Journal of The Korean Data Analysis Society (June 2025), 27(3), 769-780.
https://doi.org/10.37727/jkdas.2025.27.3.769

자율주행 도메인에서 LoRA 기반 sLLM 파인튜닝 연구*

김다연[1], 전수영[2], 주아림[3]

요 약

최근 자율주행 기술의 발전으로 다양한 센서 및 주행 데이터가 급증하고 있으며, 이러한 데이터를 효율적으로 처리하고 실시간으로 응답할 수 있는 지식 처리 시스템의 필요성이 커지고 있다. 특히, 복잡한 자율주행 상황에서 차량이 내리는 의사결정을 인간이 이해할 수 있도록 설명하는 능력은 안전성과 신뢰성 확보를 위한 필수 요소이다. 기존의 RAG(retrieval-augmented generation) 방식은 다양한 도메인 간 일반화 성능이 우수하지만, 특정 도메인 내에서는 정확성과 응답 일관성 측면에서 한계를 가진다. 이러한 문제를 해결하기 위해 본 연구에서는 소형 초거대 언어모델에 LoRA(low-rank adaptation) 기반 파인튜닝을 적용한 LoRA FT-Driver 모델을 제안한다. 제안된 모델은 전체 파라미터의 약 0.01%만을 조정하여 적은 자원으로도 높은 정확도와 일관된 응답을 제공할 수 있으며, 모델 규모가 경량화되어 자율주행차량과 같은 임베디드 시스템에 적합하다. 자율주행 설명 데이터셋(BDD-X)을 사용한 질적·정량적 실험 결과, RAG 대비 우수한 성능을 보였다. 또한, 교통 법규 위반 유형 분류와 같은 확장된 실험에서도 뛰어난 도메인 전이 능력을 입증하였다.

주요 용어 : 자율주행, RAG, Fine-tuning, LoRA, Prompt engineering.

1. 서론

최근 자율주행 연구 동향에 따르면, 자율주행 기술은 인간 운전자의 편의성과 안전성 확보를 중심으로 연구가 활발히 진행되고 있으며, 이는 자율주행 시스템의 실용화에 있어 핵심적인 요소로 부각되고 있다(Jin et al., 2023). 자율주행 차량의 사고 원인은 대부분 자율주행 자동차의 특정 행동에 대해 후행 차량이 알아차리지 못해 발생한 사고이다(Kim, Lee, Yeon, 2024). 이를 개선하기 위해 자율주행 자동차의 행동을 예측하고 정당화하는 모델이 필요하다. 특히 자율주행 시스템이 복잡한 상황에서 내리는 판단에 대해 인간이 이해할 수 있는 방식으로 설명하는 것은 실시간 상호작용 및 안전성 확보를 위한 핵심 요소이며, 실제로 자율주행에서 발생할 수 있는 다양한 인간적 오류에 대응할 수 있는 설명 가능한 시스템의 필요성이 강조되고 있다(Jeong, Choi, 2024). 자율주행 환경에서 운전자의 제어권 인수 반응 시간은 시각자극-수동반응 과제보다 청각자극-구두반응 과제에서

*본 논문은 2024년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (RS-2024-00352792).
[1]30019 세종시 세종로 2511, 고려대학교 빅데이터사이언스학과 석사과정. E-mail: dayun0405@korea.ac.kr
[2]30019 세종시 세종로 2511, 고려대학교 빅데이터사이언스학부 교수. E-mail: scheon@korea.ac.kr
[3](교신저자) 30019 세종시 세종로 2511, 고려대학교 경제통계학과 박사수료. E-mail: jooalim@korea.ac.kr
[접수 2025년 3월 31일; 수정 2025년 4월 22일; 게재 확정 2025년 4월 25일]

더 빠른 반응을 보인 것으로 보고되었다(Go, Lee, 2023). 이는 자율주행 시스템의 실시간 상호작용을 위한 사용자 인터페이스 설계에 있어 음성 기반 접근이 보다 효과적일 수 있음을 시사한다.

기존 연구는 주로 RAG 방식을 활용하여 도메인 간 일반화 성능을 강조했으나, 이는 모델 자체의 도메인 특화 지식을 내재화하기 어렵고, 정확성과 응답 일관성 면에서 한계를 보였다(Balaguer et al., 2024). 특히, 실시간 정보 검색의 불안정성은 자율주행처럼 정확한 실시간 응답이 중요한 도메인에서 단점으로 작용할 수 있다. 한편, 소규모 모델을 도메인 특화 데이터로 파인튜닝할 경우 모델이 해당 도메인의 지식을 효율적으로 학습하여 보다 정확하고 일관된 응답을 제공할 수 있다는 점이 제시되었다(Soudani, Kanoulas, Hasibi, 2024).

이에 본 연구는 소형 초거대언어모델(small large language model, sLLM)에 LoRA 기반의 파인튜닝을 적용한 LoRA FT-Driver 모델을 제안하여, 자율주행 도메인에서 정확하고 일관된 실시간 응답 시스템을 구축하고자 한다. LoRA 파인튜닝은 매우 적은 수의 파라미터만을 조정함으로써 효율적인 학습을 가능하게 하고, 차량 내 제한된 자원에서도 안정적인 운영이 가능한 장점이 있다.

