사물인터넷 기반 LLM 토큰 자원 관리 및 AI 실습 지원 플랫폼 설계에 관한 연구

이근호*
백석대학교 컴퓨터공학부 교수

A Study on the Design of an IoT-Based LLM Token Resource Management and AI Practice Support Platform

Keun-Ho Lee*
Professor, Division of Computer Engineering, Baekseok University

요 약 본 연구는 최근 대학 교육 환경에서 확산되고 있는 생성형 인공지능 기반 실습 교육의 효율적 운영을 위해, 사물인터넷(IoT) 기반의 LLM 토큰 자원 관리 및 AI 실습 지원 플랫폼 설계를 제안한다. 생성형 AI API는 종량제 방식으로 비용이 발생하고, 대규모 동시 접속 환경에서는 호출 제한 및 자원 병목 문제가 발생함에 따라 안정적인 교육 환경 구축이 요구된다. 이를 해결하기 위해 본 연구에서는 중앙 집중형 토큰 관리 시스템과 분산 사용자 환경을 결합한 IoT 기반 플랫폼 구조를 설계하였다. 제안 시스템은 토큰의 효율적 배분, 실시간 사용량 모니터링, 지능형 알림 기능을 통해 학습 연속성을 보장하며, 다양한 LLM API를 통합 관리할 수 있도록 설계되었다. 또한, 데이터 보안 강화를 위해 API 프록시 기반 접근 제어 및 보안 격리 환경을 적용하여 산학협력 프로젝트에서도 안전한 데이터 활용이 가능하도록 하였다. 특히, 학습 성과와 토큰 사용량 간의 상관관계 분석을 통해 교육 효과를 정량적으로 평가할 수 있는 기능을 포함하였다. 본 연구에서 제안한 플랫폼은 AI 실습 교육의 비용 효율성, 안정성, 확장성을 동시에 확보할 수 있는 통합 운영 모델로서, 향후 대학 중심 AI 교육 인프라 구축에 중요한 기반이 될 것으로 기대된다.

주제어 : 사물인터넷, 대형언어모델, 토큰 자원 관리, 생성형 인공지능 교육, AI 실습 플랫폼

Abstract This study proposes the design of an Internet of Things (IoT)-based LLM token resource management and AI practice support platform for the efficient operation of generative artificial intelligence-based practical education, which is rapidly expanding in recent university learning environments. Since generative AI APIs incur costs under a pay-as-you-go model and often suffer from call limitations and resource bottlenecks in large-scale concurrent usage environments, there is a growing need for a stable and sustainable educational infrastructure. To address these issues, this study designs an IoT-based platform architecture that combines a centralized token management system with a distributed user environment. The proposed system is designed to support efficient token allocation, real-time usage monitoring, and intelligent notification functions, thereby ensuring continuity in learning activities while enabling the integrated management of various LLM APIs. In addition, to strengthen data security, the platform applies API proxy-based access control and a secure isolated environment, allowing safe data utilization even in industry-academic collaboration projects. In particular, the system includes a function for quantitatively evaluating educational effectiveness by analyzing the correlation between learning outcomes and token usage. The platform proposed in this study is expected to serve as an integrated operational model that simultaneously secures cost efficiency, stability, and scalability in AI practical education, and to provide an important foundation for building university-centered AI educational infrastructure in the future.

Key Words : IoT, LLM, Token Resource Management,Generative AI Education, AI Practice Platform

이 논문은 2026학년도 백석대학교 학술연구비 지원을 받아 작성되었음
*교신저자 : 이근호(leekeunho1004@gmail.com)
접수일 2026년 04월 02일 수정일 2026년 04월 14일 심사완료일 2026년 04월 20일

