형 모델을 운용해야 하는 제약이 상존한다. 특히 이 분야는 선행 연구가 많지 않아, 제한된 자원으로도 실무에 필요한 수준의 성능을 달성할 수 있는지 불확실성이 크다. 이러한 문제를 해결하기 위해서는 기존 파운데이션 모델을 기반으로 도메인 지식을 반영한 강화 학습을 수행할 필요가 있으며, 그 핵심은 적절한 보상함수의 설계와 이를 실험적으로 검증할 수 있는 시뮬레이션 프레임워크의 마련에 있다. 본 연구는 사이버범죄 수사 추론 특화 Large Language Model(LLM) 학습을 위해 보상 시뮬레이션 프레임워크와 벤치마크를 함께 제안하고, 제한된 환경에서도 특화 모델 개발의 방향성과 실현 가능성을 체계적으로 검증할 수 있는 기반을 제공하고자 한다.

Ⅱ. 본론

도메인 특화 LLM을 구현하기 위해 단순 지도학습만을 적용하는 것은 한계가 분명하다. 사이버범죄 수사 추론은 복잡한 의사결정과 맥락 통합이 요구되는 영역으로, 정답 예시를 반복 학습하는 방식만으로는 실제 수사관의 추론 과정을 충실히 재현하기 어렵다. 이 한계를 보완하기 위해 인적 피드백을 이용한 강화 학습이 활발히 논의되어 왔으나, 전통적인 Proximal Policy Optimization(PPO) 기반 접근은 대규모 자원과 많은 피드백 데이터를 필요로 한다는 점에서 구축형 환경에는 부담이 크다. 이에 본 연구는 여러 응답 후보를 그룹 단위로 비교해 상대적 우위를 학습하는 방식의 Group Relative Policy Optimization(GRPO)[2]를 채택하였다. GRPO는 정답이 단일하지 않고 다양한 경로의 합리적 추론이 가능한 과제를 다루기에 적합하며, 절대 점수 대신 응답들 간의 상대적 선호를 학습함으로써 효율적인 미세조정이 가능하다는 장점이 있다. 또한, 보상함수 시뮬레이션의 효과적 구현을 위해 실제 수사 시나리오를 반영한 벤치마크를 구축하였다. 이 벤치마크는 총 여덟 개의 주요 문항으로 구성되며, 각 문항에는 다시 상황파악, 주요정보파악, 다음 수사 프로세스 추천의 세 가지 항목에 대해 각기 일반 응답과 전문가 응답을 제시하여 구성된다. 문항은 사기 수법의 식별과 패턴화, 계좌 명의자와 실제 사용자의 분리 판단, 피해자 진술과 기술 단서·연계 사건의 통합 정리, 가용 증거에 근거한 다음 절차의 우선순위 도출, 통화 및 거래내역 기반의 조직 구조 추론, IP 패턴과 현금 인출 및 현지 결제 내역을 활용한 위치 추정, 그리고 메신저·클라우드 접속 기록을 이용한 공범 식별 등으로 구성된다. 이러한 구성은 실제 수사관이 수행하는 상황 파악, 핵심 정보 정리, 그리고 실행 가능한 대응 제안이라는 세 가지 과업 유형을 폭넓게 포괄한다. 이러한 추론을 수행할 수 있도록 8개의 질의에 대해 별도의 컨텍스트를 포함한다. 컨텍스트는 10문장 수준의 텍스트로써, 본 추론에 활용할 수 있는 요약된 수사 정보를 의미한다. 추후 실제적인 사이버범죄 수사추론 시스템 개발을 위해서는 다양한 곳에서 데이터나 분석 결과를 조회하여 양질의 컨텍스트를 만드는 것이 성능을 향상시키는 주요 요소중 하나가 될 수 있으나, 현재 단계에서는 컨텍스트를 활용하여 모델이 얼마나 추론을 잘 할 수 있는지를 평가하는 것이 중요하므로, 이러한 컨텍스트는 충분히 양질의 컨텍스트가 사전에 생성되었다고 가정한다.