2장에서는 기존 방법론에 대한 설명, 3장에서는 본 연구에서 제안하는 LoRA FT-Driver 모델에 대한 설명, 4장에서는 실 자료분석 결과, 그리고 5장에서는 결론을 서술한다.

2. 자율주행분야에서의 기존 대형 언어모델

RAG-Driver는 자율주행 분야에서 설명 가능성과 신뢰성을 향상시키기 위해 제안된 다중 모달 대형 언어모델로, 주행 영상 입력을 바탕으로 차량의 행동 설명, 정당화, 그리고 제어 신호를 생성하는 것을 목표로 한다(Yuan et al., 2024). 이 모델의 핵심 아이디어는 세 가지로, 첫 번째는 설명 가능성 강화이다. 기존의 비주얼 중심 설명 방식은 사용자에게 일방적이고 이해하기 어려운 결과만을 제공하는 한계가 있었다. RAG-Driver는 자연어 설명을 통해 차량의 행동을 직관적으로 이해할 수 있도록 설계되었으며, 이는 자율주행 시스템의 투명성과 신뢰성을 높이는 데 중요한 역할을 한다. 두 번째는 RA-ICL(retrieval-augmented in-context learning)의 검색기반 접근법이다. 인-컨텍스트 학습(in-context learning, ICL)은 테스트 쿼리에 소수의 예시를 함께 제공하여, 이를 문맥 정보로 활용하는 방식이다. LLM은 이 문맥적 예시로부터 유추된 유사성을 기반으로 파라미터 업데이트 없이 즉시 새로운 입력에 대한 출력을 생성한다. RAG-Driver는 유사한 주행 시나리오를 벡터스토어에서 효율적으로 검색하고 이를 현재의 상황에 문맥적 정보로 활용하여, 모델의 예측을 개선하는 방식이다. 이는 특히 훈련 데이터 밖의 새로운 환경에서도 높은 일반화 성능을 제공한다. 이와 같은 검색기반 접근법은 복잡한 운전 상황에서의 예측 보강 효과를 보여준다. 마지막으로 다중 모달 처리이다. 영상과 텍스트를 결합한 다중 모달 입력을 통해, 모델은 다양한 센서 데이터를 효과적으로 융합하여 운전 상황을 종합적으로 이해할 수 있다.

자율주행 시스템의 효과적인 의사결정을 위해서는 다양한 주행 환경에서의 설명 가능한 데이터를 활용하는 것이 필수적이다. 해당 연구에서는 자율주행 행동의 정당화 및 설명 생성을 위한 벤치마크로 널리 사용되는 BDD-X(Berkeley DeepDrive eXplanation; Kim et al., 2018) 데이터셋을 채택하였다. BDD-X 데이터셋은 약 6,970개의 비디오로 구성되어 있으며, 비디오는 다양한 주행 조건에서 촬영되었고, 각 행동에는 설명이 주석으로 달려 있다.

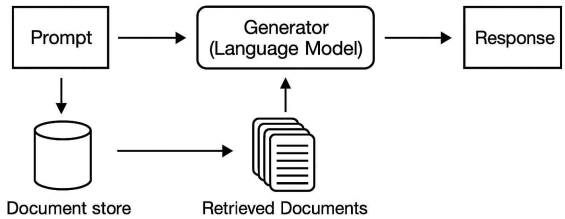

Figure 1. RAG Process

이러한 복합 정보를 바탕으로 모델이 운전자의 의사결정을 이해하고 예측할 수 있도록 설계되어 있다. 각 문장을 벡터화하기 위해 문장 임베딩 모델(sentence-BERT; Reimers, Gurevych, 2019)를 사용하여 구축하였다.

해당 연구에서의 핵심 방법론 RAG는 외부 지식 데이터베이스에서 문서를 검색하여 이를 입력 쿼리에 추가하는 방식으로, 외부 지식과 특정 도메인 지식을 효과적으로 반영할 수 있다(Lewis et al., 2020). 프로세스는 Figure 1과 같으며, RAG는 검색(retrieval) 단계와 생성(generation) 단계를 결합하여 다음과 같은 방식으로 질의에 응답한다.

(a) 검색단계: 입력 쿼리에 가장 유사한 문서들을 검색하여 컨텍스트를 구성한다.
(b) 생성단계: 구성된 컨텍스트를 기반으로 사전학습된 언어모델을 통해 최종응답을 생성한다.

이러한 접근법은 모델이 최신 정보나 특정 도메인 지식을 활용하여 보다 정확하고 신뢰성 있는 응답을 생성할 수 있도록 한다. 특히, RAG는 오픈 도메인 질의응답과 같은 지식 집약적인 작업에서 우수한 성능을 보였다.