1. 서론

최근 생성형 인공지능(Generative AI) 기술의 발전과 함께 대형언어모델(Large Language Model, LLM)을 활용한 교육 환경이 대학 전반으로 확산되고 있다. 특히 ChatGPT, Claude, Gemini 등 다양한 LLM 기반 API를 활용한 실습 중심 교육이 증가하면서, 학생들은 실제 산업 수준의 AI 서비스를 직접 구현하고 활용하는 경험을 쌓고 있다. 이러한 변화는 기존 이론 중심 교육에서 벗어나 실무 중심 교육으로의 전환을 가속화하고 있다. 그러나 생성형 AI API는 종량제 방식으로 운영되며, 사용량에 따라 비용이 증가하는 구조를 갖는다. 이에 따라 학생 개인이 직접 API를 사용할 경우 비용 부담이 발생하고, 이는 학습 기회의 불균형 및 창의적 시도 위축으로 이어질 수 있다. 또한 무료 API의 경우 호출 횟수 제한 및 성능 제약으로 인해 대규모 동시 실습 환경에서 안정적인 운영이 어렵다는 문제가 존재한다. 이와 더불어, 산학협력 기반 프로젝트에서는 실제 기업 데이터를 활용한 실습이 이루어지는데, 이 과정에서 데이터 보안 및 개인정보 보호 문제 또한 중요한 이슈로 부각되고 있다. 기존 방식처럼 학생 개인 계정을 통한 API 접근은 데이터 유출 위험을 증가시키며, 통합적인 관리 및 통제의 필요성이 제기된다. 따라서 본 연구에서는 이러한 문제를 해결하기 위해 사물인터넷(IoT) 기반의 LLM 토큰 자원 관리 및 AI 실습 지원 플랫폼을 제안한다. 제안 시스템은 중앙집중형 토큰 관리 구조와 분산 사용자 환경을 결합하여, 다수의 사용자가 동시에 안정적으로 AI 서비스를 활용할 수 있도록 설계되었다. 또한 토큰 사용량 관리, 실시간 모니터링, 보안 제어 기능을 통합함으로써 효율적이고 안전한 AI 교육 환경을 제공하는 것을 목표로 한다.

2장에서는 생성형 인공지능 기반 교육 시스템, LLM API 자원 관리, IoT 기반 자원 관리 기술에 대한 관련 연구를 분석하고 기존 연구의 한계를 도출한다. 3장에서는 제안하는 IoT 기반 LLM 토큰 자원 관리 및 AI 실습 지원 플랫폼의 전체 구조와 핵심 기능을 설명한다. 4장에서는 제안 시스템의 운영 방식 및 적용 시나리오를 기반으로 한 실험 환경과 평가 방법을 제시한다. 5장에서는 시스템 적용 결과를 분석하고 토큰 자원 관리 효율성 및 교육적 효과를 검증한다. 마지막으로 6장에서는 본 연구의 결론을 도출하고 향후 연구 방향을 제시한다. 본 연구의 학술적 기여는 다음과 같다.

첫째, 기존 LLM 자원 관리 연구가 시스템 효율 개선에 집중된 것과 달리, 본 연구는 교육 환경을 고려한 정책 기반 토큰 자원 관리 모델을 제안한다. 둘째, IoT 기반 자원 관리 구조를 LLM API 자원 관리에 적용하여 논리적 AI 자원과 사용자 환경을 통합 관리하는 새로운 구조를 제시한다. 셋째, 토큰 사용량과 학습 성과 간의 관계를 분석하여 생성형 AI 기반 교육 효과를 정량적으로 평가할 수 있는 모델을 제안한다.

2. 관련연구

2.1 생성형 인공지능 기반 교육 시스템

최근 생성형 인공지능과 대형언어모델(LLM)의 발전으로 대학 교육에서는 대화형 학습 지원, 자동 피드백, 코드 생성 등 다양한 AI 기반 학습 환경이 빠르게 확산되고 있다[1-5]. 이러한 기술은 학습자의 수준에 맞는 맞춤형 응답을 제공하여 자기주도 학습과 실습 중심 교육을 촉진한다 [1,4,5]. 특히 ChatGPT를 중심으로 학습 경험과 교육 혁신에 대한 연구가 활발히 진행되고 있다 [2-5]. 그러나 기존 연구는 주로 학습 효과와 활용성에 초점을 두고 있으며, 비용 관리, 대규모 동시 사용, 안정적 운영 구조와 같은 실제 교육 환경 문제는 충분히 반영하지 못하고 있다[1-3]. 또한 학습 의존성 및 평가 공정성 문제도 제기되고 있어 통합 운영 모델의 필요성이 강조된다[2,5].

