Journal of Digital Policy
Vol. 2. No. 4, pp. 1-10, 2023
ISSN 2951-245X
https://doi.org/10.23149/JDP.2023.2.4.001

ChatGPT의 경찰 관련 교통법규 응답 능력에 대한 탐색적 연구 - 운전면허 학과시험과 도로교통사고감정사 1차 시험을 대상으로 -

이상엽
경찰대학교 경찰학과 교수

An Exploratory Study on ChatGPT's Performance to Answer to Police-related Traffic Laws: Using the Driver's License Test and the Road Traffic Accident Appraiser

Sang-yub Lee
Professor, Department of Police, Korean National Police University

요 약 본 연구는 경찰교통에서의 효과적 ChatGPT 활용 방안 도출을 위한 사전 연구로서 운전면허 학과시험과 도로교통사고감정사 시험에 대한 ChatGPT의 응답을 분석하였다. ChatGPT가 뛰어난 성능과 접근성으로 여러 분야에서 기대를 받고 있으나 경찰 교통법규와 같이 고도의 정확성이 요구되는 분야에서는 사전에 그 성능과 한계를 탐색할 필요가 있다. 이에 본 연구에서는 운전면허 학과시험 문제은행과 도로교통사고감정사 1차 시험을 대상으로 파이썬 코드로 OpenAI API를 이용해 30회의 반복 실험으로 ChatGPT의 응답을 수집하고 응답 결과를 바탕으로 시험별·연도별·내용 영역별 정답률, 일관성 능력을 분석하였다. 분석 결과 첫째, 운전면허 학과시험 및 도로교통사고감정사 1차 시험의 평균 정답률은 각 44.60%, 35.45%로 합격기준보다 낮았다. 연도별로는 2022년 이후 정답률이 평균 정답률을 하회했다. 둘째, 영역별 정답률은 29.69%~56.80%로 나타나 큰 편차를 보였다. 셋째, 정답을 맞힌 경우 95% 이상 일관되게 같은 응답을 출력하였다. ChatGPT의 효과적 활용을 위해서는 사용자의 전문 지식, 평가 데이터 및 방법 마련, 양질의 교통법규 말뭉치 설계와 주기적 학습이 필요하다고 판단된다.

주제어 : 경찰교통, ChatGPT, 교통법규 질의응답, 운전면허 학과시험, 도로교통사고감정사 시험

Abstract This study conducted preliminary study to identify effective ways to use ChatGPT in traffic policing by analyzing ChatGPT's responses to the driver's license test and the road traffic accident appraiser test. I collected ChatGPT responses for the driver's license test item pool and the road traffic accident appraiser test using the OpenAI API with Python code for 30 iterative experiments, and analyzed the percentage of correct answers by test, year, section, and consistency. First, the average correct answer rate for the driver's license test and the for road traffic accident appraisers test was 44.60% and 35.45%, respectively, which was lower than the pass criteria, and the correct answer rate after 2022 was lower than the average correct answer rate. Second, the percentage of correct answers by section ranged from 29.69% to 56.80%, showing a significant difference. Third, it consistently produced the same response more than 95% of the time when the answer was correct. To effectively utilize ChatGPT, it is necessary to have user expertise, evaluation data and analysis methods, design a quality traffic law corpus and periodic learning.

Key Words : Traffic Policing, ChatGPT, Traffic-law QA, driver's license test, road traffic accident appraiser test

*Corresponding Author : Sang-yub Lee(yubii02@gmail.com)

Received October 27, 2023
Revised November 16, 2023
Accepted December 28, 2023
Published December 28, 2023

1. 서론

1.1 연구의 배경 및 목적

2022년 11월 OpenAI에서 개발한 ChatGPT가 공개되었다. 기존의 딥러닝 모델과는 달리 일반인도 자연어로 질의를 입력해 손쉽게 사용이 가능한 높은 접근성과 번역, 요약, 코딩, 창작 등 텍스트를 사용하는 광범위한 부분에서 강력한 성능을 보여줌으로써 2개월 만에 사용자 1억 명을 돌파하며 폭발적 관심을 받았다[1]. 자연스럽게 사회, 경제, 의료, 마케팅 등 많은 분야에서 생산성을 향상시킬 것이라는 기대가 이어지며[2,3] 대통령이 직접 언급할 정도로 주목을 받고 있다[4].

대용량 데이터, 연산능력 향상, 알고리즘 개발과 함께 인공지능이 급격히 발전하면서 경찰교통 분야에서는 그동안 교통사고 조사[5], 단속[6,7], 교통사고 예측[8.9] 등 주로 개발된 인공지능 모델들을 업무에 활용하는 방안을 제시하는 연구들이 활발히 이루어져 왔다. 이제 ChatGPT라는 새로운 모델이 공개되고 큰 반향을 일으킬 정도로 놀라운 성능과 우수한 접근성을 보여준 만큼 그 활용 방안을 고민할 시점으로 생각된다. ChatGPT의 기본 입력과 출력이 현재까지 텍스트인 점을 고려하면 대표적인 활용 방안으로 경찰 관련 교통법규 질의응답을 생각할 수 있다. 그러나 ChatGPT는 방대한 언어 데이터를 학습한 후 다음 단어를 예측하는 확률기반의 모델이므로[10,11] 응답이 정확한 사실인지 담보하기 어렵다. 문학, 미술처럼 창의적 작업에서는 이것이 장점이 될 수도 있으나 법률처럼 정확한 정보 제공이 필요한 영역에서는 활용에 앞서 그 성능과 한계를 사전에 탐색할 필요가 있다.