3. LoRA FT-Driver

본 연구에서는 RAG-Driver의 접근법과 달리, 텍스트 기반 sLLM을 중심으로 연구를 진행한다. 특히 LoRA(low-rank adaptation) 기반 파인튜닝(Hu et al., 2021)을 활용하여 단일 모달 텍스트 데이터로도 자율주행 도메인에 특화된 정밀한 응답을 생성할 수 있음을 보이고자 한다. LoRA 기반 파인튜닝은 모델 내에 직접 도메인 특화된 지식을 내재화하기 때문에 외부 검색 없이 즉시 추론이 가능하여, 실시간 환경에서 더욱 우수한 일관성과 응답속도를 보장할 수 있다. 이로써, 복잡한 멀티모달 모델에 비해 구조가 간소하고 계산 효율성이 높은 모델을 구현하여, 향후 음성 기반 사용자 인터페이스 등 다양한 응용 서비스로 확장할 수 있는 기반을 마련할 수 있다.

3.1. 데이터셋 전처리

학습 데이터로는 자율주행 차량의 행동을 설명하는 데 사용되는 BDD-X 데이터셋을 활용하였다. 각 샘플은 운전 중 발생할 수 있는 다양한 상황과 이에 대한 적절한 운전 행동 또는 지침으로 구성되어 있다. 데이터 전처리 과정은 텍스트 정제, 토큰화 및 벡터화, 벡터스토어 구축으로 이루어져 있다.

텍스트 정제 과정에서는 불필요한 공백, 특수문자 등을 제거하여 텍스트를 정제한다. 그리고 정제된 텍스트를 문장 단위로 토큰화하고, 각 문장을 임베딩하여 벡터 표현으로 변환한다. 이를 위해 사전 학습된 언어 모델을 활용한다. 문장을 고정 길이의 고차원 벡터 표현으로 변환하여, 의미적 유사성을 수치적으로 반영하도록 한다. 벡터화된 문장을 효율적으로 검색하기 위해 FAISS(Facebook AI similarity search)를 기반으로 벡터스토어를 구축한다. 이 벡터스토어는 임베딩된 벡터 간의 유사성을 빠르게 검색할 수 있도록 최적화되어 있으며, 추후 모델의 추론 과정에서 질의와 가장 유사한 문장을 신속히 검색하여 모델의 응답 정확도와 일관성을 향상시키는 데 활용된다.

3.2. 모델 학습

대규모 언어모델을 특정 도메인이나 작업에 효과적으로 적용하기 위해서는 파인튜닝이 필수적이다. 그러나 기존의 전통적인 파인튜닝 방식은 모델 전체의 파라미터를 조정해야 하므로 방대한 계산 자원과 메모리를 필요로 하며, 이는 비효율성을 초래할 수 있다. 이러한 한계를 극복하기 위해 Hu et al.(2021)은 파라미터의 효율성을 크게 향상시키는 방법을 적용했다.

LoRA는 사전 학습된 언어 모델의 가중치를 고정한 상태에서 각 트랜스포머(transformer) 레이어에 소규모의 저랭크(low-rank) 행렬을 추가하여 제한된 수의 파라미터만 학습시키는 방식으로 파인튜닝을 수행한다. Figure 2는 LoRA의 기본 구조를 나타낸 것으로, 기존의 사전 학습된 가중치 행렬(W)에 새롭게 학습 가능한 저랭크 행렬(A,B)을 추가하여 가중치를 근사하는 방식을 사용한다(Hu et al., 2021). 이는 아래 수식 (1)과 같이 표현할 수 있다.

$$W = W_0 + BA, B \in R^{d \times r}, A \in R^{r \times k}, r \ll d,k \tag{1}$$

수식 (1)에서 d와 k는 원본 가중치 행렬의 차원이며, r은 LoRA의 랭크로 설정된 작은 값(본 연구에서는 $r=8$)을 의미한다. 이렇게 설정된 r값은 충분한 표현력을 제공하면서도 파라미터의 수를 효과적으로 제한하여 계산 효율성과 메모리 절약을 동시에 달성한다. LoRA에서는 초기 행렬 A는 랜덤한 가우시안 분포로, B는 0으로 초기화하여, 학습 시작 시 보정값 $\triangle W = BA$가 0으로 설정되어 있다. 이후 점진적으로 학습하며 최적의 표현을 찾는다. 저랭크 가중치 A,B는 Adam 옵티마이저를 통해 최적화된다. 파라미터 업데이트 규칙은 Adam 옵티마이저의 일반적인 규칙을 따른다. 사전학습 가중치 W_0는 고정되고, 오직 A,B만 최적화된다.

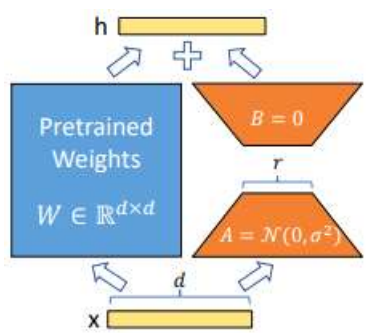

Figure 2. Architecture of LoRA(Hu et al., 2021)

$$A \leftarrow A - \eta \frac{\partial L}{\partial A}, B \leftarrow B - \eta \frac{\partial L}{\partial B} \tag{2}$$