보상함수는 의미적 정합성과 내용 충실성, 그리고 언어적 완결성을 동시에 반영하도록 설계되었다. 먼저 모델 응답이 전문가 답변과 의미적으로

얼마나 근접한지를 평가하여, 단순 표현의 유사성을 넘어 추론의 방향성과 결론의 타당성을 측정한다. 다음으로 사건 분석에 필수적인 키워드의 반영 정도를 점검하여, 모델이 핵심 개념과 단서를 충분히 포괄하는지를 확인한다. 키워드는 데이터셋의 모범응답을 기반으로 주요단어를 뽑아 선정되어있다. 응답에서 해당 키워드를 얼마나 많이 포함하는지를 계산하여 점수화하는 방식이다. 마지막으로 응답의 길이와 구성, 한국어 사용의 적정성, 자연스러운 종결 등 문장 수준의 품질을 평가하여, 보고서 형식으로 바로 활용 가능한 응답을 유도한다. 사이버범죄 수사의 특수성을 고려해 최소 길이와 최적 범위를 상향 조정하고 한국어 비중 기준을 강화하는 등 도메인 맞춤형 기준을 적용하였다.

표1에서는 각 모델의 벤치마크 결과를 나타내고 있다. 전문가 응답을 정답의 기준으로 설정하고, gpt-4o, gpt-4o-mini, gpt-oss-20b[3]의 성능을 동일 벤치마크에서 비교하였다. gpt-4o와 gpt-4o-mini는 API 기반 LLM의 가장 대표적인 서비스로써 성능의 기준을 삼기 위해 선정하였다. gpt-oss-20b 는 OpenAI에서 최근에 출시한 오픈소스 Smaller Large Language Model (sLLM)이다. 특히, gpt-oss-20b는 4.5bit 양자화 튜닝으로 인해 24GB 이하의 게이밍 GPU에서도 구동이 가능한 것이 장점으로, 향후 사이버범죄 수사추론 특화 모델을 개발할 시에 적극적으로 고려될 수 있는 수준의 모델이다.

모델들의 성능을 평가하기 위해서는 벤치마크를 기반으로 보상 시뮬레이션을 구동하였을 때의 결과가 정답 데이터의 보상 시뮬레이션 결과가 얼마나 차이가 나는지를 비교해야 한다. 각 LLM 모델이 추론에 사용한 프롬프트는 상황파악, 주요정보, 다음 절차 추천으로 구분하여 구성하였으며, 핵심정보를 요약하고 이해하고 정리하여 연관성을 파악하라는 내용이 포함되어 있다. 또한 벤치마크에 포함된 컨텍스트를 프롬프트에 포함하여 LLM 모델이 활용할 수 있도록 하였다. 종합 성능에서 gpt-4o가 가장 정답에 근접했으나, 유사 수준에는 못 미치는 결과가 나왔다. 정답의 평균점수가 0.969점 인데 반해 gpt-4o의 평균점수는 0.807이었다. gpt-4o-mini는 전반적으로 gpt-4o와 큰 차이가 나지 않았다. 대부분의 모델 성능 벤치마크에서 gpt-4o이 gpt-4o-mini 보다 성능이 좋을 수 밖에 없는데도 해당 벤치마크에서 큰 차이가 나지 않는 것은 본 벤치마크가 특별히 어렵고 복잡한 역할을 LLM에게 기대하기보다 도메인 특화된 요구사항을 기대하는 것으로 해석할 수 있다. 그럼에도 불구하고, gpt-oss-20b는 다른 두 모델에 비해 현저한 성능저하가 관찰 되었다. 평균점수는 0.697 수준이었으며, 오차의 표준편차도 0.089로 다른 모델에 비해 상대적으로 일관성도 떨어졌다. gpt-4o-mini 정도의 성능만 되어도 크게 무리가 없는 수준의 복잡도를 지닌 태스크에 대해서도, 모델 성능의 한계로 인해 gpt-oss:20b는 현저한 성능차이를 보이고 있다.