그러나 기존의 연구 동향을 분석한 결과, 경찰 관련 교통법규 질의응답 능력에 대한 실증적 근거를 제공한 연구는 아직 전례가 없는 것으로 판단되었다. 따라서 본 연구에서는 ChatGPT의 경찰 관련 교통법규 질의응답 능력을 실험해 성능과 한계를 탐색하고 향후 개선 가능성을 제시해보고자 한다. 이런 의미에서 본 연구는 경찰교통 분야에서 ChatGPT를 활용하기 위한 사전 연구로 볼 수 있다.

1.2 연구의 범위 및 방법

경찰 관련 교통법규로 도로교통법, 교통사고처리 특례법과 특정범죄 가중처벌 등에 관한 법률 중 도주차량, 위험운전치사상, 어린이보호구역 사고 조항[1]을 중심으로 살펴보았다. 그 이유는 자동차관리법, 자동차손해배상 보장법, 도로법이 국토교통부 소관인 것과 달리 도로교통법과 교통사고처리 특례법은 경찰청 소관 법률이기 때문이다. 특정범죄 가중처벌 등에 관한 법률은 비록 법무부 소관 법률이나 해당 조항은 교통사고와 관련된 것으로 도로교통법, 교통사고처리 특례법 위반 중 일부 행위에 대해 가중처벌만을 규정하고 있어 연구 범위에 포함했다. 질의응답 데이터로는 2018 ~ 2023년도 운전면허 학과시험 문제은행과 2018 ~ 2022년도 도로교통사고감정사 1차 시험 중 '교통관련 법규' 과목을 사용하였다. 운전면허 학과시험 문제은행은 일반 시민을 대상으로 하고 있어 일반인과 비교가 가능하고 문제은행 형식으로 풍부한 평가 문제를 제공하기 때문이다. 도로교통사고감정사 시험은 전문가 자격 인증을 위한 시험으로 운전면허 학과시험보다 난이도가 높은 문제에서 성능을 평가하기 위해 사용하였다.

파이썬 코드로 OpenAI에서 제공하는 API를 이용하는 함수를 만들어 시험별로 30회씩 질의응답을 반복해 실험하였다. 응답 결과를 바탕으로 먼저 시험별 · 연도별 정답률을 분석하였고, 내용 영역별로 정답률을 분석하였다. 마지막으로 응답의 일관성 여부를 평가하였다.

2. 이론적 배경

2.1 ChatGPT[2]

ChatGPT는 InsturctGPT와 같이 인간 피드백을 통한 강화학습(RLHF, Reinforcement Learning from Human Feedback)을 사용해 학습한 챗봇으로, 사용자가 입력한 문장을 이해하고 대화하는 것처럼 의사소통을 할 수 있다. InstructGPT는 GPT(Generative Pre-trained Transformer) 계열 모델에 RLHF 방법을 적용해 일종의 파인튜닝한 모델로 대화에 좀 더 적합한 모델이다. GPT는 OpenAI에서 개발한 자연어 생성 모델로 뉴스 기사, 백과사전, 웹사이트, 책, 소설, 논문, 블로그 등에서 수집된 대규모의 자연어 데이터에서 주어진 텍스트의 다음 단어를 예측하는 태스크를 학습함으로써,

1) 정확한 조문 명은 같은 법 제5조의3(도주차량 운전자의 가중처벌), 제5조의 11(위험운전 등 치사상), 제5조의 13 어린이 보호구역에서 어린이 치사상의 가중처벌)이다.

2) 현재 ChatGPT는 GPT-3.5와 GPT-4를 서비스하고 있으며 GPT-3.5는 누구나 무료로 사용이 가능하나, GPT-4는 한 번 이상 유료 결제를 한 사람만 사용 가능하며 웹페이지에서는 3시간당 25개의 메시지만 가능하다.

사람이 사용하는 자연어의 특징을 분석하여 텍스트를 생성할 수 있다[2]. RLHF 방법은 다음의 순서로 이루어진다. 1) 샘플링된 프롬프트에 대해 평가자가 원하는 아웃풋을 작성하고, 이 데이터로 지도학습 방식으로 기본 모델을 파인튜닝한다. 2) 프롬프트에 대해 여러 개의 아웃풋을 샘플링하고 평가자가 아웃풋 간에 상대적 비교를 통해 순위를 정한 후 이 데이터로 보상모델을 학습한다. 3) 샘플링된 프롬프트에서 아웃풋을 작성토록 하고 2단계에서 학습한 보상모델이 평가한 값을 이용해 PPO (proximal policy optimization)란 강화학습 알고리즘으로 1) 번의 모델을 업데이트한다[12]. OpenAI 블로그[13]에 따르면 ChatGPT의 학습방법도 위와 동일하나 대화 형태의 데이터셋을 위 1, 2단계에 적용한 차이가 있다고 한다.

2.2 선행 연구

시험 문제 등 질의응답 데이터셋을 활용해 ChatGPT의 성능을 평가하는 연구는 법률, 교육, 의료, 회계 등 다양한 분야에서 이루어지고 있다.