매 학습 단계마다, 수식 (2) 규칙에 따라 A,B가 업데이트된다(Hu et al., 2021). η는 Adam 옵티마이저를 통해 결정되는 학습률이며, L은 조건부 언어 모델링(conditional language modeling) 손실이다. 본 연구에서 사용한 손실 함수는 조건부 언어모델링 손실로 다음과 같다.

$$L(\Theta) = -\sum_{(x,y)\in Z} \sum_{t=1}^{|y|} \log\left(p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t|x, y < t)\right) \tag{3}$$

수식 (3)에서 Φ_0는 원본 사전학습된 모델의 파라미터를 나타내며, $\Delta\Phi(\Theta)$는 LoRA를 통해 추가된 학습 가능한 파라미터를 나타낸다. 이 식은 LoRA를 통해 추가된 소수의 파라미터(Θ)만 최적화하며 전체 모델의 안정성과 효율성을 높인다(Hu et al., 2021).

이 방식을 사용하면 전체 파라미터의 약 0.01% 수준의 파라미터만 조정하여 기존의 파인튜닝 방식과 동등하거나 더욱 우수한 성능을 달성할 수 있다(Hu et al., 2021). 또한, 학습에 필요한 GPU 메모리 사용량을 기존 파인튜닝 대비 약 3분의 1로 감소시켜, 추가적인 추론 지연을 발생시키지 않아 효율적인 모델 적용이 가능하도록 한다. 이러한 효율적인 파인튜닝 방법은 자율주행과 같은 특정 도메인에 언어 모델을 적용할 때에도 제한된 자원으로 효과적인 성능 개선을 가능하게 해준다.

따라서 본 연구에서는 LoRA 기반의 파인튜닝 방식을 적용하여, 자율주행 차량 내에서 신속하고 효율적으로 동작할 수 있는 도메인 특화 지식 생성 모델 LoRA FT-Driver를 구축하였다. 파인튜닝 절차는 (a) LoRA 어댑터 추가, (b) 하이퍼파라미터 설정, (c) 프롬프트 설계 및 최적화로 진행하였다.

학습 가능한 파라미터 수를 줄이면서도 성능을 유지하기 위해 모델의 각 트랜스포머 레이어에 LoRA 어댑터를 삽입하고 학습률, 배치 크기, 에포크 수 등 파인튜닝에 필요한 하이퍼파라미터를 설정한다. 모델의 질의 응답 성능을 향상시키기 위해 효과적인 프롬프트를 설계하였다. 이를 통해 모델이 사용자 질의에 대해 보다 정확하고 일관된 응답을 생성할 수 있도록 한다.

본 연구에서는 Microsoft에서 2025년에 공개한 sLLM인 Phi-4-mini-instruct를 파인튜닝 대상으로 사용하였다. 해당 모델은 3.8B 파라미터 규모의 트랜스포머 구조를 기반으로 한다. 이 모델은 사용자 지시를 정확히 따르는 능력과 논리적 추론 성능, 안정성 측면에서 우수한 성능을 보이며, 최대 128K 토큰에 이르는 긴 문맥을 처리할 수 있는 구조와 빠른 응답 속도로 동작할 수 있도록 최적화되어 있다. LoRA는 트랜스포머의 쿼리 및 값 모듈에 적용되었으며, 적은 수의 학습 파라미터만으로도 충분한 표현력을 확보할 수 있도록 균형 있게 설계되었다.

이러한 접근법을 통해, 본 연구는 간단한 구조와 빠른 응답 속도를 갖춘 모델 LoRA FT-Driver를 구현하였으며, 이는 자율주행 도메인에 특화된 지식 생성에 효과적으로 활용될 수 있다.

4. 실증자료분석

4.1. BDD-X 데이터

본 연구에서는 자율주행 행동의 정당화 및 설명 생성을 위한 벤치마크로 널리 사용되는 BDD-X(Kim et al., 2018)의 테스트 데이터를 사용하였다. BDD-X 데이터셋은 다양한 주행 조건(주/

야간, 고속도로/도시/시골, 여름/겨울 등)에서 촬영된 비디오로 평균적으로 3~4개의 행동(가속, 감속, 우회전 등)을 포함하며, 각 행동과 설명이 주석으로 달려 있다. BDD-X 데이터 셋은 Table 1과 같이 구성 되어있다.

이 데이터의 텍스트 정보를 바탕으로 모델이 운전자의 의사결정을 이해하고 예측할 수 있는지 확인했다. 모델의 성능 평가를 위하여 ROUGE(recall-oriented understudy for gisting evaluation)와 BLEU(bilingual evaluation understudy)를 사용하였다. 또한 실시간 의사결정 지원을 위한 자율주행 차량 환경 특성상 응답 시간 역시 중요한 지표로 추가하였다. ROUGE는 모델이 생성한 후보 문장과 사람이 작성한 참조 문장 사이의 유사도를 평가하는 데 사용되는 대표적인 자동 평가 지표이다(Lin, 2004). 본 연구에서는 그중에서도 ROUGE-1과 ROUGE-L을 사용하여 평가하였다. BLEU는 생성된 문장이 참조 문장과 얼마나 유사한지 n-gram 정밀도를 기반으로 평가하는 대표적인 자동 평가 지표로, 단어 순서나 구문 구조의 유사성 평가에 적합하다(Papineni et al., 2002). 응답 시간은 모델이 특정 쿼리에 대해 응답을 생성하는 데 걸리는 시간을 의미한다.

