

Large Language Model을 활용한 네트워크 최적화 연구 동향

손 석 빙^{*}, 김 중 현[◦], 조 창 식^{*}, 박 수 현^{**}

Trends in Network Optimization Using Large Language Models

Seok Bin Son^{*}, Joongheon Kim[◦], Changsik Cho^{*}, Soohyun Park^{**}

요 약

Large Language Model (LLMs)은 네트워크 최적화, 분산 컴퓨팅, 자율 시스템 등 다양한 분야에서 혁신을 이끌고 있다. 본 논문은 LLM이 네트워크 최적화, 엣지 및 분산 컴퓨팅, 자율 시스템 제어에서 수행하는 역할과 성능향상에 미친 영향을 분석한다. 그러나 LLM의 잠재력에도 불구하고, 보안 취약점, 에너지 효율성, 멀티모달 데이터 통합 등 해결해야 할 과제가 남아 있다. 이를 해결하기 위해 본 논문은 다양한 향후 연구 방향을 제시한다. LLM은 다양한 산업에서 핵심 기술로 자리 잡을 가능성을 보여주며, 지속적인 연구를 통해 그 활용도를 극대화할 수 있을 것으로 기대된다.

키워드 : Large Language Model, 네트워크 최적화, 엣지 및 분산 컴퓨팅, 자율 시스템 제어

Key Words : Large Language Model, Network Management and Optimization, Edge and Distributed Computing, Autonomous Systems Control

ABSTRACT

Large Language Models (LLMs) are at the forefront of innovation across diverse domains, including network optimization, edge and distributed computing, and autonomous systems. This paper examines the critical role of LLMs in enhancing network management and optimization, enabling efficient edge and distributed computing, and advancing autonomous systems and control. It also evaluates their impact on performance improvement across these areas. Despite their significant potential, LLMs face several challenges, such as security vulnerabilities, limited energy efficiency, and inadequate integration of multimodal data. To overcome these limitations, this paper proposes a range of future research directions. With continued research and development, LLMs are poised to become foundational technologies across industries, enabling their full potential to be realized.

* 본 연구는 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No. 2021-0-00766, 신경망 응용 자동생성 및 실행환경 최적화 배포를 지원하는 통합개발 프레임워크 기술개발) 및 (No. RS-2024-00435652, 6GARROW: 6G Ai-native 통합 RAN-Core 네트워크).

◆ First Author : Korea University Department of Electrical and Computer Engineering, lydiasb@korea.ac.kr, 학생회원

◦ Corresponding Author : Korea University Department of Electrical and Computer Engineering, joongheon@korea.ac.kr, 종신회원

* Electronics and Telecommunications Research Institute On-Device Artificial Intelligence Research Division, cscho@etri.re.kr, 정회원

** Sookmyung Women's University Division of software, soohyun.park@sookmyung.ac.kr, 정회원

논문번호 : 202501-019-C-RU, Received January 17, 2025; Revised February 4, 2025; Accepted February 4, 2025

I. 서 론

인공지능 기술은 최근 대형 언어 모델(LLM, Large Language Models)의 등장과 함께 빠른 속도로 발전하고 있다^[1,2]. 기존의 전통적인 기계학습 모델이 특정 작업을 위해 대규모 데이터와 복잡한 특징 추출 과정을 필요로 했던 반면, LLM은 수십억에서 수천억 개의 파라미터를 학습하여 언어적 패턴을 자율적으로 습득한다. 이를 통해 LLM은 최소한의 지시만으로도 다양한 문제를 해결하는 능력을 제공한다^[3]. 이러한 특성은 LLM이 자연어 처리뿐 아니라 다양한 도메인으로 인공지능 활용 범위를 확장하는 데 기여하고 있다^[4,5].

LLM의 핵심 강점은 단순 텍스트 생성 능력을 넘어 선 고도화된 추론 및 문맥 이해 능력에 있다^[6]. 기존 언어 모델은 확률적 예측 기반으로 텍스트를 생성하거나 분류에 활용되었지만^[7], LLM은 단어와 문맥을 동시에 학습하여 복잡한 질문에 대한 정교한 답변과 생성 결과를 제공할 수 있다. 이는 네트워크 관리^[8-13], 엣지 및 분산 컴퓨팅^[14-18], 자율 시스템 제어^[19,20] 등 기술적으로 복잡한 영역에서도 높은 부가 가치를 창출할 잠재력을 지닌다. 실제로 LLM은 음성 인식 개선, 대규모 트래픽 제어, 자율 시스템 작업 계획 등 기존 AI 접근법으로 해결하기 어려웠던 문제들에 대한 효과적인 대안을 제공하고 있다.

현대 네트워크 환경의 복잡성이 증가하며 LLM의 중요성은 더욱 부각되고 있다. 5G/6G 시대의 도래로 실시간 데이터 교환과 지역 시간 최소화 요구가 높아짐에 따라 기존의 규칙 기반 네트워크 관리 방식은 한계에 직면하고 있다^[21,22]. 이에 LLM은 변화하는 환경에 적응하며 효율적인 의사결정을 내릴 수 있는 지능형 알고리즘으로 주목받고 있다. 예를 들어, LLM은 네트워크 관리에서 장애 예측, 작업 분산, 협업 제어 등 다양한 영역에서 학습된 문맥 정보와 지식을 바탕으로 유연하고 효과적인 메커니즘을 제공한다^[12,13]. 자율 시스템에서도 LLM은 자연어 기반의 복잡한 명령을 해석하여 작업 계획으로 변환하고, 상황 변화에 따라 동작을 동적으로 수정하는 데 중요한 역할을 하고 있다^[19,20]. 특히 멀티 에이전트 시나리오에서 에이전트 간 자연어 대화를 지원하여 협업과 부하 분산, 예측 불가능한 상황 변화에 유연하게 대응할 수 있는 가능성을 보여준다.

한편, LLM 도입은 성능향상과 편의성을 제공하지만 몇 가지 기술적 한계도 존재한다. LLM은 자율 시스템과 엣지 컴퓨팅 등 다양한 분야에서 활용 가능성성이 크지만, 여전히 해결해야 할 한계가 존재한다. 우선, 대규모 데이터 학습 과정에서 보안 취약점이 발생할 가능성이

표 1. LLM 응용 분야별 분류
Table 1. Classification of LLM Application Domains

	네트워크 최적화	엣지 및 분산 컴퓨팅	자율 시스템 제어
목적	LLM을 통신 네트워크 설계, 트래픽 관리, 자원 최적화 등 네트워크 문제 해결에 활용	LLM을 엣지 및 클라우드 환경에서의 작업 분산 처리와 자원 활용 최적화에 적용	LLM을 로봇 및 자동화 시스템의 작업 계획, 협업, 제어에 활용
논문	[8], [9], [10], [11], [12], [13]	[14], [15], [16], [17], [18]	[19], [20]

있으며, 막대한 연산 자원과 전력을 소모해 효율성이 제한적이다. 또한, 텍스트 외에 이미지, 음성, 센서 데이터를 통합적으로 처리하는 멀티모달 데이터 처리 능력이 부족하며, 일반화 능력 부족, 비결정성 문제, 시뮬레이션과 현실 간 격차와 같은 자율 시스템 내 협업 및 상호 운용성에서도 한계를 보인다. 이러한 문제들은 LLM의 잠재력을 완전히 발휘하는 데 장애가 되고 있어 이를 극복하기 위한 연구가 필요하다^[23-31].