먼저 법률 분야는 상대적으로 해외에서 활발히 연구되고 있다. Bommarito II 등[14]은 gpt-3.5에게 미국 변호사 시험 출제기관인 NCBE(National Conference of Bar Examiner)가 출제한 미국 변호사 시험 객관식 모의고사를 치르게 한 결과, 50.3%의 정답률을 보여 실제 사람 응시자의 정답률인 68%보다는 낮은 수치를 기록했지만 증거(evidence)와 불법행위(torts) 영역에서 평균 합격률에 도달했다고 보고하였다. Chalkidis[15]은 법률 업무 관련 자연어 처리 벤치마크인 LexGLUE 데이터셋[3]을 이용해 gpt-3.5-turbo 모델의 성능을 평가하였다. 평균 micro-F1-score는 49.0%로 높지 않았으나 일부 하위 데이터셋에서는 70.1%의 micro-F1-score를 기록하였다. Choi 등[16]은 객관식과 주관식 문제로 구성된 미네소타 대학교 로스쿨 4개 법률 과목의 실제 시험에 대한 ChatGPT의 응답을 실제 학생들의 답안과 함께 채점해 비교하였다. 평균적으로 C+ 학생 수준의 성적을 얻어 점수는 낮았지만 네 과목 모두 합격점수를 통과하였다. 법률 분야 국내 연구로는 박성미 등[17]의 연구가 있다. gpt-4에게 2023년도 법학적성시험(LEET)의 '추리 논증' 영역 시험 문제를 풀게 한 결과, 표준 점수 49.4점으로 누적비율 89.7%에 해당하였다. 풀이 과정에서 상식추론 능력 부재, 판단기준 제시 부재 등의 문제점을 확인하고 아직 논증 타당성을 대체하기에는 부족하며 활용 가능성을 높이기 위해 법률 관련 사전학습, 전문가 피드백 시스템, 법 영역 특화기술 개발을 제안하였다.

국내에서는 교육 분야에서 관련 연구가 활발하며, 이외에도 의료, 회계 등 분야에서 관련 연구가 시작되는 것으로 보인다. 먼저 교육 분야를 살펴보면, 백미경[18]은 한국어 교육에서 교육자료로 ChatGPT를 활용할 수 있는지 살펴보기 위해 한국어 능력 수준과 한국어 오류에 대해 분석하였다. 한국어 능력 시험(TOPIK) 읽기 영역 기출문제 및 관용구[4] 목록을 이용해 평가한 결과, 한국어 능력 시험은 정답률이 약 80%를 보였으나 추론능력이 요구되는 문항에서 상대적으로 정답률이 낮았고 관용구도 약 56%의 정답률을 보여 ChatGPT의 한국어 능력을 3~4급 정도의 중급 수준으로 판단하였다. 권서경·이영태[11]는 ChatGPT의 영어 독해 문항 풀이 능력을 탐색하였다. 2021 ~ 2023학년도 대학수학능력시험 영어 읽기 문항과 ETS의 TOEFL iBT reading 영역 Practice Test 세트 3개에 대해 GPT-3.5, 4로 실험한 결과 GPT-3.5는 대학수학능력시험에서는 약 69%의 정답률을 보여 2등급 수준을, TOEFL iBT에서는 73%의 정답률을 보였으며 GPT-4는 대학수학능력시험에서는 약 93%의 정답률을 보여 1등급 수준을, TOEFL iBT에서는 93%의 정답률을 보였다. 권오남 등[2]은 수학교육에서 ChatGPT의 활용방안 도출을 위한 기초 연구로서 ChatGPT의 수학적 성능을 평가하였다. 2018 ~ 2020 국가수준 학업성취도 평가와 2021 ~ 2023 대학수학능력시험 수학 문제를 ChatGPT 3.5-turbo 모델에 입력하고 정답률, 풀이 과정 정확도, 오류 유형을 분석한 결과, 학업 성취도 평가는 정답률 37.1%, 풀이 과정 정확도 3.44, 절차적 오류 비중 71.01%, 기능적 오류[5] 비중 28.99%로 나타났으며 대학수학능력시험은 정답률 15.97, 풀이 과정 정확도 2.49, 절차적 오류 비중 47.15%, 기능적 오류 비중 52.85%로 차이를 보였다.

3) LexGLUE 데이터셋은 총 7개의 문서 분류 하위 태스크별로 데이터셋이 구성되어 있다. 데이터셋에 대한 세부적인 정보는 Hugging Face 사이트 (https://huggingface.co/datasets/lex_glue) 참조.

4) 두 개 이상의 단어로 이루어져 있으나 그 단어들의 의미만으로 전체의 의미를 알 수 없는 어구를 말한다. 예를 들어 '배가 아프다'는 '남이 잘되어 심술이 나다'는 의미이다.

5) 절차적 오류는 다음 단계로의 식을 연결 짓는 과정이나 계산상의 오류, 기능적 오류는 텍스트를 인식, 판단, 출력하는 과정에서 발생하는 오류를 뜻한다고 한다.

유재진[19]은 ChatGPT를 지리교육에 활용할 수 있는지 탐구하기 위해 개념 확인을 위한 지리 교과 관련 일반적 질문, 문제해결 능력 확인을 위한 심화 질문, 오개념 관련 질문과 추가적 질의응답을 실시하였다. 실험 결과, ChatGPT가 출력한 응답이 학습자의 의문을 해소할 수 있고 추가적인 질의를 통해 심화 문제도 해결이 가능한 것을 확인하고 교사는 교수·학습 모형, 방법 수립 및 교재 개발에, 학습자는 개별화 학습 보조 도구로 사용 가능함을 시사점으로 제시하였다.