2.2 LLM API 자원 관리 및 비용 최적화

LLM API는 토큰 기반 종량제 구조로 인해 교육 환경에서 비용 부담이 크게 증가할 수 있다 [6-10]. 이에 따라 메모리 관리, 스케줄링, 추론 최적화 등 다양한 효율화 기법이 제안되었으며, ORCA, PetS, Paged Attention 기반 기술 등이 대표적이다[6-8]. 또한 최근 연구에서는 모델 경량화 및 시스템 최적화의 중요성이 강조되고 있다[9,10]. 하지만 이러한 연구는 주로 시스템 성능 향상에 집중되어 있으며, 사용자 유형과 학습 단계에 따른 정책 기반 자원 배분이나 교육 운영 관점의 관리 모델은 충분히 다루지 못하고 있다 [6-10].

2.3 IoT 기반 자원 관리 기술

IoT 환경에서는 중앙 관리 서버와 분산 디바이스 간 협력을 통해 자원을 효율적으로 관리하며, 실시간 모니터링, 접근 제어, 보안 기술이 핵심적으로 활용된다 [11-15]. 이러한 구조는 다수의 사용자가 동시에 LLM API를 사용하는 교육 환경과 유사한 특성을 가진다.

그러나 기존 연구는 센서 네트워크, 엣지 컴퓨팅 등 물리적 자원 중심으로 발전되어 왔으며, LLM 토큰과 같은 논리적 AI 자원 관리에 대한 적용은 제한적이다.

2.4 기존 연구의 한계

기존 연구는 생성형 AI 교육, LLM 자원 관리, IoT 기술이 각각 독립적으로 발전해 왔으나, 이를 통합적으로 고려한 연구는 부족하다. 생성형 AI 교육 연구는 운영 비용 문제를 충분히 반영하지 못하며 [1-5], LLM 자원 관리 연구는 교육기관 수준의 정책 기반 관리로 확장되지 못하였다[6-10]. 또한 IoT 연구는 AI 자원 관리 적용이 제한적이다[11-15]. 따라서 본 연구에서는 IoT 기반 구조를 활용하여 LLM 토큰 자원을 통합 관리하고, 비용 효율성, 확장성, 보안성을 동시에 확보할 수 있는 AI 실습 지원 플랫폼을 제안한다 [1-15].

3. 제안 방법

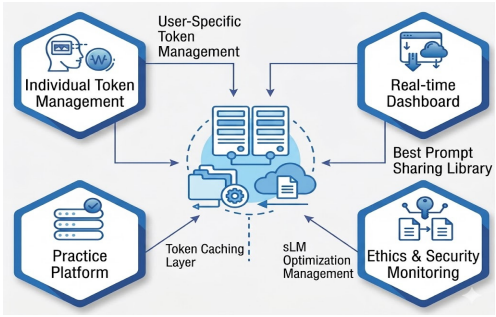
3.1 시스템 개요

본 연구에서는 생성형 인공지능 기반 교육 환경에서 발생하는 비용, 자원 관리, 보안 문제를 해결하기 위해 사물인터넷(IoT) 기반 LLM 토큰 자원 관리 및 AI 실습 지원 플랫폼을 제안한다. 제안 시스템은 IoT의 중앙 관리 + 분산 사용자 구조를 적용하여, 다수의 사용자가 동시에 다양한 LLM API를 활용할 수 있도록 설계되었다. 기존의 개별 API 키 방식과 달리 중앙 관리 서버를 통해 모든 요청을 통합 처리함으로써 자원 효율성과 보안성을 동시에 확보한다.

본 시스템은 IoT의 주요 특징인 분산 사용자 환경, 중앙 게이트웨이 기반 자원 제어, 엣지 처리(sLM 기반), 실시간 상태 모니터링 구조를 반영하여 설계되었다. 특히 sLM 기반 로컬 처리는 Edge Computing 개념을 적용한 것으로, 중앙 서버 의존도를 줄이고 시스템 확장성을 향상시키는 역할을 수행한다.

3.2 전체 시스템 아키텍처

제안하는 IoT 기반 LLM 토큰 자원 관리 시스템은 그림 1과 같이 통합 관리 플랫폼을 중심으로 구성된다. 시스템은 사용자 인터페이스, 토큰 관리, 분석 및 보안 기능이 통합된 중앙-분산 구조를 가진다. 사용자는 Practice Platform을 통해 AI 실습을 수행하며, 모든 요청은 중앙 플랫폼으로 전달된다.