따라서 본 논문은 네트워크 관리 최적화, 엣지 및 분산 컴퓨팅, 자율 시스템이라는 세 가지 주요 영역을 중심으로 LLM이 어떤 방식으로 응용되고 있으며, 어떠한 효과를 가져오는지에 대해 종합적으로 살펴본다. 또한 이러한 LLM을 활용할 수 있는 라이브러리에 대해서 조사하고, 각 분야에서 LLM 적용 시 직면하는 기술적 도전과제 및 향후 연구 방향을 정리함으로써, 효율적인 LLM 활용 방안을 제시하고자 한다.

본 논문에서 2장은 LLM의 개념 및 작동 원리에 대해서 살펴보았다. 3장은 LLM을 통한 네트워크 관리 최적화 연구 동향을, 4장은 엣지 및 분산 컴퓨팅에서의 LLM 활용 및 연구 성과를, 5장은 자율 시스템과 제어 분야의 LLM 적용 사례를 다룬다. 또한 6장에서는 LLM을 지원하는 라이브러리에 대해서 기술하였다. 그리고 7장에서는 이러한 연구 결과를 종합하여 향후 연구 과제를 논의하고, 8장에서 결론을 제시한다.

II. LLM의 개념 및 작동원리

LLM은 대량의 비지도 학습 데이터를 기반으로 사전 학습된 딥러닝 모델로, 자연어의 문맥적 의미를 학습하여 인간과 유사한 수준의 텍스트 생성과 이해를 수행한다. 이 모델들은 언어 간의 복잡한 의존 관계를 효과적으로 학습하며, 이를 기반으로 다양한 작업에서 우수한

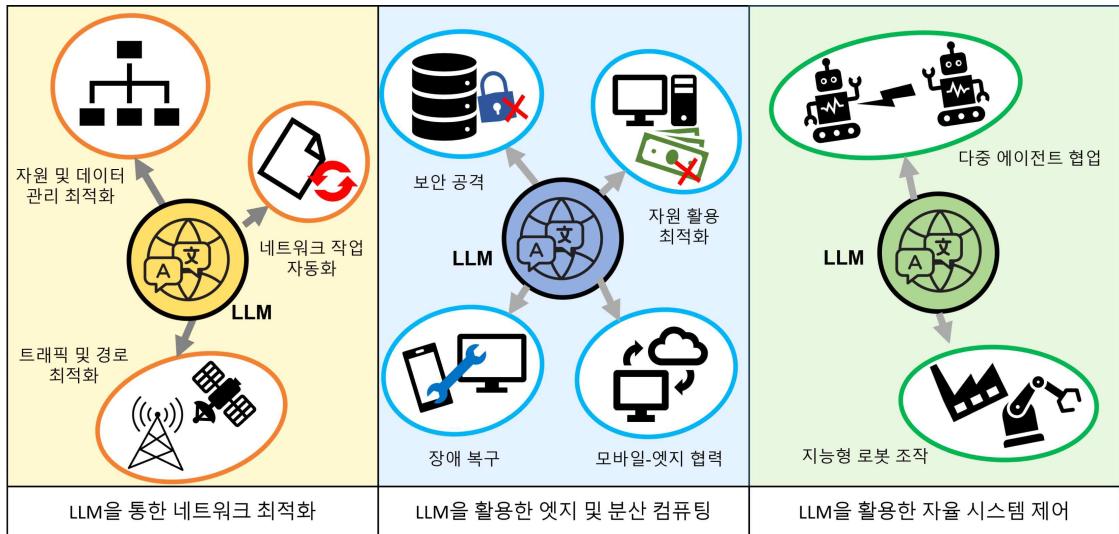


그림 1. LLM을 활용한 네트워크 최적화, 엣지 및 분산 컴퓨팅, 자율 시스템 제어 분야의 주요 응용 사례

Fig. 1. Key Application Cases of LLM-based Network Optimization, Edge and Distributed Computing, and Autonomous System Control

성능을 발휘한다^[12]. 대표적인 모델인 Generative Pre-trained Transformer (GPT)와 Bidirectional Encoder Representations from Transformers (BERT)는 각각 생성과 이해에 특화된 구조를 가진다^[8,10]. LLM의 핵심은 다음과 같은 특징을 포함한다. 먼저, 대규모 데이터 학습을 통해 다양한 언어 패턴과 뉴앙스를 학습한다. 또한, 사전 학습(pre-training) 후 특정 작업에 대한 미세 조정(fine-tuning)을 통해 특화된 성능을 제공한다^[16]. 그리고, 긴 문맥을 동적으로 해석하여 문맥적 이해를 증대시키고, 확률적 생성 방식으로 다양한 텍스트 결과를 생성한다.

LLM의 작동 원리는 학습 단계와 텍스트 생성 과정으로 나눌 수 있다. 학습 단계에서 모델은 대규모 데이터로부터 언어의 일반적 구조와 패턴을 학습하며, 미세 조정을 통해 특정 작업이나 도메인에 특화된 능력을 갖춘다^[18]. 트랜스포머 아키텍처는 LLM의 핵심 요소로, 다중 헤드 어텐션 메커니즘을 활용해 입력 데이터 내의 단어 간 관계를 학습한다^[16]. 이는 단어 간의 장거리 의존성을 모델링하고 병렬 처리로 효율성을 높인다. 텍스트 생성 과정에서는 모델이 주어진 입력을 바탕으로 다음 단어를 확률적으로 생성하며, 반복적으로 이를 수행하여 최종 출력을 완성한다^[8,12]. 이 과정에서 온도 조정이나 샘플링 기법을 사용해 다양성과 창의성을 조절할 수 있다. 또한, LLM은 일반화된 언어 이해를 통해 다양한 주제와 작업에 적응할 수 있는 유연성을 보유한다^[10].

이와 같은 특징을 가진 LLM은 텍스트 생성, 번역,

요약, 감정 분석과 같은 자연어 처리 작업은 다양한 분야에서 활용되고 있다. 본 논문에서는 네트워크 최적화, 엣지 및 분산 컴퓨팅, 자율 시스템 제어에서의 LLM의 역할에 초점을 두었다.

III. LLM을 통한 네트워크 최적화

네트워크 최적화는 디지털 인프라의 안정성과 성능을 보장하기 위해 필수적인 기술 분야로, 최근 LLM의 도입을 통해 크게 발전하고 있다^[8,13]. LLM은 대규모 데이터를 학습하여 네트워크 운영에서 발생하는 복잡한 문제를 해결하는 데 뛰어난 성능을 발휘하며, 트래픽 관리, 자원 최적화, 네트워크 설계 등의 다양한 응용에서 핵심적인 역할을 수행하고 있다. 본 장에서는 네트워크 최적화의 주요 하위 분야로 표 2와 같이 자원 및 데이터 관리 최적화, 트래픽 및 경로 최적화, 네트워크 작업 자동화를 선정하고, 각 분야에서의 LLM이 수행하는 역할에 관해서 기술하였다.