표2 는 각 모델들이 세 가지 채점 항목에서 각 어떤 점수를 보이고 있는지를 나타내고 있다. 키워드 탐색에서 세 모델 다 정답 대비 적절한 성능을 보이지 못하는 것을 알 수 있다. 이는 사이버범죄 수사 도메인 특유의 단어와 표현방법을 모델들이 잘 쓰지 않기 때문에 나오는 결과로 보인다. 정답과의 유사도 점수는 gpt-4o와 gpt-4o-mini가 비슷했고, 문장 구조에 대해서는 gpt-4o가 gpt-4o-mini보다 다소 나은 결과를 보였다. 문장을 만드는 기본 능력에 있어서 gpt-4o가 다소 나은 성능을 보였다고 해석할 수 있다. gpt-4o-oss-20b의 경우, 유사도나 문장 구조에 비해 오히려 키워드 항목에서는 다른 두 모델에 비해 상대적으로 적은 차이를 보였는데, 이는 학습되지 않은 도메인 특유의 표현을 하는 것에 있어서는 다른 대규모 모델들과 큰 차이가 없었음을 의미한다. gpt-oss-20b는 단순 사실 확인형 과제에서는 준수한 결과를 보였지만, 분산된 단서를 결합해 결론을 도출하는 고난도 종합 추론 과제에서는 성능 저하가 두드러졌다.

<표1. 전문가 응답 및 각 LLM 모델들의 벤치마크 평균점수>

지표 \ 모델	전문가 응답	gpt-4o	gpt-4o-mini	gpt-oss:20b
평균점수	0.969	0.807	0.800	0.697
오차	-	0.162	0.169	0.272
오차 표준편차	-	0.043	0.047	0.089

<표2. 전문가 응답 및 각 LLM 모델들의 벤치마크 항목 별 점수>

항목 \ 모델	전문가 응답	gpt-4o	gpt-4o-mini	gpt-oss:20b
유사도	1.00	0.862	0.852	0.750
키워드	1.00	0.633	0.653	0.613
문장 구조	0.90	0.899	0.868	0.702

Ⅲ. 결론

본 연구는 사이버범죄 수사 추론에 특화된 LLM 개발을 목표로, 보상 시뮬레이션 프레임워크와 실제 수사 시나리오를 반영한 벤치마크를 제안하고 그 타당성을 실험적으로 확인하였다. 분석 결과는 상용 모델이 문장 수준의 품질에서는 높은 완성도를 보이지만, 정답에 상응하는 논리적 추론과 핵심 정보의 포괄성에서 여전히 격차가 존재함을 보여준다. 이에 따라 향후 연구에서는 벤치마크의 문항 범위와 난이도를 확대하고 항목별 평가 기준을 정교화하여, 다양한 유형의 수사 추론을 안정적으로 평가할 수 있는 체계를 구축할 것이다. 동시에 제안한 시뮬레이션을 학습 루프의 중심에 두고 도메인 피드백을 체계적으로 반영함으로써, 제한된 자원 환경에서도 일관되게 동작하는 수사 특화 모델을 단계적으로 구현할 계획이다. 아울러 실무 적용성을 높이기 위해, 수사에 필요한 정보를 능동적으로 수집·정제하는 도구 활용과 절차적 플래닝 역량을 모델 수준과 시스템 수준에서 통합하는 별도의 연구를 병행할 필요가 있다. 이를 통해 모델의 추론 능력뿐 아니라 실제 업무 흐름에서의 효율성과 신뢰성을 함께 제고할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

이 논문은 2025년도 행정안전부의 재원으로 과학치안진흥센터 사이버범죄 수사단서 통합분석 및 추론시스템 개발 사업의 지원을 받아 수행된 연구임(No. RS-2025-02218280)

참 고 문 헌

[1] J. Kim, et al., “A Conversational AI Agent System Based on Large Language Models and Retrieval-Augmented Generation to Support Investigation,” Journal of Data Forensics Research, vol. 1, no. 1, pp. 47-62, 2024.

[2] Z. Shao, et al., “DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models,” arXiv preprint arXiv:2402.03300, 2024.

[3] OpenAI, et al., “gpt-oss-120b & gpt-oss-20b Model Card,” arXiv preprint arXiv:2508.10925, 2025.