LoRA FT-Driver 모델의 결과는 학습을 진행하지 않은 sLLM 모델에 BDD-X 데이터셋을 RAG로 연결한 모델과 비교를 진행하였고, 이때 RAG 모델은 참조문서 k를 1, 3, 5로 나누어 비교하였다. Table 2과 같이 모든 질적 평가 지표(ROUGE-1: 0.1495, ROUGE-L: 0.1399, BLEU: 0.0178)에서 LoRA FT-Driver 모델이 가장 높은 성능을 보였다. 반면, RAG 기반 모델은 참조 문서 수(k값)에 따라 성능이 달라졌으며, 최고 성능을 보인 RAG($k=1$)의 경우에도 ROUGE-1 값이 0.0888로 파인튜닝 모델 대비 현저히 낮은 성능을 보였다. 특히, RAG 방식은 k값이 증가함에 따라 오히려 성능이 저하되는 경향을 보였다(RAG($k=3$): 0.0770, RAG($k=5$): 0.0693). 이러한 결과는 검색된 외부 문서가 쿼리와의 관련성이 낮을 때, 정보의 혼잡성이 증가해 생성된 텍스트의 정확성이 떨어지기 때문으로 해석할 수 있다(Balaguer et al., 2024).

Table 1. Characteristics of the BDD-X Dataset

Attribute	Value / Info	Description
total number of videos	6,970	real-world driving videos totaling approximately 77 hours
total number of frames	~8,400,000 frames	each video is around 40 seconds on average
total number of annotated actions	26,228	each action is paired with a description and an explanation
average number of actions/video	~3.8 actions	each video contains multiple annotated driving situations
driving conditions	day/night, mostly urban	covers diverse lighting and road environment conditions
frequent keywords (action)	car, stop, driving, stopped, etc.	commonly appear in describing vehicle behavior and driving actions

Table 2. Performance evaluation results by model

Model	ROUGE-1	ROUGE-L	BLEU	Latency
LoRA FT-Driver	0.1495	0.1399	0.0178	1.145
RAG(k=1)	0.0888	0.0745	0.0051	1.099
RAG(k=3)	0.0770	0.0646	0.0041	1.253
RAG(k=5)	0.0693	0.0589	0.0033	1.438

평균 응답 시간 측면에서도 LoRA FT-Driver 모델이 평균 1.145초로 가장 빠른 응답 속도를 나타냈다. RAG 기반 모델은 문서 검색 및 재순위화 과정으로 인해 응답 시간이 더 길어졌으며, k값이 증가할수록 더 많은 시간이 소요되었다(RAG($k=1$): 1.099초, RAG($k=3$): 1.253초, RAG($k=5$): 1.438초). 자율주행 차량과 같은 실시간 의사결정이 중요한 환경에서는 응답 지연 시간이 매우 중요하기 때문에, 이러한 결과는 LoRA FT-Driver 모델의 우수성을 뒷받침하는 근거가 된다(Soudani, Kanoulas, Hasibi, 2024).

본 연구에서 사용한 방법은 최적화된 파인튜닝 방식이 외부 데이터를 실시간으로 참조하는 RAG 방식보다 정확성과 속도 측면에서 더 효과적이다(Soudani, Kanoulas, Hasibi, 2024). 이는 파인튜닝 기법을 통해 모델이 도메인 특화 지식을 내재화하여 질의 응답의 일관성과 정확성이 향상되었기 때문으로 해석할 수 있다. 또한, RAG 방식은 컨텍스트가 적절한 경우 효과적이나, 외부 참조 품질이 낮을 경우 성능 저하를 겪을 수 있다(Balaguer et al., 2024).

이러한 결과를 종합적으로 볼 때, 본 연구에서 제안하는 LoRA 기반 파인튜닝 방식은 제한된 자원과 빠른 응답성을 요구하는 자율주행 차량 환경에 더욱 적합하며, 향후 텍스트 기반 인터페이스와 같은 실질적인 차량 내 서비스 개발로 확장될 수 있는 가능성을 제시한다.

정량적인 평가지표 외에도, LoRA FT-Driver 모델이 실제 시나리오에 대해 얼마나 적절하고 상황에 맞는 운전 지침을 생성하는지를 확인하기 위해 질적 평가를 수행하였다. BDD-X 데이터셋의 테스트 데이터셋 중 5개의 샘플을 랜덤하게 추출한 후, 각각의 시나리오에 대한 레퍼런스 응답과 LoRA FT-Driver 모델의 생성 결과를 비교하였다.

Table 3에서 볼 수 있듯이, 전반적으로 모델은 시나리오에 기반한 자연스러운 응답을 생성하였으며, 일부 응답에서는 상황에 대한 추가적인 추론을 통해 더욱 상세한 지시를 제공하였다. 특히, 교차로 상황(시나리오 3)이나 교통량이 없는 상황(시나리오 4)에서는 상황을 정교하게 해석하고 적절한 운전 지침을 생성하는 모습을 확인할 수 있었다.