3.1 자원 및 데이터 관리 최적화

자원 및 네트워크 최적화는 대역폭, 연산능력 등과 같은 네트워크 자원을 효율적으로 사용하는 데 중점을 둔다. 이는 데이터 처리량을 줄이고 네트워크 자원을 절약하여 비용 절감과 성능향상을 동시에 달성하는 것을 목표로 한다. 특히, 제한된 컴퓨팅 환경에서 음성 인식 정확도를 높이기 위해 LLM을 활용하여 단어 오류율(WER; Word Error Rate)을 많이 감소시킨 연구가

표 2. LLM을 활용한 네트워크 최적화 연구 분석

Table 2. Analysis of Network Optimization Research Leveraging LLM

논문	세부 응용 분야	연구목표	LLM 활용방식	주요 기여
[8]	자원 및 관리 최적화	네트워크 자원 관리 최적화	LLM 임베딩을 활용한 음성 인식 및 자원 최적화	제한된 환경에서도 효율적인 자원 활용 및 정확도 향상
[9]		네트워크 작업의 도메인 지식 학습	저랭크 적응을 통한 LLM 최적화	네트워크 데이터의 효율적 학습과 비용 절감
[10]		의료데이터 활용 최적화	멀티모달 학습 기반 의료 AI 모델	팬데믹 상황에서 제한된 데이터 활용
[11]	트래픽 및 경로 최적화	통합 네트워크(ISATN) 최적화	트래픽 분석 및 최적 경로 설정	네트워크 자원 활용 및 안정성 강화
[12]	네트워크 작업 자동화	네트워크 작업 자동화	3GPP 문서 분류 및 작업 그룹화	네트워크 설계 자동화 및 분류 성능향상
[13]		ChatNet을 활용한 네트워크 효율화	LLM 기반 작업 세분화 및 실행 계획 수립	네트워크 작업 효율성 개선

주목받고 있다^[8]. 이 연구에서는 LLM의 텍스트 임베딩(Embedding)을 자동 음성 인식(ASR; Automatic Speech Recognition)의 학습 단계에 멀티태스크 학습으로 통합하였다. 텍스트 임베딩은 자연어 처리에서 단어를 실수 벡터로 변환하여 컴퓨터가 이해할 수 있는 형태로 표현하는 기술로, 단어의 의미와 같은 다양한 정보를 포함한다. 이를 ASR 모델에 적용함으로써, 모델이 단순히 음성을 텍스트로 변환하는 것에서 나아가 사전 학습된 언어 임베딩을 예측할 수 있도록 확장되었다. 구체적으로, [8]에서는 기존 음성 인식 손실 함수에 LLM 임베딩 예측 손실 함수를 추가하고, LLM 임베딩을 예측하는 회귀 네트워크(Regression Network)를 통합하였다. 이를 통해 모델이 언어적 지식을 효과적으로 학습할 수 있도록 설계되었다. 또한, 제한된 컴퓨팅 환경을 고려하여 LLM 관련 컴포넌트는 학습 단계에서만 사용하며, 추론 단계에서는 제거하여 추가적인 계산 비용 없이 모델이 동작할 수 있도록 구현하였다. 또 다른 연구에서는 저차원 적응(DL-LRNA)을 활용하여 LLM의 전체 파라미터를 조정하지 않고도 네트워크 작업에 필요한 지식을 효율적으로 학습하는 방법을 제안하였다^[9]. 이 방법은 멀티모달 인코더와 네트워킹 헤드를 사용하여 네트워크 데이터를 LLM이 이해 가능한 형태로 변환하고, LLM은 사전 학습된 지식을 활용하여 네트워크 도메인에서 일반화된 솔루션을 제공하도록 설계되었다. 이를 통해 학습 데이터 분포와 다른 테스트 환경에서도 뛰어난 일반화 성능을 달성할 수 있었다. 특히, LLM의 파라미터를 동결하고 추가로 소규모 저랭크 행렬을 학습하여 네트워크 작업에 필요한 도메인 지식을 습득함으로써, 추가적인 설계 비용 없이 리소스를 절감하는 동시에 높은 성능을 유지하는 방법론을 제

시하였다. 또 다른 연구에서는 LLM을 활용해 팬데믹 초기와 같은 데이터 부족 상황에서도 효과적인 의료 AI 모델을 제안하며, 자원 및 데이터 관리 최적화에 중점을 두었다^[10]. LLM은 방대한 텍스트 데이터에서 학습한 언어적 맥락을 활용해 의료 텍스트 임베딩을 생성하며, 이를 의료 이미지와 결합해 데이터의 효율적 활용을 가능하게 한다. Patient-Level Contrastive Learning (PCL)을 통해 이미지 데이터를 학습하고, Masked Language Modeling (MLM)과 Sentence Reconstruction (SR)으로 라벨 없는 텍스트 데이터를 학습함으로써, 적은 양의 데이터로도 높은 성능을 달성한다. 연구 결과, 적은 라벨 데이터만으로도 기존 전체 데이터를 사용하는 방법과 유사한 성능을 기록했으며, COVID-19 진단에서는 기존 대비 성능향상을 보였다. 이 연구는 LLM을 활용해 의료데이터와 자원을 최적화하여, 제한된 환경에서도 신속하고 정확한 대응이 가능함을 입증하였다.

3.2 트래픽 및 경로 최적화

LLM을 활용한 트래픽 및 경로 최적화는 실시간 트래픽 분석을 통해 네트워크 자원을 동적으로 배분하고 최적 경로를 설정하여 네트워크의 안정성과 효율성을 극대화하는 것을 목표로 한다. 최근에는 그림 2와 같이 위성-항공-지상 통합 네트워크(ISATN; Integrated Satellite-Aerial-Terrestrial Network)에서 데이터를 분석하여 5G 및 6G 환경에서 안정성과 효율성을 강화하려는 연구가 등장하였다^[11]. 이 연구에서는 LLM을 활용하여 실시간 네트워크 데이터와 과거 데이터를 분석하고, 트래픽 패턴을 예측함으로써 자원 활용을 최적화하여 네트워크 효율성을 향상하는 방법을 제안하였다.

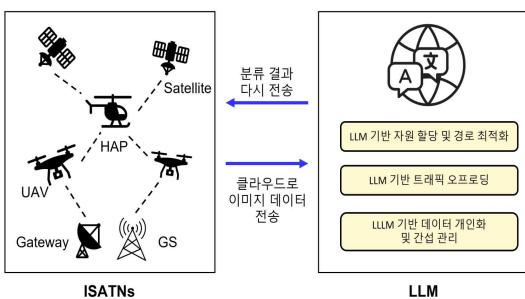


그림 2. ISATNs에서의 LLM 적용 연구 [11]
Fig. 2. LLM Application Studies in ISATNs [11]

또한, 네트워크 장애와 성능 병목 현상을 사전에 예측하고 완화하는 방안을 제시하였다. 구체적으로, ISATN에서 LLM은 사용자 간 최적 전력 할당과 간섭 관리 정책을 예측 및 생성하여 비직교 다중 접속(NOMA; Non Orthogonal Multiple Access)의 효율성을 극대화하는데 활용되었다. 또한, 스펙트럼 센싱 데이터를 분석하여 주파수 대역 활용 패턴을 학습하고, 실시간으로 최적 주파수 대역을 추천하는 시스템을 구현하였다. 더 나아가, LLM은 네트워크 트래픽 데이터와 환경 요인을 분석하여 최적의 라우팅 경로를 생성하고, 실시간으로 경로를 조정함으로써 네트워크 트래픽 관리 효율을 개선하고 자원 활용을 최적화하였다. 이를 통해 실시간 문제 해결과 트래픽 분산이 가능해져 서비스 중단을 최소화하는 성과를 거두었다.