의료 분야에서 허선[10]은 ChatGPT가 실제 증례에서 정확하게 진단을 내릴 수 있는지 확인하기 위해 무작위로 선별한 case report 논문 10편에 대해 진단과 치료 방법을 나열토록 하였다. 그 결과 증상, 소견, 과거력만으로는 10건 중 4건, 진단 검사 결과를 포함했을 때는 10건 7건에 대해 정확한 진단명을 제시하였으며 치료 방법은 10건 중 4건에 대해서만 모든 내용을 적절하게 제시하는 것으로 나타났다. 회계 분야에는 윤양인[20]이 ChatGPT의 회계 분야 성능을 평가하고 회계교육과 관련한 시사점을 모색하였다. 일반적인 회계 개념 평가와 거래에 대한 분개(회계처리) 수행능력 평가를 실시한 결과, 주요 개념 설명이나 간단한 거래에 대한 분개는 적절히 수행하였으나 복잡한 거래, 복수의 사실관계 종합이 필요한 상황에서의 회계처리에서는 한계를 보였다. 이를 바탕으로 회계교육 측면에서 기존 단답형 서술이나 단순 분개 수행 과제의 재검토, 종합적 회계 사고능력 평가 문항 개발, 부분 점수 부여에 대한 재검토를 시사점으로 제시하였다.

2.3 시사점

선행 연구를 검토한 결과 질의응답 데이터셋을 활용해 ChatGPT의 능력을 평가, 분석하는 연구가 활발히 이루어지고 있으나, 국내에서 법률 분야에 대한 연구는 상대적으로 부족하며 특히 교통 법규 관련한 연구는 전례가 없음을 확인할 수 있었다. 또한 ChatGPT가 매회 동일한 응답을 하지 않는다는 점을 고려했을 때 대부분 연구가 1회 실험 결과에 기초하고 있어 평가 결과의 신뢰성을 파악하기 어렵다는 한계가 있다. 본 연구는 교통 법규에 한정해 ChatGPT의 질의응답 능력을 평가하였으며, 시험별 30회 반복 실험을 통해 통계적 신뢰도를 확인하였다는 차별점이 있다.

3. 연구방법

3.1 연구 자료

3.1.1 운전면허 학과시험

운전면허 학과시험은 도로교통법 제83조에 의거해 도로교통공단이 실시하는 시험으로 자동차등 및 도로교통에 관한 법령에 대한 지식, 자동차등의 관리방법과 안전운전에 필요한 점검의 요령 등[6] 운전자가 필수적으로 알아야 할 사항들을 문답 형식으로 진행하는 필기시험이다[21]. 출제비율은 자동차등 및 도로교통에 관한 법령에 대한 지식 95%, 자동차등 관리방법과 안전운전에 필요한 점검의 요령 5%로 하며 출제유형은 문장형 4지선1형과 4지선2형, 사진형, 일러스트형, 안전표지형, 동영상형으로 구성된다[22]. 합격점수 기준은 100점 만점에 제1종은 70점 이상, 제2종은 60점 이상이며 합격률은 2020년 기준으로 85.8%이다[23]. 2010년부터 자격시험의 신뢰성 향상과 초보 운전자의 운전능력 향상을 위해 문제은행을 구축해 공개하고 있다[21]. 현재 도로교통공단 안전운전 통합민원 사이트에 공개된 자료는 1·2종 보통, 대형·특수 학과시험과 이륜자동차 학과시험이다. 본 연구에는 2018년 ~ 2023년 1·2종 보통, 대형·특수 학과시험 중 ChatGPT로 입력이 가능한 문장형 문제만을 사용하였다. 이륜자동차 학과시험은 난이도가 상대적으로 낮을 뿐만 아니라[21] 국가통계포털[24]에 따르면 운전면허소지자 약 3370만 중 2종 소형면허 소지자는 약 12,000명에 불과하기 때문이다.

문장형 문제 수는 2018년 ~ 2019년은 700문제, 2020년부터는 680문제이다. 이 중 문제별 지문과 해설을 기준으로 도로교통법, 교통사고처리 특례법, 특정범죄 가중처벌 등에 관한 법률 관련 문제만을 선별하였다. 연도별 문제 수는 Table 1과 같다.

3.1.2 도로교통사고감정사 1차 시험

도로교통사고감정사는 경찰청으로부터 인증받은 공인자격으로 과학적이고 체계적인 조사 및 분석으로 공정한 사고조사를 하기 위해 도입되었으며 도로교통공단이 발급기관이다[25].

6) 세부 출제 항목은 도로교통법 시행령 제46조, 제47조 참조

Table 1. Number of Driver's License Test Questions

Years	Total	Excluded	Analyzed
2018	700	267	433
2019	700	262	438
2020	680	246	434
2021	680	251	429
2022	680	238	442
2023	680	232	448
Total	4120	1496	2624

시험은 1차 객관식 시험과 2차 주관식 시험으로 나누어지며 1차 시험 과목은 교통관련 법규, 교통사고 조사론, 교통사고 재현론, 차량 운동학으로 과목당 25문제, 총 100문제이다. 합격점수 기준은 과목 평균 60점 이상, 각 과목 40점 이상이며 합격률은 2022년 기준 23.83%이다[26]. 본 연구에는 기출문제 중 2018년 ~ 2022년[7] 1차 시험 '교통관련 법규' 과목 문제를 사용하였다. 해당 연도 문제 중 텍스트 기반이 아닌 문제, 도로교통법, 교통사고처리 특례법, 특정범죄 가중처벌 등에 관한 법률 외의 문제는 제외하였다. 연도별 문항 수는 Table 2와 같다.