반면, 시나리오가 다소 모호하거나 짧은 경우(시나리오 1, 5)에는 일반화된 응답이나 과도한 추론이 포함되기도 하였다. 예를 들어, 시나리오 5에서는 '좌측 차로의 느린 차량을 추월하기 위해'라는 비교적 간단한 상황에 대해 모델은 다른 차량이 진입하는 복잡한 상황을 가정하여 설명하였다. 이는 파인튜닝 데이터에 포함된 복합적인 시나리오를 일반화하는 과정에서 발생한 것으로 추정된다.

Table 3. Response examples from LoRA FT-Driver model

Number	Scenario	Reference response	Model-Generated Response
1	because the lane to the left is clear	The car is swerving around another car.	The car merges into another road and continues straight ahead at night with no traffic around it.
2	to allow a vehicle to merge	The car slows.	The car is slowing down and steering right slightly...
3	Traffic ahead of the car is slowly advancing...	The car is advancing forward slowly.	The vehicle continues to move forward at slow speed...
4	because there is no traffic	The car drives down the street.	The car accelerates to normal speed and continues...
5	to pass the cars moving slow in the left lane	The car maintains a steady speed.	The car is merging into traffic from right...

이러한 질적 평가는 파인튜닝된 모델이 단순히 응답 길이를 채우는 수준을 넘어 상황 인식 및 추론 능력을 어느 정도 학습하였음을 보여준다. 그러나 일부 불필요한 과추론은 향후 학습 데이터의 다양성과 정교함을 통해 보완될 수 있을 것으로 판단된다.

RAG-Driver에서 제시한 방식과 유사하게 BDD-X 데이터셋을 이용하여 성능을 평가했으며, 설명 가능성 및 응답의 질을 ROUGE 및 BLEU 점수를 통해 평가했다. 실험 결과 RAG-Driver에서 제시한 RA-ICL 기반의 일반화 성능 대비 본 연구가 LoRA 기반 파인튜닝을 통해 보다 높은 도메인 특화 성능을 보였음을 입증했다. RAG-Driver는 제로샷 일반화에서 강력한 성능을 보였으나, 본 연구는 제로샷 환경보다 도메인 특화 환경에서 더욱 우수한 성능을 나타내어, 실제 서비스 적용 시 효과적일 수 있음을 강조하였다.

4.2. Traffic violations 데이터

본 실험에서는 Kaggle에 공개된 Traffic violations 데이터셋을 활용하였다(Kaggle, 2022). 해당 데이터셋은 미국 내 실제 교통 법규 위반 사례들을 기반으로 하며, 각 레코드는 위반 내용과 함께 차량 정보, 운전자 특성, 위반 유형 등 다양한 텍스트 및 범주형 데이터를 포함하고 있다. 위반 유형은 Table 4에서 볼 수 있듯이 경고(warning), 인용(citation), 차량 정비 위반(SERO; Safety Equipment Repair Order) 세 가지로 분류 되어있다.

본 연구에서는 텍스트 기반 분류 성능 비교를 위해 위반 유형을 분류 목적의 정답 레이블로 설정하고, 텍스트 설명을 설명변수로 활용하였다.

해당 과제를 통해 기존 자율주행 주행 설명 데이터를 학습한 LoRA FT-Driver 모델이 다른 교통 관련 도메인에서도 얼마나 일반화된 성능을 보이는지를 검증하고자 하였다. 향후 본 모델이 차량 내에서 온디바이스로 탑재되어 다양한 교통 문맥을 처리해야 함을 고려할 때, 실제 교통 법규 위반 분류 태스크는 유의미한 테스트베드가 될 수 있다.

파인튜닝된 모델과 RAG 기반 모델(RAG($k=1$), RAG($k=3$), RAG($k=5$))을 비교하기 위해, 각 클래스(warning, citation, SERO)당 예시 8쌍씩 총 24개의 문장을 프롬프트에 제공하여, 모델이 해당 입력 문장의 위반 유형을 예측하도록 구성하였다.

실험은 이전과 동일한 방식으로 진행되었으며, LoRA FT-Driver 모델은 자율주행 주행 설명 텍스트를 기반으로 사전 파인튜닝된 모델이다. RAG 모델은 동일한 자율주행 설명 텍스트 문서를 벡터스토어에 저장하고, 질의와 유사한 문서를 검색하여 이를 기반으로 응답을 생성하도록 구성하였다. 모델의 성능 평가는 정확도(accuracy)와 클래스 별 F1-score의 평균인 Macro F1-score, 응답시간으로 설정했다.