3.3 네트워크 작업 자동화

LLM을 활용한 네트워크 작업 자동화는 네트워크 설계 및 관리 작업을 자동화하여 운영 효율성을 높이고 사람의 개입을 최소화하는 것을 목표로 한다. 최근 연구에서는 3GPP 기술문서를 LLM을 활용해 통신 도메인

에 맞게 작업 그룹으로 분류하여 자율적이고 적응 가능한 통신 네트워크 설계를 지원하는 방안을 제안하였다 [12]. 이 연구는 LLM의 사전 학습된 지식을 활용하여 3GPP 기술문서 내 단어의 양방향 문맥을 이해하고,통신 도메인 단어의 의미를 학습하여 작업 그룹을 정확히 식별할 수 있도록 설계되었다. 이를 통해 높은 분류 성능을 달성했으며, 파라미터 수를 줄인 상태에서도 기존 모델과 유사한 성능을 유지하며 우수성을 입증하였다. 또 다른 연구에서는 ChatNet을 제안하여 네트워크 작업을 세부적인 실행 계획으로 분해하고, 명령어 생성과 시뮬레이터 연계를 통해 작업 수행 효율을 높이고자 하였다 [13]. 이 과정에서 인터넷 데이터를 사용하여 LLM을 대규모로 학습하여 일반적인 이해 능력을 개발한 뒤, 네트워크 도메인 데이터를 활용해 네트워크 작업에 최적화된 모델로 개선하였다. ChatNet은 네트워크 상태와 제약조건을 바탕으로 작업을 계획하고 실행하면서, 외부 도구와의 연계를 통해 작업 효율성을 더욱 높였다. 이를 통해 LLM이 네트워크 작업 효율성을 향상하고 인간의 개입을 최소화할 수 있음을 입증하였다.

IV. LLM을 통한 엣지 및 분산 컴퓨팅

엣지 및 분산 컴퓨팅은 현대의 5G/6G 네트워크 환경에서 필수적인 요소로, 데이터 처리와 자원 활용을 최적화하며 사용자 경험을 향상하는 데 중점을 둔다 [32, 33]. LLM은 자연어 처리 능력과 멀티모달 데이터 분석 역량을 활용하여 엣지 컴퓨팅과 클라우드 컴퓨팅 간 작업 분산처리, 자원 최적화, 사용자 요구사항의 실시간 처리 등 다양한 응용 분야에서 혁신적인 변화를 이끌고 있다 [14-18]. 본 장에서는 LLM이 엣지 및 분산 컴퓨팅에서 수행하는 주요 역할을 표 3과 같이 보안 공격 및 장애 복구, 자원 활용 및 최적화, 모바일-엣지 협력의 세 가지

표 3. LLM을 활용한 엣지 및 분산 컴퓨팅 연구 분석

Table 3. Analysis of Edge and Distributed Computing Research Leveraging LLM

논문	세부 응용 분야	연구목표	LLM 활용방식	주요 기여
[14]	보안 공격 및 장애 복구	연합학습의 보안 취약점 탐구	FILM을 통해 민감 데이터 복구 공격 수행	연합학습 환경에서 프라이버시 취약점 분석
[15]		엣지 네트워크 장애 복구	LLM 기반 장애 복구 경로 생성	장애 복구 효율성 및 안정성 강화
[16]	자원 활용 최적화	자원 활용 최적화	사용자 요청 분석 및 작업 분산	5G 환경에서 효율적인 자원 관리
[17]		클라우드-엣지 컴퓨팅 자원 분배	작업 오프로딩 및 자원 최적화	액티브 추론 기반 자원 최적화
[18]	모바일-엣지 협력	모바일-엣지 협력 모델 설계	LLM 기반 작업 분배 및 협력	지연 시간 감소 및 모바일 기기 자원 활용

측면에서 논의한다.

4.1 보안 공격 및 장애 복구

LLM은 엣지 및 분산 컴퓨팅 환경에서 다양한 문제를 해결하거나 새로운 보안 취약점을 밝혀내는데 중요한 역할을 한다. 최근 연구에서는 연합 학습(FL; Federated Learning) 환경에서 발생할 수 있는 데이터 유출 및 보안 취약점을 강조한 공격 기법과 엣지 네트워크에서의 장애 복구 방안을 제안하였다. FILM (Federated Inference Leakage Mitigation)은 연합학습 환경에서 클라이언트의 민감 정보를 복원할 수 있는 새로운 공격 기법으로, FL의 보안 취약점을 부각한 사례이다^[14]. FILM은 LLM의 메모리 능력과 언어 모델의 확률 분포를 활용하여 학습된 민감 데이터를 재구성한다. 이를 통해 단어와 문장의 순서를 복원할 수 있으며, 큰 배치 크기인 최대 128개의 문장에서도 높은 정확도로 데이터를 복구할 수 있음을 입증하였다. 이 연구는 FL 환경에서 보안 강화와 민감 데이터 보호를 위한 추가적인 대책 마련의 필요성을 제기하였다. 한편, 엣지 네트워크 환경에서는 LLM을 활용하여 장애 복구와 네트워크 안정성을 향상하는 연구가 주목받고 있다. LLM-ENFT 시스템은 LLM 기반 자율 에이전트를 사용하여 네트워크 장애를 감지, 분석, 복구하며, 네트워크 트래픽 상태와 장애 지표를 바탕으로 최적의 복구 경로를 생성한다^[15]. LLM-ENFT는 사전 학습된 도메인 지식과 정의된 도구를 활용하여 장애 상황에서도 네트워크 흐름을 신속히 복원하고, 관리 및 복구 효율성을 극대화하는 데 중점을 둔다. 이를 통해 엣지 네트워크 환경에서 서비스 중단을 최소화하고, 사용자 경험을 유지하며, 안정성을 강화할 수 있음을 입증하였다.

4.2 자원 활용 최적화

엣지 및 분산 컴퓨팅 환경에서 LLM은 네트워크와 클라우드 간의 자원을 효율적으로 분배하고 최적화하여 연산 비용을 절감하고 시스템 성능을 향상하는 데 핵심적인 역할을 한다. 특히, 5G 네트워크와 같은 복잡한 환경에서 LLM은 사용자 요구를 분석하여 태스크를 세분화하고, 엣지 서버와 클라이언트 간의 작업 부하를 효율적으로 분산함으로써 자원 활용을 극대화한다^[16]. LLM은 사용자의 자연어 요청을 분석하여 이를 구체적인 하위 작업으로 분해하고, 엣지 서버의 자원 상태를 실시간으로 감시하여 작업 오프로딩과 자원 할당을 동적으로 최적화한다. 이를 통해 작업 지향적 데이터 압축을 수행하여 전송 비용을 절감하며, GPT 기반 템플릿 코드 생성 및 연합학습 설정 자동화를 통해 민감 데이터

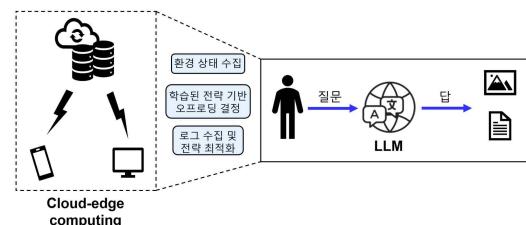


그림 3. Cloud-edge 컴퓨팅에서의 LLM 적용 [16]
Fig. 3. Application of LLM in Cloud-Edge Computing [16]

를 보호하면서도 사용자 요구를 충족시키는 서비스를 제공한다. 이러한 자원 활용 전략은 연산 비용을 줄이고 대기 시간을 최소화하여 네트워크의 효율적인 운영을 가능하게 합니다. 또한, 자원 제한이 큰 엣지 컴퓨팅 환경에서 지연 시간을 최소화하고 추론 정확도를 극대화하기 위해 액티브 추론 기반 알고리즘을 활용한 자원 최적화 연구도 주목받고 있다^[17]. 액티브 추론 기반 알고리즘은 보상 없는 기준을 활용하여 기존 강화학습보다 빠르게 수렴하며 더 높은 일반화 성능을 제공한다. 이 알고리즘은 그림 3과 같이 LLM이 처리하기 어려운 대규모 추론 작업을 엣지 서버로 분산하여 처리함으로써 모바일 장치에서의 자원 제약 문제를 해결하고, 엣지 서버의 자원을 최대한 활용하여 제한된 환경에서도 LLM 기반 서비스를 제공할 수 있도록 지원한다.