Table 2. Number of Road traffic Accident Appraiser Test Questions

Years	Total	Excluded	Analyzed
2018	25	1	24
2019	25	1	24
2020	25	0	25
2021	25	1	24
2022	25	1	24
total	125	5	121

3.2 자료 수집 및 분석

3.2.1 자료 수집

OpenAI에서 제공하는 API를 활용하여 파이썬으로 함수를 만들어 시험별로 30회씩 실시하였다. 문제별로 지문과 보기를 구분해 gpt-3.5-turbo 모델에 한 번에 입력하여 응답을 요구하는 방식이며 하나의 결과를 도출한 뒤에는 ChatGPT를 초기화한 후 다음 문제를 입력하는 과정을 반복하였다. 시스템 메시지(system message)와 프롬프트에는 정답에 해당하는 선택지의 번호만 출력하도록 지시사항을 기재하였으며, 온도(temperature) 파라미터는 0으로 설정하였다. 온도 파라미터는 모델 출력의 무작위성과 창의성을 조절하는 파라미터로 값이 높을수록 다양하고 창의적인 결과를 얻을 수 있으며 낮을수록 모델이 각 스텝별로 가장 확률이 높은 단어를 선택해 일관된 결과를 얻을 수 있다.[8]

```python
# GPT predicition 생성
for idx, example in tqdm(enumerate(dataset)):
  problem = dataset[idx]['question']
  option = dataset[idx]['option']
  if len(predictions) and predictions[idx]['prediction'] is not None:
    dataset[idx]['prediction'] = predictions[idx]['prediction']
    print(f'Predictions for example #{idx} is already available!')
    continue

  # API call
  try:
    params = {'model' : "gpt-3.5-turbo",
              'messages' :[
    {"role": "system", "content": system_message},
    {"role": "user", "content": user_message.format(problem, option)}
    ],
              'temperature' : 0}

    response = openai.ChatCompletion.create(**params)
    dataset[idx]['prediction'] = response['choices'][0]['message']['content']

  except Exception as inst:
    print(inst)
    dataset[idx]['prediction'] = None
```

Fig. 1. Python Code for Collecting GPT Responses

3.2.2 자료 분석

수집한 자료에 대해 먼저 시험별, 연도별로 정답률을 분석하였다. 둘째, 세부적인 분석을 위해 내용 영역을 구분하고 영역별로 ChatGPT의 응답 능력에 차이가 있는지를 탐색하였다. 내용 영역 분류는 법률의 장이나 조문을 기준으로 하였다. 구체적으로 도로교통법은 제1장 ~ 제14장까지 장별로 구분하였고[9], 교통사고처리 특례법은 장의 구분이 없으므로 제1조(목적)과 제2조(정의)를 총칙, 제3조(처벌의 특례) ~제4조(보험 등에 가입된 경우의 특례) 제1항을 처벌 특례, 제4조 제2, 3항을 손해배상, 제5조(벌칙)과 제6조(양벌규정)를 벌칙으로 구분하였다. 특정범죄 가중처벌 등에 관한 법률은 제5조의 3을 도주, 제5조의11을 위험운전치사상, 제5조의 13을 어린이보호구역으로 구분하였다. 출제문제별 문항 수는 Table 3과 같다. 셋째, 응답의 일관성을 평가하기 위해

7) 2023년 도로교통사고감정사 시험은 2023.8.27. 시행 예정이다.

8) 다만 OpenAI Platfrom에 따르면 0일 때 가장 결정론적인(deterministic) 결과를 출력하나 여전히 작은 변화가 있을 수 있다고 한다.(https://platform.openai.com/docs/models/gpt-3-5 참조)

9) 제1장 총칙, 제2장 보행자의 통행방법, 제3장 차마 및 노면전차의 통행방법, 제4장 운전자 및 고용주 등의 의무, 제5장 고속도로 및 자동차전용도로에서의 특례, 제6장 도로의 사용, 제7장 교통안전교육, 제8장 운전면허, 제9장 국제운전면허증 제10장 자동차운전학원, 제11장 도로교통공단, 제12장 보칙, 제13장 벌칙, 제14장 범칙행위의 처리에 관한 특례.

Table 3. Number of Road Traffic Accident Appraiser Test Questions

Section[10]		Total	Driver's License Test	Road Traffic Accident Appraiser Test
Road Traffic Act	General Provisions	339	323	16
	Pedestrians Walk	221	220	1
	Ways for Motor Vehicles	1052	1031	21
	Obligations of Drivers	327	319	8
	Expressways	54	54	0
	Traffic Safety Education	32	28	4
	Drivers' Licenses	449	421	28
	International Drivers' Licenses	20	20	0
	Penalty Provisions	93	84	9
	Treatment of Violation of Regulations	140	135	5
Act On Special Cases Concerning The Settlement Of Traffic Accidents	General Provisions	9	7	2
	Special Cases for Punishment	97	80	17
	Compensation for Damages	4	0	4
	Penalty Provisions	2	0	2
Act On The Aggravated Punishment Of Specific Crimes	Driver of Hit-and-Run Vehicle	3	0	3
	Dangerous Driving	4	1	3
	Protection Area for Children	6	3	3

정답률이 100%인 문항의 비율을 살펴보았다.

4. 연구 결과

4.1. 시험별, 연도별 정답률

시험별 ChatGPT의 정답률은 Table 4와 같다. 운전면허 학과시험 2,624개 문제 중 평균적으로 1,211.48개를 정답으로 응답해 44.60%의 정답률을 기록하였다. 랜덤하게 정답을 제출했을 때 정답률(이하 임의추측 정답률) 약 25%보다 높지만 제2종 운전면허 학과시험 합격기준 60점에는 미치지 못하는 점수이다. 또한 도로교통사고감정사 시험에서는 121개 문제 중 평균 42.90개를 정답으로 응답해 정답률이 35.45%이다. 역시 임의추측 정답률보다 높지만 과목별 과락 점수인 40점에 미치지 못했다. 운전면허 학과시험에 비해 도로교통사고감정사 시험의 정답률이 약 10% 낮은 것은 운전면허 학과시험은 운전자로서 도로교통에 참여할 수 있는 자격을 갖추었는지 측정하기 위한 목적이나 도로교통사고감정사 시험은 교통사고감정 전문가 배출을 목적으로 하는 데서 오는 시험 난이도의 차이가 반영된 것으로 보인다.