[Fig. 1] Architecture overview of the Integrated Management Platform

이 과정에서 Individual Token Management 및 User-Specific Token Management 기능을 통해 사용자별 토큰 사용량과 정책이 적용된다. 중앙의 플랫폼은 API 프록시 역할을 수행하며, Token Caching Layer를 통해 중복 요청을 최소화하고, sLM Optimization Management를 통해 일부 요청을 로컬 처리함으로써 비용과 시스템 부하를 줄인다. 또한 Real-time Dashboard를 통해 시스템 상태와 사용량을 실시간으로 모니터링할 수 있으며, Best Prompt Sharing Library를 통해 학습 효율을 향상시킨다. 더불어 Ethics & Security Monitoring 기능을 통해 사용자 활동을 분석하고 보안 및 윤리 기준을 유지한다. 제안 아키텍처는 IoT 기반 자원 관리 구조를 적용하여 확장성, 비용 효율성, 보안성을 동시에 확보할 수 있는 통합 AI 실습 플랫폼을 제공한다.

3.3 토큰 자원 관리 모델

본 연구에서는 교육 환경에 최적화된 정책 기반 토큰 자원 관리 모델을 제안한다.

〈Table 1〉 Token Allocation Policy by User Type

User Type	Target Group	Monthly Token Allocation	Purpose
Basic Learners	Freshman-Sophomore	Low (50K-200K)	Basic practice
Advanced Learners	Junior-Senior	Medium (200K-500K)	Project-based learning
Researchers	Graduate students	High (500K-1M+)	Research and thesis
Industry Projects	Team-based	Very High	Real-world data analysis

표 1은 사용자 유형에 따른 토큰 배분 정책을 나타낸다. 제안된 모델은 학습자의 수준과 역할에 따라 토큰을 차등적으로 할당하여, 효율적인 자원 활용과 교육 효과를 동시에 달성하도록 설계되었다. 기초 학습자(1~2학년)는 기본적인 AI 실습 수행을 위해 비교적 낮은 수준(50K~200K)의 토큰이 할당되며, 심화 학습자(3~4학년)는 프로젝트 기반 학습을 위해 중간 수준(200K~500K)의 토큰을 부여받는다. 또한 대학원생과 같은 연구자는 논문 작성 및 고급 연구 수행을 위해 높은 수준(500K~1M 이상)의 토큰이 할당된다. 산학협력 프로젝트의 경우에는 실제 기업 데이터 기반 분석이 요구되므로 팀 단위로 매우 높은 수준의 토큰이 지원된다. 이와 같은 차등 배분 정책을 통해 학습 목적과 활용 수준에 맞는 자원 제공이 가능하며, 기존의 일괄적 토큰 배분 방식보다 자원 활용의 효율성을 향상시킬 수 있다.

3.4 실시간 모니터링 및 학습 분석 기능

제안 시스템은 토큰 자원 관리의 효율성을 향상시키고 교육 효과를 분석하기 위해 실시간 모니터링 및 학습 분석 기능을 제공한다. 모든 사용자 활동과 토큰 사용 데이터는 중앙 플랫폼에 저장되며, 이를 기반으로 다양한 분석이 수행된다. 관리자 및 교수자는 Real-time Dashboard를 통해 개인, 학과, 프로젝트 단위의 토큰 사용량을 실시간으로 확인할 수 있으며, 전체 시스템의 자원 사용 현황을 직관적으로 파악할 수 있다. 이를 통해 특정 사용자 또는 그룹에서 발생하는 과도한 사용이나 비정상적인 패턴을 신속하게 탐지할 수 있다. 또한 시스템은 토큰 사용 이력과 학습 결과 데이터를 연계하여 분석함으로써, 토큰 사용량과 학습 성과 간의 상관관계를 도출할 수 있도록 설계되었다. 예를 들어, 프로젝트 완성도, 과제 점수, 실습 수행 횟수 등의 데이터를 활용하여 AI 활용 수준과 학습 성과 간의 관계를 정량적으로 평가할 수 있다. 더불어 Best Prompt Sharing Library 기능을 통해 우수한 프롬프트 사례를 저장 및 공유함으로써, 학습자 간 지식 공유를 촉진하고 반복적인 시행착오를 줄일 수 있다. 이러한 기능은 단순한 자원 관리 시스템을 넘어, 학습 품질을 향상시키는 교육 지원 도구로서의 역할을 수행한다.