4.3 모바일-엣지 협력

LLM은 모바일 기기와 엣지 서버 간의 협력을 통해 분산된 작업을 효율적으로 처리하고, 모바일 환경의 제약을 극복하는 데 중요한 역할을 한다^[18]. 이를 위해 모바일 기기에는 소규모 LLM을 배치하여 실시간 자료수집과 기본적인 의사결정을 담당하도록 하고, 엣지 서버에는 대규모 LLM을 배치하여 복잡한 연산과 글로벌 데이터를 활용한 고도화된 의사결정을 수행하도록 설계되었다. 모바일-엣지 협력 모델은 각 환경의 강점을 활용하여 효율성을 극대화한다. 모바일 LLM은 사용자 환경에 맞춘 개인화된 정보와 의사결정을 실시간으로 제공하며, 엣지 LLM은 모바일 에이전트를 지원하여 더욱 복잡한 작업을 처리하고 글로벌 데이터 기반 분석을 수행한다. 이 과정에서 모델 캐싱 알고리즘(AoT; Age of Thought)과 단계별 추론 방식(CoT; Chain of Thought)과 같은 최적화 기법이 적용되었다. 모델 캐싱 알고리즘을 통해 자주 사용되지 않거나 중요도가 낮은 데이터를 제거함으로써 처리 속도를 높이고 네트워크 비용을 절감하고, 단계별 추론 방식을 통해 복잡한 문제 해결 능력을 향상해 성능을 극대화한다. 결과적으로, 모바일과 엣지의 긴밀한 협력은 실시간 데이터 처리 속도

를 높이고 대기 시간을 줄이며, 모바일 기기의 자원 제약을 효과적으로 극복할 수 있도록 지원한다. 이러한 협력은 모바일 환경에서의 효율성과 사용자 경험을 향상하는 데 있어 중요한 역할을 수행하였다.

V. LLM을 활용한 자율시스템 제어

자율 시스템 제어 분야에서 LLM의 도입은 로봇의 작업 처리와 협업 능력을 혁신적으로 변화시키고 있다 [19,20]. LLM은 자연어 이해와 추론 능력을 통해 복잡한 작업 환경에서의 적응성과 효율성을 강화하며, 특히 새로운 개념 학습 및 실시간 피드백 통합을 가능하게 한다. 이를 통해 자율 시스템은 기존 방식의 한계를 극복하고 더욱 정교하고 유연한 제어를 실현할 수 있다. 본 장에서는 표 4와 같이 지능형 로봇 조작, 다중 에이전트 협업 측면에서 LLM이 적용된 연구를 기술하고자 한다.

5.1 지능형 로봇 조작

GraspGPT는 그림 4와 같이 LLM을 활용해 지능형 로봇 조작의 새로운 가능성을 제시한 모델이다 [19]. LLM은 자연어 입력을 기반으로 새로운 작업과 물체에 대한 개념을 생성하며, 이를 로봇의 물체 조작 과정에 통합한다. GraspGPT는 3D 포인트 클라우드 데이터와 LLM에서 생성된 개념 설명을 결합하여 학습 데이터에 포함되지 않은 새로운 물체나 작업에 대해 조작 자세 (grasp pose)를 예측한다. 특히, TaskGrasp 데이터셋을 활용해 물체와 작업 간의 의미적 연관성을 분석하고, 이를 기반으로 로봇이 적응적으로 조작 자세를 결정할 수 있도록 지원한다. 이 접근방식은 기존의 데이터 의존적 로봇 조작과 달리, LLM의 추론 능력을 통해 새로운

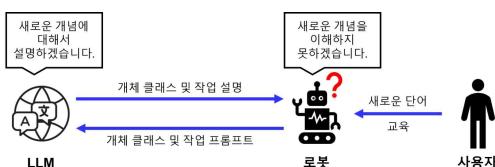


그림 4. GraspGPT 동작 방식 [19]
Fig. 4. Operation of GraspGPT [19]

표 4. LLM을 활용한 자율시스템 및 제어 연구 분석

Table 4. Analysis of Autonomous System and Control Studies Leveraging LLM

논문	세부 응용 분야	연구목표	LLM 활용방식	주요 기여
[19]	지능형 로봇 조작	로봇 조작 및 작업 적응	GraspGPT를 활용한 물체 조작	새로운 작업 및 환경에서 적응성 향상
[20]	다중 에이전트 협업	다중 로봇 협업 최적화	LLM 기반 대화 및 서브 태스크 계획	동적이고 유연한 다중 에이전트 협업

환경에서도 효과적인 작업 수행을 가능하게 한다. GraspGPT는 실험 결과에서 새로운 작업과 물체에 대한 높은 정확도와 성공률을 기록하며, LLM이 지능형 로봇 조작에서 핵심적인 역할을 할 수 있음을 입증했다.

5.2 다중 에이전트 협업

RoCo는 LLM을 활용하여 다중 로봇 협업의 효율성을 극대화한 시스템으로, 자율 시스템 제어 분야에서 새로운 가능성을 제시한다 [20]. 각 로봇은 고유의 LLM 에이전트를 가지며, 자연어 대화를 통해 작업 전략을 논의하고 하위 작업을 분배하며, 환경에서 실시간 피드백을 반영해 계획을 지속해서 개선한다. 이를 통해 기존의 단일 로봇 중심 제어 방식에서 벗어나 다중 에이전트 간의 동적이고 유연한 협업을 구현한다. RoCo의 핵심은 다중 에이전트 대화, 서브 태스크 계획, 그리고 동작 계획의 세 가지 단계로 이루어진다. 먼저, LLM 에이전트는 각 로봇의 관찰 정보와 역할에 기반하여 대화를 통해 작업 전략을 수립한다. 대화 종료 시 각 로봇이 수행할 하위 작업과 경로를 포함한 계획이 생성되며, 이 과정에서 로봇 간의 상호작용을 극대화한다. 이후 서브 태스크 계획 단계에서는 LLM이 생성한 작업 계획이 환경 검증을 통과해야 실행 가능하며, 실패 시 피드백을 기반으로 대화를 재개하고 계획을 수정한다. 마지막으로, 서브 태스크 계획이 검증되면 다중 로봇 표본 기반 경로 생성 알고리즘(RRT; Rapidly-exploring Random Tree)을 활용해 최적 경로를 탐색하며, 각 로봇이 자신의 목표 구성에 도달할 수 있도록 지원한다. 실험 결과, RoCo는 벤치마크에서 높은 성공률을 기록하였으며, LLM 대화 에이전트를 통해 각 로봇의 고유 제약조건을 효율적으로 반영하는 데 성공했다. 이를 통해 LLM을 활용할 경우 다중 로봇 협업 시스템이 더욱 정교하고 효율적으로 작동할 수 있음을 입증했다.