연도별로는 운전면허 학과시험은 2022, 2023년 정답률이 평균 정답률보다 낮았으며, 도로교통사고감정사 시험은 2018, 2022년 정답률이 평균 정답률을 하회해 두 시험 모두 2022년 이후 성적이 저조하게 나타났다. ChatGPT의 학습데이터가 2021년 9월까지인 점을 원인으로 조심스럽게 추측할 수 있겠으나 2022년 이후 시행된 시험 횟수가 충분하지 않으므로 향후 이 부분은 추가적인 연구가 필요하다.

Table 4. Average Performance by Test and Year

Section		Min	Max	Mean	Std
Driver's License Test	2018	44.57	45.50	44.97	0.27
	2019	45.21	46.58	45.95	0.35
	2020	44.70	47.24	45.71	0.74
	2021	49.42	50.82	49.97	0.34
	2022	42.31	43.89	42.99	0.41
	2023	37.72	38.84	38.28	0.27
Road Traffic Accident Appraiser Test	2018	29.17	33.33	31.39	2.11
	2019	33.33	41.67	35.97	2.32
	2020	41.67	41.67	41.67	0.00
	2021	33.33	41.67	37.92	2.97
	2022	20.83	20.83	20.83	0.00

4.2 내용 영역별 정답률

내용 영역별 ChatGPT의 정답률은 Table 5와 같다.

10) 한국법제연구원의 영문법령 번역을 따르되 지나치게 긴 명칭은 의미를 알 수 있는 범위 내에서 축약했다.

교통사고처리 특례법-총칙(88.89%), 도로교통법-고속도로(56.80%), 교통사고처리 특례법-처벌 특례(52.43%) 순으로 정답률이 높았으며, 특정범죄 가중처벌 등에 관한 법률-도주(0%), 교통사고처리 특례법-벌칙(0%), 교통사고처리 특례법-손해배상(25%) 순으로 정답률이 낮았다. 문제 수가 10문제 미만인 영역을 제외하면 도로교통법-고속도로(56.80%), 교통사고처리 특례법-처벌 특례(52.43%), 도로교통법-보행자 통행방법(50.18%) 순으로 정답률이 높았으며, 도로교통법-교통안전교육(29.69%), 도로교통법-운전면허(34.09%), 도로교통법-범칙 특례(36.28%) 순으로 정답률이 낮았다. 이러한 정답률의 차이는 해당 영역의 특성상 세부적인 내용을 별표에 규정하는데 기인한 것으로 추측할 수 있다. 예를 들어 교통안전교육의 과목·내용·방법·시간(도로교통법 시행규칙 별표 16), 운전면허별 운전할 수 있는 차의 종류(도로교통법 시행규칙 별표 18), 운전면허 벌점(도로교통법 시행규칙 별표 28), 범칙행위별 범칙금액(도로교통법 시행령 별표 8, 9, 10)은 별표에서 세부 내용을 규정하고 있다. 그런데 별표는 법령과 달리 PDF 파일이나 한글 파일, 이미지 파일 형식으로 법제처에서 제공하고 있어 학습 데이터로 구축하기 위해서 법률 본문과 달리 별도의 작업이 필요한 어려움이 있고 이 점이 학습에 영향을 미쳤을 수 있다. 즉, 웹페이지에서 쉽게 접근이 가능한 법률 본문 텍스트만으로 풀이가 가능한 내용 영역과 텍스트가 아닌 표의 형식으로 작성된 파일을 다운로드 등의 방식으로 별도로 수집해야 풀이가 가능한 내용 영역의 차이가 나타난 것으로 볼 수 있다.

4.3. 일관성 능력

동일한 질의에 대해 매번 다른 응답을 한다면 그 응답을 신뢰하기 어려우므로 일관성이 담보되어야 한다. Table 4에서 정답률의 표준편차가 3%p 내로 매우 작은 값을 확인할 수 있었지만 그 자체로 일관성을 담보할 수 없으므로 문제별로 일관된 응답을 하는지 살펴보았다. 30회 실험 중에 정확도가 100%인 문제, 즉 일관되게 정답을 출력한 문제의 비율은 Table 6과 같다. 정답률과 비교했을 때 최대 4.93%p 차이로 정답을 제시한 경우 95% 이상 같은 응답을 출력하는 것으로 나타났다. 일정 정도 같은 질문에 일관성 있는 응답을 기대할 수 있는 것으로 판단된다.

Table 5. Average Performance by Section

Section		Mean	Question	Correct answer[11]
Road Traffic Act	General Provisions	42.59	339	144.38
	Pedestrians Walk	50.18	221	110.90
	Ways for Motor Vehicles	45.92	1052	483.05
	Obligations of Drivers	48.52	327	158.66
	Expressways	56.80	54	30.67
	Traffic Safety Education	29.69	32	9.50
	Drivers' Licenses	34.09	449	153.09
	International Drivers' Licenses	40.65	20	8.13
	Penalty Provisions	44.47	93	41.36
	Treatment of Violation of Regulations	36.28	140	50.79
Act On Traffic Accidents	General Provisions	88.89	9	8.00
	Special Cases for Punishment	52.43	97	50.86
	Compensation for Damages	25.00	4	1.00
	Penalty Provisions	0.00	2	0.00
Act On Specific Crimes	Driver of Hit-and-Run Vehicle	0.00	3	0.00
	Dangerous Driving	50.00	4	2.00
	Protection Area for Children	33.33	6	2.00

11) 정답 문제 수는 30회 실험했을 때 평균 정답 수이다. 예를 들어 도로교통법 총칙(general provisions) 영역 339문제를 30회 실험했을 때 평균적으로 144.38문제는 정답으로 응답한다는 의미이다.