3.5 보안 및 윤리 기반 관리 구조

제안 시스템은 생성형 AI 활용 과정에서 발생할 수 있는 데이터 유출 및 오남용 문제를 해결하기 위해 보안 및 윤리 기반 관리 구조를 적용하였다. 우선 모든 API 요청은 중앙 플랫폼을 통해 처리되며, 개별 사용자 API 키 사용을 제한함으로써 외부로의 직접 접근을 차단한다. 이러한 구조는 API 프록시 기반 접근 제어 방식으로, 인증되지 않은 요청이나 비정상적인 사용을 효과적으로 방지할 수 있다. 또한 사용자 권한에 따라 접근 가능한 기능을 제한하는 역할 기반 접근 제어(RBAC)를 적용하여, 데이터 접근 범위를 최소화하였다. 특히 산학협력 프로젝트 수행 시에는 기업 데이터 보호를 위해 별도의 보안 격리 환경을 구성하여 외부 유출 가능성을 차단한다. 시스템은 Ethics & Security Monitoring 기능을 통해 사용자 활동을 지속적으로 분석하며, 비정상적인 프롬프트 사용, 과도한 요청, 부적절한 콘텐츠 생성 시도를 탐지한다. 이를 통해 AI 윤리 기준을 유지하고, 교육 환경에서의 책임 있는 AI 활용을 유도한다. 또한 모든 요청 및 응답 로그는 암호화하여 저장되며, 필요 시 감사 추적이 가능하도록 설계되었다. 이러한 보안 구조는 IoT 기반 시스템에서 요구되는 데이터 보호 및 접근 제어 요구사항을 반영한 것으로, 안전한 AI 실습 환경을 제공한다.

4. 적용 시나리오 및 평가 방법

4.1 평가 환경 구성

본 연구에서는 제안 시스템의 적용 가능성을 검토하기 위해 대학의 생성형 AI 실습 운영 환경을 반영한 평가 환경을 설정하였다. 평가 환경은 중앙 통합 관리 서버, 외부 LLM API 연동 환경, 그리고 웹 기반 사용자 접속 환경으로 구성하였다. 중앙 서버 환경은 Intel i7급 CPU, 32GB RAM, 1TB SSD 수준의 단일 서버를 기준으로 가정하였으며, 제안 시스템의 토큰 관리, 캐싱, 모니터링 및 접근 제어 기능이 해당 서버에서 운영되는 상황을 반영하였다. 또한 일부 요청은 sLM 기반 로컬 처리 환경으로 분산되는 구조를 포함하도록 설정하였다. 네트워크 환경은 대학 내 유선 및 무선 네트워크가 혼합된 일반적인 교육용 접속 환경을 가정하였으며, 외부 LLM API와의 연동은 안정적인 인터넷 연결 상태를 전제로 하였다. 사용자는 별도의 개발 환경이 아닌 웹 기반 AI 실습 플랫폼을 통해 시스템에 접속하는 형태로 설정하였고, 학생 및 교수 사용자가 동일 플랫폼에서 프롬프트를 입력하고 응답을 수신하는 구조를 반영하였다. 이러한 평가 환경은 실제 대학의 생성형 AI 실습 수업, 프로젝트 수업, 산학협력 실습 환경에서 발생할 수 있는 운영 조건을 고려하여 구성하였다.