Table 4. Characteristics of the Traffic violations dataset

Attribute	Value / Info	Description
Total number of samples	~70,000	based on real-world traffic violations in the state of Maryland
Top 3 violation types	warning: 34,382 / citation: 32,452 / SERO: 3,506	target variable 'Violation.Type' includes these major categories
Number of features	21 features	includes vehicle info, driver profile, and textual violation descriptions
Frequent keywords (description)	failure, driving, mph, vehicle, etc.	related to driving behavior, vehicle status, and types of violations

Table 5. Performance comparison of traffic violation classification

Label	Accuracy	Macro F1	Latency (sec)
LoRA FT-Driver	0.745	0.731	0.346
RAG(k=1)	0.505	0.492	0.404
RAG(k=3)	0.540	0.531	0.339
RAG(k=5)	0.537	0.532	0.359

실험 결과, 파인튜닝 기반 모델이 모든 성능 지표에서 우수한 결과를 보였다(Table 5). 특히 정확도와 Macro F1 지표에서 각각 0.745, 0.731로 RAG 계열 모델보다 현저히 높은 수치를 기록하였다. 반면 RAG 모델은 k 값에 따라 성능이 불안정하게 변화하였으며, $k=1$에서는 가장 낮은 정확도(0.505)를, $k=3$과 $k=5$에서는 비슷한 수준의 성능을 나타냈다. 또한 RAG($k=3$)에서는 문서 검색 과정 중 메모리 과부하가 발생하는 등 실험 환경에서도 비효율적인 면모가 드러났다.

이러한 결과는 LoRA FT-Driver 모델이 사전에 자율주행 주행 설명 데이터를 학습함으로써, 교통 위반 분류와 같은 연관된 태스크에도 도메인 적응력이 높게 나타났기 때문으로 해석된다. 반면 RAG 모델은 관련 문서를 실시간으로 검색하더라도 도메인 불일치로 인해 유의미한 정보 검색에 실패한 사례가 많았다.

5. 결론

본 연구에서는 sLLM에 LoRA 기반의 파인튜닝을 적용한 LoRA FT-Driver 모델을 제안하였다. 제안된 모델이 사용되어지는 sLLM은 차량 내부의 제한된 임베디드 환경에서도 적은 리소스로 안정적으로 구동할 수 있어 차량 내 설치 비용과 운영 비용을 최소화할 수 있는 이점이 있다. 또한 LoRA FT-Driver 방법론은 자율주행 관련 텍스트 데이터를 효과적으로 학습하여 사용자가 요구하는 정확한 도메인 특화 정보를 안정적으로 제공한다. 더불어 본 연구의 LoRA 기반 파인튜닝 방식은 적은 데이터만으로도 높은 성능을 유지할 수 있어, 자율주행 차량 내 제한된 데이터 환경에서도 사용자의 개인 맞춤형 서비스를 효과적으로 구축할 수 있다.

본 연구의 주요 한계는 학습에 사용된 데이터의 양과 범위가 제한적이었다는 점이다. 본 실험에 사용된 BDD-X 데이터셋은 실제 운전 영상을 기반으로 운전자의 시나리오와 의도를 주석으로 기록한 데이터로 구성되어 있다. 이러한 데이터는 영상 기반의 주행 상황을 텍스트 형태로 변환하기 위해 높은 품질의 수작업 주석 작업이 필요하며, 이로 인해 데이터 수집 및 확장이 시간적·비용적 측면에서 상당한 제약을 받는다. 이와 같은 제약은 다양한 도로 환경과 복잡한 운전 시나리오를 충분히 포괄하기 어려운 한계로 작용할 수 있으며, 결과적으로 모델의 일반화 성능과 응답 다양성 측면에서 제한을 줄 수 있다.

향후 연구에서는 BDD-X와 같은 수작업 주석 기반 데이터셋의 한계를 극복하기 위해, 운전 영상에서 시나리오 설명을 자동으로 추출할 수 있는 비전-언어 모델 기반 주석 자동화 기법을 도입하는 방안을 고려할 수 있다. 또한, 다양한 도로 환경과 상황을 반영한 멀티모달 데이터셋을 구성함으로써, 모델의 일반화 능력을 높이고 보다 정교하고 설명력 있는 응답 생성을 목표로 하는 연구도 진행될 수 있다.

이와 더불어, 실제 차량 내 인공지능 서비스에의 적용을 고려한다면 음성 기반 인터페이스와의 연계 역시 중요한 향후 과제가 될 수 있다. 현재는 텍스트 형태로 생성된 운전 지시를 사용자에게 제공하고 있으나, 이를 자연스러운 음성으로 변환하는 TTS(text-to-speech) 기술과 연동함으로써 보다 직관적이고 사용자 친화적인 피드백을 제공할 수 있다. 나아가, 운전자가 음성으로 상황을 설명하면 시스템이 이를 인식하여 적절한 지시를 제공하는 STT(speech-to-text) 기반 입력 방식과의 통합도 가능하다. 이러한 음성 기반 양방향 상호작용 시스템은 실제 주행 환경에서의 사용성을 높이는 데 기여할 수 있으며, 자율주행 차량 내 AI 보조 시스템의 실용화를 위한 중요한 기반 기술로 작용할 것이다.

References

Balaguer, M. A. de L., Benara, V., Cunha, R. L. de F., Estevão Filho, R. de M., Hendry, T., Holstein, D., Marsman, J., Mecklenburg, N., Malvar, S., Nunes, L. O., Padilha, R., Sharp, M., Silva, B., Sharma, S., Aski, V., Chandra, R. (2024). RAG vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture, *Arxiv Preprint*, arXiv:2401.08406. DOI: https://doi.org/10.48550/arXiv.2401.08406.