Table 6. Average Performance by Test and Year

Section		Correct Answer Rate(a)	100% Correct Answer Rate(b)	(a) - (b)
Driver's license test	2018	44.97	43.19	1.79
	2019	45.95	43.15	2.80
	2020	45.71	40.78	4.93
	2021	49.97	48.25	1.72
	2022	42.99	40.27	2.72
	2023	38.28	37.05	1.23
Road traffic accident appraiser test	2018	31.39	29.17	2.22
	2019	35.97	33.33	2.64
	2020	41.67	41.67	0.00
	2021	37.92	33.33	4.58
	2022	20.83	20.83	0.00

5. 결론 및 제언

본 연구는 ChatGPT의 경찰 관련 교통법규 질의응답 성능과 한계를 탐색하기 위해 운전면허 학과시험, 도로교통사고감정사 1차 시험에 대한 풀이 능력을 실험하였다. OpenAI에서 제공하는 API를 활용해 시험별로 30회 반복 실험을 통해 시험별 · 연도별 정답률, 내용 영역별 정답률, 일관성 능력을 분석하였다. 분석 결과 첫째, 운전면허 학과시험 및 도로교통사고감정사 1차 시험의 평균 정답률은 각각 44.60%, 35.45%로 나타나 임의 추측 정답률보다 높은 점수를 기록했지만 모두 각 시험의 합격기준에는 미치지 못했다. 연도별로는 2022년 이후 정답률이 평균 정답률을 하회했다. 둘째, 내용 영역별로는 정답률이 29.69% ~ 56.80%의 범위로 나타나 영역별로 큰 편차를 보였다. 도로교통법-고속도로(56.80%), 교통사고처리 특례법-처벌 특례(52.43%), 도로교통법-보행자통행방법(50.18%) 순으로 정답률이 높았으며, 도로교통법-교통안전교육(29.69%), 도로교통법-운전면허(34.09%), 도로교통법-범칙 특례(36.28%) 순으로 정답률이 저조하였다. 셋째, 정답을 제시한 경우 95% 이상 같은 응답을 출력하는 것으로 나타나 일정 정도 일관성 응답을 기대할 수 있었다.

연구 결과를 토대로 ChatGPT를 효과적으로 활용하기 위해 다음과 같은 제언을 제시할 수 있다. 첫째, 교통법규 질의응답과 같이 전문성과 정확성이 요구되는 영역에서는 ChatGPT의 활용에 신중해야 하며, 응답의 신뢰성을 판단할 수 있는 사용자의 전문적인 지식이 중요하다. 인공지능 기술의 발전으로 ChatGPT의 성능이 향상될 수 있을 것이나 효과적으로 활용되는지는 사용자 전문성에 영향을 받을 것이다. 둘째, ChatGPT의 성능과 한계를 분석하기 위한 평가데이터 및 방법이 마련되어야 한다. 현재 해당 분야에서 성능을 평가하는 데이터셋과 평가방법은 기존에 사람을 평가하기 위한 시험에 대부분 의존하고 있다. 그러나 모라벡의 역설[12]처럼 사람이 느끼는 문제별 난이도와 인공지능이 겪는 난이도가 동일하다고 담보할 수 없다. 또한 각종 시험은 그 목적별로 다루는 출제 범위가 제한적이기 때문에 전체 내용을 포괄하기 어렵다. 따라서 경찰 관련 교통법규 영역에서도 성능을 정확하게 평가하기 위한 데이터셋과 평가방법이 마련되어야 한다. 셋째, 활용 가능성을 높이기 위한 추가학습이 필요하다. 양질의 교통 법규 말뭉치를 설계하고 이를 사전학습이나 파인튜닝에 이용한다면 성능 향상을 기대할 수 있다. 데이터셋 설계와 학습 과정에 해당 분야 전문가가 참여해야 하며 법개정사항이 반영되도록 주기적으로 업데이트해야 한다.

'AI will not replace you. A person using AI will.'[27] ChatGPT의 등장으로 사람들이 기대와 우려가 교차하던 2023년 초에 회자된 한 트윗이다.[13] 조직과 사람 모두 빠르게 발전하는 인공지능을 얼마나 잘 활용하는지가 향후 경쟁력을 좌우할 가능성이 높다. 본 연구가 경찰교통 분야에서 인공지능을 잘 활용하기 위한 사전 연구로서 일조하길 기대한다.

REFERENCES

[1] Science & Technology Policy Institute(2023), The Generative AI Era Triggered by ChatGPT. *What the Future Holds and How to Respond,* 02-09.

[2] O. N. Kwon, S. J. Oh, J. Yoon, K. Lee & B. C. Shin. (2023), "Analyzing Mathematical Performances of ChatGPT: Focusing on the Solution of National Assessment of Educational Achievement and the College Scholastic Ability Test. *Communications of Mathematical Education, 37(2),* 233-256. DOI : 10.7468/jksmee.2023.37.2.233

[3] Y. Lee, C. Kim & H. Ahn. (2023). A Study on the ChatGPT: Focused on the News Big Data Service and ChatGPT Use Cases. *Journal of the Korea*

12) 인간에게 어려운 일이 컴퓨터에게는 쉽고 반대로 인간에게 쉬운 일이 컴퓨터에게는 어렵다는 역설

13) Santiago의 트윗 계정에서 확인할 수 있다 (https://twitter.com/svpino/status/1610984481342771200 참조).