4.2 적용 시나리오 및 재현 조건

평가 시나리오는 실제 대학 실습 환경에서 자주 발생하는 생성형 AI 활용 상황을 반영하여 구성하였다. 사용자는 웹 기반 플랫폼을 통해 프롬프트를 입력하고, 시스템은 이를 중앙 관리 계층에서 검증한 후 외부 LLM API 또는 sLM 로컬 처리 계층으로 전달하는 흐름을 따른다. 이를 위해 프롬프트 유형은 코드 생성, 질의응답, 요약의 세 가지로 구분하였으며, 각 유형별로 동일한 프롬프트 세트를 기준으로 비교 분석이 가능하도록 설정하였다. 재현성을 높이기 위해 각 비교군의 분석은 동일한 프롬프트 세트를 기반으로 수행되도록 하였다. 또한 동시 사용자 증가에 따른 시스템 부하 환경을 반영하기 위해 사용자 수는 10명, 30명, 50명의 세 단계로 구분하여 설정하였다. 각 사용자는 동일한 시점에 실습 플랫폼에 접속하는 상황을 가정하였으며, 요청은 동시성 환경을 모사하기 위해 1초 간격으로 순차적으로 발생하도록 구성하였다. 즉, 각 사용자 그룹은 동일한 프롬프트 유형과 입력 순서를 유지한 상태에서, 동일한 시간 간격으로 요청을 생성하도록 설정하였다. 또한 각 사용자 수 조건에 대해 동일한 시나리오를 반복 적용할 수 있도록 구성함으로써, 토큰 사용량, API 호출 횟수, 응답 시간, 처리량, 비용 절감률 등의 지표를 비교 가능하도록 하였다. 이를 통해 사용자 수 증가에 따른 시스템의 구조적 차이와 운영 특성을 일관된 조건에서 분석할 수 있도록 하였다. 비교 분석의 일관성을 확보하기 위해 각 시나리오는 동일한 모델 환경(GPT-4 및 GPT-3.5 기반 API), 동일한 프롬프트 유형, 동일한 요청 간격을 전제로 하였으며, 제안 방식에서는 캐싱 및 sLM 기반 로컬 처리 비율을 함께 반영하여 운영 구조의 차이를 비교할 수 있도록 하였다. 이러한 설정을 통해 제안 시스템과 기존 방식 간의 구조적 차이와 운영 효과를 보다 명확하게 분석하고자 하였다.

5. 적용 시나리오 기반 분석

본 장에서는 제안 시스템의 적용 시나리오 기반 분석을 통해 토큰 자원 관리 효율성과 운영 측면의 효과를 검토한다. 이를 위해 개별 API 키 방식, 기관형 API Gateway 방식, 오픈소스 LLM 기반 방식, 클라우드 정책 기반 방식과 비교하여 분석을 수행하였다.

5.1 토큰 사용 효율성 분석

표 2는 사용자 수 증가에 따른 평균 토큰 사용량 변화를 나타낸다. 제안 시스템은 모든 사용자 구간에서 기존 방식 대비 낮은 토큰 사용량을 유지하는 구조를 보였으며, 사용자 수가 증가할수록 토큰 절감 효과가 더욱 뚜렷해지는 경향을 나타냈다.

분석 결과, 제안 방식은 평균 토큰 사용량을 기존 대비 약 28% 절감할 수 있는 구조로 나타났다. 또한 기존 방식에서 높게 나타나는 중복 요청 비율은 캐싱 기능 적용을 통해 감소될 수 있는 것으로 분석되었다.

이러한 결과는 Token Caching Layer를 통한 중복 요청 제거와 sLM 기반 로컬 처리 기능이 결합된 구조에 기인한 것으로 판단된다. 특히 동시 사용자 수 증가 상황에서도 토큰 사용량 증가 폭이 상대적으로 완만하게 유지되는 특성을 보여, 대규모 실습 환경에서의 자원 활용 효율성을 확보할 수 있는 구조로 분석되었다.

〈Table 2〉 Comparison of Token Usage Efficiency

Category	Conventional Method	Proposed Method	Reduction
Average Token Usage	100%	72%	Approx. 28% reduction
Duplicate Request Ratio	High	Low	-
Caching Effect	Not applied	Applied	-

5.2 시스템 성능 및 처리 효율 분석

표 3은 시스템 응답 시간과 처리 효율 측면에서의 비교 결과를 나타낸다. 분석 결과, 제안 방식은 기존 방식 대비 응답 시간 감소와 처리량 증가 측면에서 효율적인 구조를 가지는 것으로 나타났다. 특히 제안 시스템은 일부 요청을 sLM 기반 로컬 처리로 분산하고, 캐싱을 통해 중복 요청을 줄임으로써 외부 API 호출 부담을 감소시키는 구조를 가진다. 이러한 구조는 다수 사용자가 동시에 접속하는 환경에서 시스템 부하를 완화하고, 안정적인 서비스 제공을 가능하게 하는 요소로 작용할 수 있다. 또한 기관형 API Gateway 방식과 비교할 때, 제안 시스템은 단순 중앙 통제 구조를 넘어 캐싱 및 로컬 처리 기능을 포함함으로써 보다 높은 처리 효율을 확보할 수 있는 구조로 분석되었다.