VI. LLM 라이브러리

LLM의 활용이 증가함에 따라, 이를 지원하는 다양한 라이브러리들이 개발되고 있다. 이 중 LangChain,

DeepSpeed, FAISS (Facebook AI Similarity Search)는 각각의 특성과 기능으로 주목받고 있다. LangChain은 LLM 기반 애플리케이션의 파이프라인을 설계하고 다양한 데이터 소스를 통합하여 워크플로우를 자동화하는 프레임워크이다. LangChain의 핵심 기능은 데이터 통합, 메모리 관리, 외부 API와의 연동 등을 포함하며, 이를 통해 PDF 문서 처리, 대화형 에이전트 개발, 질의응답 시스템과 같은 다양한 응용 사례에서 활용되고 있다^[34]. 예를 들어, LangChain을 사용한 고객 서비스 자동화 연구에서는 맞춤형 LLM을 구축해 응답의 품질과 속도를 동시에 개선한 사례가 있다^[35].

DeepSpeed는 Microsoft에서 개발한 대규모 딥러닝 모델의 학습 및 추론 최적화 프레임워크로, 특히 ZeRO (Zero Redundancy Optimizer) 기술을 통해 대규모 모델 학습 시 메모리 사용량을 최소화한다. DeepSpeed는 모델 병렬화, 데이터 병렬화, 파이프라인 병렬화를 결합하여 학습 속도를 크게 향상시키며, 이러한 기술은 초대 규모 언어 모델의 훈련에서 활용되었다^[36]. 또한, DeepSpeed-FastGen은 LLM의 고속 텍스트 생성 기능을 제공하여 기존 시스템 대비 최대 2.3배의 처리량 향상과 2배 낮은 자연 시간을 달성한 것으로 보고되었다^[37].

FAISS는 벡터 데이터를 기반으로 유사성 검색과 클러스터링을 수행하는 라이브러리로, LLM이 생성한 임베딩 벡터를 효율적으로 처리하는 데 사용된다. FAISS는 GPU와 CPU를 활용해 대규모 벡터 데이터를 빠르게 검색할 수 있도록 최적화되어 있으며, 검색 엔진, 추천 시스템, 문서 검색 등 다양한 애플리케이션에서 활용되고 있다. 이러한 기술적 강점은 FAISS가 대규모 데이터 세트에서 벡터 유사성 검색을 효율적으로 수행할 수 있도록 지원한다^[38].

결론적으로, LangChain, DeepSpeed, FAISS는 LLM 생태계에서 각각 중요한 역할을 담당하며, LLM의 확장성과 응용 가능성을 크게 확장하고 있다. LangChain은 워크플로우 자동화, DeepSpeed는 대규모 모델 학습 최적화, FAISS는 벡터 검색 효율화에 중점을 두고 있어, 이들 라이브러리는 LLM 기술의 발전과 응용 확장을 지원하는 데 핵심적인 역할을 수행할 것이다.

VII. LLM을 활용한 향후 연구 방향

앞서 LLM이 네트워크 관리 최적화, 엣지 및 분산 컴퓨팅, 자율 시스템 제어 등에서 다양한 응용 가능성과 기술적 진보를 가져오는 것을 살펴보았다. 또한 다양한 라이브러리를 활용하여 LLM의 연구가 활발히 진행되

고 있음을 확인하였다. 이 같은 연구성과를 바탕으로, 향후 연구 방향을 제안한다.

7.1 LLM의 보안 강화 및 프라이버시 보호

LLM은 방대한 양의 데이터를 학습하는 과정에서, 모델 내부에 특정 텍스트나 개인 정보를 저장하여 보안 및 프라이버시를 침해할 가능성이 있다. 최근에는 LLM으로부터 개인 정보를 추출할 수 있음을 보여, 기존 보안 대책으로는 완전히 해결하기 어려운 새로운 취약점이 존재함을 시사하였다^[14,23]. 따라서 이를 방지하기 위해서는 LLM 학습 과정에서의 비식별화(Anonymization), 차등 프라이버시(Differential Privacy), 암호화 기반 기법 등 다각적인 접근이 필요하다. 또한 LLM의 자연어 이해·생성 능력을 활용하여 공격 시그니처, 악성 코드 패턴 등과 같은 보안 위협을 실시간으로 탐지하고 자동 대응하는 시스템 구현이 요구된다. 이를 통해 네트워크·엣지 환경에서 발생할 수 있는 사이버 공격에 신속히 대응할 수 있을 것이다.

7.2 LLM의 보안 강화 및 프라이버시 보호

대규모 파라미터로 인해 LLM은 막대한 연산 자원과 전력을 소모한다. 따라서 모델 경량화, 분산 학습 기법을 통해 에너지 효율을 높이는 연구가 필요하다. 이에 따라 최근에는 MoE(Mixture of Experts) 방식을 제안하여 동일한 파라미터 규모에서도 일부 전문가(Expert)만 활성화해 연산을 수행함으로써 계산 비용을 효과적으로 낮출 수 있는 연구가 등장하였다^[24]. 이러한 연구를 바탕으로 LLM을 엣지·분산 환경에서의 적용 가능성을 높이고, 모바일 장치나 센서 네트워크 같은 자원 제약 환경에도 유리하도록 해야한다. 또한, 양자화·프루닝·지식 증류 등의 방법을 통해 경량화된 LLM 모델을 개발하여 엣지 장치에서도 추론이 가능하도록 설계하는 연구가 필요하다. 이를 통해 에너지 소비를 줄이면서도, 실시간 의사결정과 응답 속도를 높일 수 있을 것이다.

7.3 멀티모달 데이터 처리 및 통합

네트워크 트래픽 로그, 텍스트, 음성, 이미지, 센서 데이터 등 다양한 형태의 정보가 실시간으로 생산되는 환경에서, LLM을 텍스트뿐만 아니라 멀티모달 데이터를 통합적으로 처리할 수 있도록 확장할 필요가 있다^[25]. 멀티모달 데이터를 활용한 BLIP-2는 이미지 인코더와 대규모 언어 모델을 연결해 텍스트와 이미지를 동시에 처리하는 대표적인 예시다^[26]. 이를 엣지·분산 컴퓨팅 환경에 적용하면, 센서 데이터나 카메라 영상 등

시각 정보와 텍스트 정보를 결합해 더욱 정교한 분석과 의사결정이 가능해질 것이다. 특히 엣지 장치나 IoT 환경에서 들어오는 다양한 멀티모달 데이터를 빠르게 처리하기 위해서는, 효율적인 스트리밍 처리 기술과 멀티모달 융합 모듈의 연구가 필요하다. 지역 최소화, 동적인 데이터 품질 관리 등이 중요한 이슈로 부상할 것이다.

7.4 자율 시스템의 협업 및 상호 운용성 향상

LLM은 자율 시스템의 협업과 상호 운용성을 강화하는 데 중요한 역할을 할 수 있는 강력한 도구이다. LLM은 자연어 기반 명령의 이해와 작업 계획 생성에 뛰어난 성능을 보여주며, 복잡한 작업을 다중 에이전트 시스템에서 효율적으로 분해하고 할당할 수 있는 가능성은 제공한다^[27]. 그러나 현재 LLM의 적용에는 몇 가지 한계가 존재한다. 일반화 능력 부족, 명령의 애매성 처리의 어려움, 비결정성 문제, 그리고 시뮬레이션과 현실 간의 격차가 대표적인 문제들이다. 이러한 문제는 복잡한 다중 로봇 환경에서의 효과적인 협업 및 자원 분배를 어렵게 만들며, 안정적이고 일관된 결과를 생성하는 데 장애가 된다. 따라서 LLM의 잠재력을 극대화하기 위해, 실제 환경의 물리적 제약을 반영한 시뮬레이션 플랫폼과 실시간 학습 기능을 개발하여 현실 적용성을 높이는 연구가 필요하다. 이질적인 에이전트 간 자원 분배와 팀 구성을 최적화할 수 있는 강화학습을 기반으로^[28-31], LLM의 강점을 활용한 협업 프레임워크를 설계함으로써 다중 로봇 간 효율적인 협력을 지원할 수 있다. 이러한 방향으로의 발전은 자율 시스템에서 LLM의 활용도를 극대화하고, 다양한 환경과 자원 제약 조건에서도 효과적으로 작동할 수 있는 시스템을 구현하는 데 기여할 것이다.