Society of Digital Industry and Information Manageme, 19(1), 139-151.
DOI : 10.17662/ksdim.2023.19.1.139

[4] D. H. Kim (2023. 1. 28). The Chosunilbo. https://www.chosun.com/politics/politics_general/2023/01/28/3DVO3AS4CVGDJMAOUCO3N3PAOI/?utm_source=naver&utm_medium=referral&utm_campaign=naver-news.

[5] G. Kim, J. Cho, S. Kim, S. Beak, S. Ryu, J. Koh & B. Kim. (2021). Deep Learning-based Real-time Traffic Accident Type and Fault Information Provision Service. *The Journal of The Institute of Internet, Broadcasting and Communication, 21(3),* 1-6
DOI : 10.7236/JIIBC.2021.21.3.1

[6] S. J. Lim & Y. S. Shin. (2020). A Study on Traffic Violation Surveillance System Based on Edge AI -Detecting a car which cuts in left-turn waiting lines. *KICS Summer Conference 2020*(pp.1315-1316)

[7] G. S. Lee. (2020). Illegal Parking Number Recognition Technology using Deep Learning Algorithm Based on Drone Image. *Journal of The Korean Cadastre Information Association, 22(3),* 20-31
DOI : 10.46416/JKCIA.2020.12.22.3.20

[8] Y. J. Roh & S. H. Bae. (2021), "Forecasting of Traffic Accident Occurrence Pattern Using LSTM", *The Journal of The Korea Institute of Intelligent Transportation Systems, 20(3),* 59-73
DOI : 10.12815/kits.2021.20.3.59

[9] J. D. Rye, S. Park, S. Park, C. Kwon & I. Yun. (2018). A Study for Development of Expressway Traffic Accident Prediction Model Using Deep Learning. *The Journal of The Korea Institute of Intelligent Transportation Systems, 17(4),* 14-25
DOI : 10.12815/kits.2018.17.4.14

[10] S. Huh. (2023). Can we trust AI chatbots' answers about disease diagnosis and patient care?. *Journal of the Korean Medical Association, 66(4)*, 218-222
DOI : 10.5124/jkma.2023.66.4.218

[11] S. K. Kwon & Y. T. Lee. (2023). Investigating the performance of generative AI ChatGPT's reading comprehension ability. *Journal of the Korea English Education Society, 22(2)*, 147-172
DOI : 10.18649/jkees.2023.22.2.147

[12] L. Ouyang et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems, 35*(pp.27730-27744).
DOI : 10.48550/arXiv.2203.02155

[13] OpenAI. (n.d). https://platform.openai.com/docs/models/gpt-3-5.

[14] M. Bommarito II & D. M. Katz. (2022). GPT takes the bar exam. arXiv preprint arXiv:2212.14402.
DOI : 10.48550/arXiv.2212.14402

[15] I. Chalkidis. (2023). ChatGPT may Pass the Bar Exam soon, but has a Long Way to Go for the LexGLUE benchmark. *arXiv preprint arXiv:2304.12202.*
DOI : 10.48550/arXiv.2304.12202

[16] J. H. Choi, K. E. Hickman, A. Monahan, & D. Schwarcz. (2023). *Chatgpt goes to law school. Available at SSRN.*
DOI : 10.2139/ssrn.4335905

[17] S. Park, J. Park & J. Ahn. (2023). Potential Applications and Implications of GPT-4 in Legal Inference Using Korean Legal Aptitude Test (LEET). *Journal of Law & Economic Regulation, 16(1),* 7-28
DOI : 10.22732/CeLPU.2023.16.1.7

[18] M. Baek. (2023). A Study on the Assessment of Korean Language Proficiency of ChatGPT - Focusing on the Reading Section of TOPIK and Idioms -. *Language Facts and Perspectives, 59,* 279-308
DOI : 10.20988/lfp.2023.59..279

[19] J. J. Yu. (2023). Application of Artificial Intelligence for Geography Education - Focusing on Question Answering on ChatGPT -. *Journal of the Association of Korean Photo-Geographers, 33(1)*, 162-173
DOI : 10.35149/jakpg.2023.33.1.011

[20] Y. Yoon. (2023). Artificial Intelligence and Accounting Education : Focusing on ChatGPT and Its Applications. *Korean Computers and Accounting Review, 21(1)*, 1-29
DOI : 10.32956/kaoca.2023.21.1.1

[21] J. Kim. (2013). A study on the Development Plan of Driver's License System - Focusing on the Test for Driver's License -. *Master's Thesis, Graduate School of Transportation / Intelligent Transport Systems*

[22] S. Y. Baik. (2018). A Study on the Driver's License Written Test System to Improve New Driver' Driving Ability. *The Korean Association of Police Science Review, 20(2)*, 31-66

[23] Korean National Police Agency. (n.d). *Korean National Police Agency 2021 White Paper,* 236

[24] Statistics Korea. (n.d). https://kosis.kr/statHtml/statHtml.do?orgId=132&tblId=DT_13201_A002&con

n_path=I3.

[25] KoROAD. (n.d). https://www.safedriving.or.kr/license/licPass.do?menuCode=MN-PO-1541.

[26] KRIVET. (n.d). https://www.pqi.or.kr/inf/qul/infQulBasDetail.do?qulId=811.

[27] Santiago. (n.d). https://twitter.com/svpino/status/1610984481342771200.

이 상 엽(Sang-yub Lee) [정회원]

· 2009년 2월 : 아주대교통ITS대학원 교통공학 석사
· 2020년 9월 ~ 현재 : 고려대 컴퓨터학과 석박사통합과정 중
· 2018년 3월 ~ 현재 : 경찰대학 경찰학과 교수

· 관심분야 : 인공지능, 교통경찰
· E-Mail : yubii2@police.go.kr