〈Table 3〉 System Performance and Processing Efficiency Compariso

Metric	Conventional Method	Proposed Method	Improvement
Average Response Time	Baseline (1.0)	0.82	Approx. 18% reduction
System Throughput	100%	125%	Approx. 25% increase
Number of API Calls	High	Reduced	Efficiency improved

5.3 비용 절감 및 교육적 효과 분석

본 연구에서는 제안 시스템의 교육적 활용 가능성을 검토하기 위해 32명의 학생(N=32)을 대상으로 사전-사후 비교를 수행하였다. 학습 몰입도는 Likert 5점 척도 기반 8개 문항으로 구성된 설문을 통해 측정하였으며, 프로젝트 완성도는 교수 평가 점수를 기준으로 분석하였다. 분석 결과, 학습 몰입도는 사전 평균 3.21점에서 사후 평균 4.02점으로 증가하는 경향을 보였으며, 프로젝트 완성도 역시 평균 78.4점에서 85.6점으로 향상되는 경향이 나타났다. 이러한 변화는 제안 시스템이 반복 실습 기회를 확대하고 프롬프트 개선 활동을 촉진하는 구조를 제공함으로써 학습 참여와 수행 수준 향상에 기여할 가능성을 시사한다. 특히 토큰 사용량과 학습 성과 간의 관계를 함께 고려할 때, 충분한 토큰 자원 제공이 학습자의 실습 수행과 문제 해결 활동을 지원하는 요소로 작용할 수 있는 것으로 분석되었다. 또한 Best Prompt Sharing Library와 같은 기능은 학습자 간 지식 공유를 촉진하여 학습 효율성을 높일 수 있는 구조로 판단된다.

〈Table 4〉 Pre- and Post-Comparison of Educational Effectiveness

Item	Pre-test	Post-test	Mean Change
Learning Engagement (5-point Likert)	3.21	4.02	+0.81
Project Completion Score (100 points)	78.4	85.6	+7.2

6. 결론 및 향후 연구

본 연구에서는 생성형 인공지능 기반 교육 환경에서 발생하는 비용, 자원 관리 및 보안 문제를 해결하기 위해 IoT 기반 LLM 토큰 자원 관리 및 AI 실습 지원 플랫폼을 제안하였다. 제안 시스템은 중앙 통합 관리 구조를 기반으로 토큰 자원을 효율적으로 배분하고, 실시간 모니터링 및 보안 기능을 통해 안정적인 AI 실습 환경을 지원할 수 있는 구조로 설계되었다. 적용 시나리오 기반 분석 결과, 제안 시스템은 기존 방식 대비 약 25~30% 수준의 비용 절감 가능성을 가지며, 캐싱 및 sLM 기반 로컬 처리 구조를 통해 시스템 처리 효율성과 응답 성능 측면에서 개선 효과를 기대할 수 있는 것으로 나타났다. 또한 사용자 유형 기반 자원 배분과 실시간 분석 기능은 교육 환경에서의 운영 효율성과 안정성을 향상시킬 수 있는 요소로 분석되었다. 더불어 제안 시스템은 반복 실습 기회 확대, 프롬프트 개선 활동 촉진, 학습 참여 유도 등의 측면에서 교육적 효과를 기대할 수 있는 구조를 가지며, 생성형 AI 기반 실습 교육 환경에서 학습 지원 도구로 활용될 가능성을 보여주었다. 그러나 본 연구는 실제 구현 기반의 장기 운영 데이터가 아닌 설계 및 적용 시나리오 기반 분석에 초점을 두고 있어, 다양한 교육 환경에서의 일반화에는 한계가 있다. 따라서 향후 연구에서는 실제 대학 및 산학협력 환경에 시스템을 적용하여 장기적인 운영 데이터를 확보하고, 비용 절감 효과와 교육적 성과를 실증적으로 검증할 필요가 있다. 또한 사용자 행동 및 토큰 사용 패턴을 반영한 지능형 자원 배분 모델, 다양한 생성형 AI 및 멀티모달 모델을 통합한 확장형 플랫폼, 그리고 학습 성과를 정량적으로 분석할 수 있는 평가 체계의 고도화가 요구된다. 이를 통해 제안 시스템은 향후 대학 중심 AI 교육 인프라 구축을 위한 실질적인 운영 모델로 발전될 수 있을 것으로 기대된다.