VIII. 결 론

본 논문에서는 LLM이 네트워크 관리 최적화, 엣지 및 분산 컴퓨팅, 자율 시스템 등에서 혁신적인 역할을 수행하며 다양한 응용 가능성을 제시함을 논의하였다. LLM은 자원 및 트래픽 최적화, 장애 복구, 다중 에이전트 협업 등에서 성능향상을 이끌었으며, 복잡한 문제 해결과 실시간 적응성을 통해 기존 기술의 한계를 극복하고 있다. 그러나 보안 취약점, 에너지 효율성, 멀티모달 데이터 통합 등 해결해야 할 과제도 여전히 존재한다. 이를 위해 보안 강화, 분산형 복구 메커니즘, 에너지 효율적 설계, 멀티모달 데이터 분석 등 구체적인 연구 방향을 제안하였다. 결론적으로, LLM은 다양한 산업

에서 핵심적인 역할을 수행할 잠재력을 지니고 있으며, 지속적인 연구와 개선을 통해 더 신뢰할 수 있는 지능형 솔루션으로 자리 잡을 것으로 기대된다.

References

- [1] Z. Wei, et al., “Emergent abilities of large language models,” *Trans. Machine Learn. Res.*, pp. 1-30, Aug. 2022. (<https://openreview.net/forum?id=yzkSU5zdwD>)
- [2] Y. Chang, et al., “A survey on evaluation of large language models,” *ACM Trans. Intell. Syst. Technol.*, vol. 15, no. 3, pp. 1-45, Mar. 2024. (<https://doi.org/10.1145/3641289>)
- [3] T. Brown, et al., “Language models are few-shot learners,” *Advances in NeurIPS*, vol. 33, pp. 1877-1901, Virtual, Dec. 2020.
- [4] Z. Liang, et al., “A survey of multimodal large language models,” in *Proc. Int. Conf. CAICE*, pp. 405-409, Xi'an, China, Jan. 2024. (<https://doi.org/10.1145/3672758.3672824>)
- [5] J. D. M. W. C. Kenton and L. K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. NAACL-HLT*, vol. 1, pp. 4171-4186, Minneapolis, MN, USA, Jun. 2019. (<https://doi.org/10.18653/v1/n19-1423>)
- [6] J. Wei, et al., “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in NeurIPS*, vol. 35, pp. 24824-24837, New Orleans, LA, USA, Nov. 2022.
- [7] S. Kombrink, T. Mikolov, M. Karafiat, and L. Burget, “Recurrent neural network based language modeling in meeting recognition,” in *Proc. Annual Conf. Int. Speech Commun. Assoc. (Interspeech)*, vol. 11, pp. 2877-288, Florence, Italy, Aug. 2011. (<https://doi.org/10.21437/Interspeech.2011-720>)
- [8] Y. Kubo, S. Karita, and M. Bacchiani, “Knowledge transfer from large-scale pretrained language models to end-to-end speech recognizers,” in *Proc. IEEE ICASSP*, pp. 8512-8516, Virtual and Singapore, May

2022.
[\(https://doi.org/10.1109/ICASSP43922.2022.9746801\)](https://doi.org/10.1109/ICASSP43922.2022.9746801)
- [9] D. Wu and F. Wang, "Netilm: Adapting large language models for networking," in *Proc. ACM SIGCOMM Conf.*, pp. 661-678, Sydney, NSW, Australia, Aug. 2024.
[\(https://doi.org/10.1145/3651890.3672268\)](https://doi.org/10.1145/3651890.3672268)
- [10] F. Liu, et al., "A medical multimodal large language model for future pandemics," *NPJ Digital Med.*, vol. 6, no. 1, pp. 226, Dec. 2023.
[\(https://doi.org/10.1038/s41746-023-00952-2\)](https://doi.org/10.1038/s41746-023-00952-2)
- [11] S. Javaid, et al., "Leveraging large language models for integrated satellite-aerial-terrestrial networks: Recent advances and future directions," *IEEE Open J. Commun. Soc.*, pp. 1-35, 2024. (Early Access)
[\(https://doi.org/10.1109/OJCOMS.2024.3522103\)](https://doi.org/10.1109/OJCOMS.2024.3522103)
- [12] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "Graspgrpt: Leveraging semantic knowledge from a large language model for task-oriented grasping," *IEEE Robotics and Automat. Lett.*, Nov. 2023.
[\(https://doi.org/10.1109/LRA.2023.3320012\)](https://doi.org/10.1109/LRA.2023.3320012)
- [13] Y. Huang, et al., "Large language models for networking: Applications, enabling techniques, and challenges," *IEEE Netw.*, pp. 1-8, 2024. (Early Access)
[\(https://doi.org/10.1109/MNET.2024.3435752\)](https://doi.org/10.1109/MNET.2024.3435752)
- [14] S. Gupta, Y. Huang, Z. Zhong, T. Gao, K. Li, and D. Chen, "Recovering private text in federated learning of language models," *Advances in NeurIPS*, vol. 35, pp. 8130-8143, New Orleans, LA, USA, Nov. 2022.
- [15] H. Fang, D. Zhang, C. Tan, P. Yu, Y. Wang, and W. Li, "Large language model enhanced autonomous agents for proactive fault-tolerant edge networks," in *Proc. IEEE Conf. Computer Commun. Wkshps (INFOCOM WKSHPS)*, pp. 1-2, Vancouver, BC, Canada, May 2024.
[\(https://doi.org/10.1109/INFOCOMWKSHPS61880.2024.10620727\)](https://doi.org/10.1109/INFOCOMWKSHPS61880.2024.10620727)
- [16] Y. Shen, et al., "Large language models empowered autonomous edge AI for connected intelligence," *IEEE Commun. Mag.*, vol. 62, no. 10, pp. 140-146, Oct. 2024.
[\(https://doi.org/10.1109/MCOM.001.2300550\)](https://doi.org/10.1109/MCOM.001.2300550)
- [17] Y. He, J. Fang, F. R. Yu, and V. C. Leung, "Large language models (LLMs) inference offloading and resource allocation in cloud-edge computing: An active inference approach," *IEEE Trans. Mobile Computing*, vol. 23, no. 12, pp. 11253-11264, Dec. 2024.
[\(https://doi.org/10.1109/TMC.2024.3415661\)](https://doi.org/10.1109/TMC.2024.3415661)
- [18] M. Xu, et al., "When large language model agents meet 6G networks: Perception, grounding, and alignment," *IEEE Wireless Commun.*, vol. 31, no. 6, pp. 63-71, Dec. 2024.
[\(https://doi.org/10.1109/MWC.005.2400019\)](https://doi.org/10.1109/MWC.005.2400019)
- [19] C. Tang, D. Huang, W. Ge, W. Liu, and H. Zhang, "Graspgrpt: Leveraging semantic knowledge from a large language model for task-oriented grasping," *IEEE Robotics and Automat. Lett.*, vol. 8, no. 11, pp. 7551-7558, Nov. 2023.
[\(https://doi.org/10.1109/LRA.2023.3320012\)](https://doi.org/10.1109/LRA.2023.3320012)
- [20] Z. Mandi, S. Jain, and S. Song, "Roco: Dialectic multi-robot collaboration with large language models," in *Proc. ICRA*, pp. 286-299, Yokohama, Japan, May 2024.
[\(https://doi.org/10.1109/ICRA57147.2024.10610855\)](https://doi.org/10.1109/ICRA57147.2024.10610855)
- [21] N.-N. Dao, N. H. Tu, T.-D. Hoang, T.-H. Nguyen, L. V. Nguyen, K. Lee, L. Park, W. Na, and S. Cho, "A review on new technologies in 3GPP standards for 5G access and beyond," *Elsevier Computer Netw.*, vol. 245, no. 110370, May 2024.
[\(https://doi.org/10.1016/j.comnet.2024.110370\)](https://doi.org/10.1016/j.comnet.2024.110370)
- [22] T.-H. Nguyen, T. P. Truong, A.-T. Tran, N.-N. Dao, L. Park, and S. Cho, "Intelligent heterogeneous aerial edge computing for advanced 5G access," *IEEE Trans. Netw. Sci. and Eng.*, vol. 11, no. 4, Jul.-Aug. 2024.
[\(https://doi.org/10.1109/TNSE.2024.3371434\)](https://doi.org/10.1109/TNSE.2024.3371434)
- [23] N. Carlini, et al., "Extracting training data