Go, Y., Lee, J. (2023). Differences in take-over and driving performance by secondary task type and road congestion in autonomous driving: focused on young and older drivers, *Journal of the Korean Data Analysis Society*, 25(3), 1177-1192. DOI: https://doi.org/10.37727/jkdas.2023.25.3.1177. (in Korean).

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2021). LoRA: low-rank adaptation of large language models, *Arxiv Preprint*, arXiv:2106.09685.
DOI: https://doi.org/10.48550/arXiv.2106.09685.

Jeong, M., Choi, B. (2024). Trend analysis of traffic psychology research through text mining: applying topic modeling, *Journal of the Korean Data Analysis Society*, 26(2), 725 – 741.
DOI: https://doi.org/10.37727/jkdas.2024.26.2.725. (in Korean).

Jin, H., Joo, A., Lee, D., Cheon, S. (2023). A trend analysis of autonomous driving research using topic modeling. *Journal of The Korean Data Analysis Society*, 25(3), 1177-1192. https://doi.org/10.37727/jkdas.2023.25.5.1671. (in Korean).

Kaggle (2022). Traffic violations dataset, kaggle datasets. https://www.kaggle.com/datasets/nikhil1e9/traffic - - violations?resource=download.

Kim, G., Lee, E., Yeon, G. (2024). A study on accident causes of self-driving vehicles to improve stability. *Journal of the Korean Data Analysis Society*, 26(5), 1345-1356. https://doi.org/10.37727/ jkdas.2024.26.5.1345. (in Korean).

Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z. (2018). Textual explanations for self-driving vehicles, *Computer Vision – ECCV 2018*, 563 – 578. DOI: https://doi.org/10.1007/978-3-030-01216-8_35.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks, *Arxiv Preprint*, arXiv:2005.11401. DOI: https://doi.org/10.48550/arXiv.2005.11401.

Lin, C.-Y. (2004). ROUGE: a package for automatic evaluation of summaries, Text Summarization Branches Out: *Proceedings of the ACL-04 Workshop*, 74-81. https://aclanthology.org/W04-1013.

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318. DOI: https://doi.org/10.3115/1073083.1073135.

Reimers, N., Gurevych, I. (2019). Sentence-BERT: sentence embeddings using siamese BERT-networks, *Proceedings of the 2019 Conference on EMNLP and the 9th International Joint Conference on NLP (EMNLP-IJCNLP)*, 3982-3992. DOI: https://doi.org/10.18653/v1/D19-1410.

Soudani, H., Kanoulas, E., Hasibi, F. (2024). Fine tuning vs. retrieval augmented generation for less popular knowledge, *Arxiv Preprint*, arXiv:2403.01432. DOI: https://doi.org/10.48550/arXiv.2403.01432.

Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., Gadd, M. (2024). RAG-Driver: generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model, *Arxiv Preprint*, arXiv:2402.10828. DOI: https://doi.org/10.48550/arXiv.2402. 10828.

LoRA-based Fine-tuning of sLLM for Domain-specific Knowledge Systems in Autonomous Driving*

Dayeon Kim[1], *Sooyoung Cheon*[2], *Ah-Rim Joo*[3]

Abstract

Recent advances in autonomous driving technology have resulted in an explosive increase in sensor and driving data, highlighting the need for efficient knowledge -processing systems capable of providing real-time responses. In particular, the ability of autonomous vehicles to explain their decision-making processes in complex driving situations to human users is crucial for ensuring safety and reliability. While existing RAG methods show strong generalization performance across various domains, they have limitations regarding accuracy and consistency in domain-specific contexts. To address this issue, this study proposes LoRA FT-Driver, a fine-tuned sLLM utilizing LoRA. The proposed model adjusts only approximately 0.01% of total parameters, ensuring high accuracy and consistent responses even with limited computational resources. Its lightweight architecture makes it well-suited for deployment in embedded systems such as autonomous vehicles. Qualitative and quantitative experiments conducted using the BDD-X dataset demonstrated that LoRA FT-Driver significantly outperforms RAG-based models. Additionally, extended experiments, such as classification tasks involving traffic violation types, confirmed the model's superior domain transfer capabilities.

Keywords : Autonomous Driving, RAG, Fine-tuning, LoRA, Prompt engineering.

*This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT).(RS-2024-00352792).

[1]Master's Course, Division of Big Data Science, Korea University, 2511 Sejong-ro, Sejong-city, 30019, Korea. E-mail: dayun0405@korea.ac.kr

[2]Professor, Division of Big Data Science, Korea University, 2511 Sejong-ro, Sejong-city, 30019, Korea. E-mail: scheon@korea.ac.kr

[3](Corresponding Author) Ph.D. Candidate, Department of Economics and Statistics, Korea University, 2511 Sejong-ro, Sejong-city, 30019, Korea. E-mail: jooalim@korea.ac.kr

[Received 31 March 2025; Revised 22 April 2025; Accepted 25 April 2025]