REFERENCES

[1] D. Lee, M. Arnold, A. Srivastava, K. Plastow, P. Strelan, F. Ploeckl, D. Lekkas and E. Palmer, "The impact of generative AI on higher education learning and teaching: A study of educators' perspectives," Computers and Education: Artificial Intelligence, Vol.6, 2024.

[2] P.S. Bhullar, M. Joshi and R. Chugh, "ChatGPT in higher education - a synthesis of the literature and a future research agenda," Education and Information Technologies, Vol.29, pp.21501-21522, 2024.

[3] M.I. Baig and E. Yadegaridehkordi, "ChatGPT in the higher education: A systematic literature review and research challenges," International Journal of Educational Research, Vol.127, 2024.

[4] W.K. Monib, A. Qazi and M.M. Mahmud, "Exploring learners' experiences and perceptions of ChatGPT as a learning tool in higher education," Education and Information Technologies, Vol.30, No.1, pp.917-939, 2025.

[5] R. Deng, M. Jiang, X. Yu, Y. Lu and S. Liu, "Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies," Computers & Education, Vol.227, 2025.

[6] G.-I. Yu, J.S. Jeong, G.-W. Kim, S. Kim and B.-G. Chun, "Orca: A Distributed Serving System for Transformer-Based Generative Models," Proceedings of the 16th USENIX Symposium on Operating Systems Design and Implementation(OSDI 2022), pp.521-538,

2022.

[7] Z. Zhou, X. Wei, J. Zhang and G. Sun, "PetS: A Unified Framework for Parameter-Efficient Transformers Serving," Proceedings of the 2022 USENIX Annual Technical Conference (USENIX ATC 2022), pp.489–504, 2022.

[8] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C.H. Yu, J.E. Gonzalez, H. Zhang and I. Stoica, "Efficient Memory Management for Large Language Model Serving with PagedAttention," Proceedings of the 29th ACM Symposium on Operating Systems Principles(SOSP 2023), pp.611–626, 2023.

[9] Z. Wan, X. Wang, C. Liu, S. Alam, Y. Zheng, J. Liu, Z. Qu, S. Yan, Y. Zhu, Q. Zhang, M. Chowdhury and M. Zhang, "Efficient Large Language Models: A Survey," Transactions on Machine Learning Research, 2024.

[10] Z. Zhou, X. Ning, K. Hong, T. Fu, J. Xu, S. Li, Y. Lou, L. Wang, Z. Yuan, X. Li, S. Yan, G. Dai, X.-P. Zhang, Y. Dong and Y. Wang, "A Survey on Efficient Inference for Large Language Models," arXiv preprint arXiv:2404.14294, 2024.

[11] A. Musaddiq, Y.B. Zikria, O. Hahm, H. Yu, A.K. Bashir and S.W. Kim, "A Survey on Resource Management in IoT Operating Systems," IEEE Access, Vol.6, pp.8459–8482, 2018.

[12] L. Kong, J. Tan, J. Huang, G. Chen, S. Wang, X. Jin, P. Zeng, M. Khan and S.K. Das, "Edge-computing-driven Internet of Things: A Survey," ACM Computing Surveys, Vol.55, No.8, pp.1-41 2022.

[13] R. Singh and S.S. Gill, "Edge AI: A survey," Internet of Things and Cyber-Physical Systems, Vol.3, pp.71–92, 2023.

[14] W. Rafique, L. Qi, I. Yaqoob, M. Imran, R.U. Rasool and W. Dou, "Complementing IoT services through software defined networking and edge computing: A comprehensive survey," IEEE Communications Surveys & Tutorials, Vol.22, No.3, pp.1761–1804, 2020.

[15] R. Roman, J. Lopez and M. Mambo, "Mobile edge computing, Fog et al.: A survey and analysis of security threats and challenges," Future Generation Computer Systems, Vol.78, Part 2, pp.680–698, 2018.

이 근 호(Keun Ho Lee) [종신회원]

■ 2006년 8월 : 고려대학교 컴퓨터학과(이학박사)
■ 2006년 9월 ~ 2010년 2월 : 삼성전자 DMC연구소 책임연구원
■ 2010년 3월 ~ 현재 : 백석대학교 컴퓨터공학부 교수

〈관심분야〉
AI보안, 침해사고대응, 융합보안, 개인정보보호, 블록체인, 산업보안, 취약점분석, 모의해킹 등