- from large language models,” in *Proc. USENIX Security Symp.*, pp. 2633-2650, Aug. 2021.
- [24] M. Artetxe, et al., “Efficient large scale language modeling with mixtures of experts,” in *Proc. Conf. EMNLP*, pp. 11699-11732, Abu Dhabi, United Arab Emirates, Dec. 2022. (<https://doi.org/10.18653/v1/2022.emnlp-main.804>)
- [25] M. Kim, T. Kim, H. Choi, S. Min, J.-H. Lim, and S. Park, “Multi-modal LLM-based fully-automated training dataset generation software platform for mathematics education,” *IEEE ICSE*, Ottawa, Canada, Apr./May 2025.
- [26] J. Li, D. Li, S. Savarese, and S. Hoi, “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models,” in *Proc. ICML*, pp. 19730-19742, Honolulu, Hawaii, USA, Jul. 2023.
- [27] S. S. Kannan, V. L. N. Venkatesh, and B.-C. Min, “SMART-LLM: Smart multi-agent robot task planning using large language models,” in *Proc. IEEE/RSJ Int. Conf. IROS*, pp. 12140-12147, Abu Dhabi, United Arab Emirates, Oct. 2024. (<https://doi.org/10.1109/IROS58592.2024.10802322>)
- [28] S. Lee, G. S. Kim, S. Park, and J. Kim, “Advanced taxiing path guidance using multi-agent reinforcement learning for air traffic management,” in *Proc. Int. Symp. Modeling and Optimization in Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, pp. 305-312, Seoul, Korea, Oct. 2024.
- [29] M. Choi, T. Xiang, and J. Kim, “Intelligent caching for seamless high-quality streaming in vehicular networks: A multi-agent reinforcement learning approach,” *IEEE Trans. Intell. Veh.*, vol. 9, no. 2, pp. 3672-3686, Feb. 2024. (<https://doi.org/10.1109/TIV.2023.3344478>)
- [30] S. Jung, W. J. Yun, M. Shin, J. Kim, and J.-H. Kim, “Orchestrated scheduling and multi-agent deep reinforcement learning for cloud-assisted multi-UAV charging systems,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 5362-5377, Jun. 2021. (<https://doi.org/10.1109/TVT.2021.3062418>)
- [31] T. T. H. Pham, W. Noh, and S. Cho, “Multi-agent reinforcement learning based optimal energy sensing threshold control in distributed cognitive radio networks with directional antenna,” *Elsevier ICT Express*, vol. 10, no. 3, pp. 472-478, Jun. 2024. (<https://doi.org/10.1016/j.icte.2024.01.001>)
- [32] J. Oh, D. Lee, D. Won, W. Noh, and S. Cho, “Communication-efficient federated learning over-the-air with sparse one-bit quantization,” *IEEE Trans. Wireless Commun.*, vol. 23, no. 10, Oct. 2024. (<https://doi.org/10.1109/TWC.2024.3432758>)
- [33] M. C. Ho, A.-T. Tran, D. Lee, J. Paek, W. Noh, and S. Cho, “A DDPG-based energy efficient federated learning algorithm with SWIPT and MC-NOMA,” *Elsevier ICT Express*, vol. 10, no. 3, pp. 600-607, Jun. 2024. (<https://doi.org/10.1016/j.icte.2023.12.001>)
- [34] S. Sriram, C. H. Karthikeya, K. P. Kishore Kumar, N. Vijayaraj, and T. Murugan, “Leveraging local LLMs for secure in-system task automation with prompt-based agent classification,” *IEEE Access*, vol. 12, pp. 177038-177049, Dec. 2024. (<https://doi.org/10.1109/ACCESS.2024.3505298>)
- [35] K. Pandya and M. Holia, “Automating customer service using LangChain: Building custom open-source GPT Chatbot for organizations,” *arXiv preprint arXiv: 2310.05421*, 2023.
- [36] S. Smith, et al., “Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model,” *arXiv preprint arXiv:2201.11990*, 2022.
- [37] C. Holmes, et al., “Deepspeed-fastgen: High-throughput text generation for llms via mii and deepspeed-inference,” *arXiv preprint arXiv:2401.08671*, 2024.

- [38] B. Sriman, S. H. Annie Silviya, Y. Mouleesh, S. Vinod, S. Nishanthini, and P. Nikitha, "Intelligent document interaction with advanced vector embeddings and FAISS-CPU indexing," in *Proc. Int. Conf. Electr., Commun. and Aerospace Technol. (ICECA)*, pp. 1-6, Coimbatore, India, Nov. 2024. (https://doi.org/10.1109/ICECA63461.2024.10800948)

손 석빈 (Seok Bin Son)



2022년 2월 : 서울여자대학교 정 보보호학과 졸업
2022년 3월~현재 : 고려대학교 전기전자공학과 석박통합과정
<관심분야> Neural Architecture Search, Distributed Learning, Large AI Model
[ORCID:0000-0002-3692-0752]

김 종현 (Joongheon Kim)



2004년 2월 : 고려대학교 컴퓨터 학과 졸업
2006년 2월 : 고려대학교 컴퓨터 학과 석사 졸업
2014년 8월 : University of Southern California Computer Science 박사 졸업
2016년 3월 : 중앙대학교 소프트웨어학대학 조교수
2019년 9월~현재 : 고려대학교 전기전자공학부 조교수/ 부교수
<관심분야> Stochastic Optimization, Mobility, Reinforcement Learning, Quantum Deep Learning
[ORCID:0000-0003-2126-768X]

조 창식 (Changsik Cho)



1993년 2월 : 경북대학교 컴퓨터 학과 졸업
1995년 2월 : 경북대학교 컴퓨터 학과 석사 졸업
2011년 8월 : 충남대학교 컴퓨터 과학과 박사 졸업
1995년 3월~현재 : ETRI AI컴퓨팅시스템SW연구실 책임연구원/실장
<관심분야> AutoML, MLOps, 노코드 신경망 개발 도구 [ORCID:0000-0002-2162-8142]

박 수현 (Soohyun Park)



2019년 2월 : 중앙대학교 컴퓨터 공학과 졸업
2023년 9월 : 고려대학교 전기전자 공학과 박사 졸업
2024년 3월 : 숙명여자대학교 소프트웨어학부 조교수
<관심분야> Connected Mobility, Distributed Computing, Stochastic Optimization, Data Science
[ORCID:0000-0002-6556-9746]