ISSN 2287-5026 (Print)
ISSN 2288-159X (Online)

Journal of the Institute of Electronics and Information Engineers

2026 8 제 63 권 8호

Vol.63, No.8 August 2026

www.theieie.org Vol.63, No.8 August 2026

The Institute of Electronics and Information Engineers (IEIE)
Room #907, The Korea Science Technology Center The first building, 22,
Teheran-ro 7 Gil, Gangnam-gu, Seoul, Republic of Korea

전자공학회 논문지
2026 8
제63권 8호
IEIE
사단법인 대한전자공학회

전자공학회논문지
•이 논문집은 한국연구재단 **우수등재학술지**임.

차 례

2026년 8월 제63권 제8호

TC / 통신

[미래네트워크]

SD / 반도체

[반도체 소자 및 재료]

[SoC 설계]

[PCB & Packaging]

CI / 컴퓨터

[멀티미디어]

AISP / 인공지능 신호처리

[영상 신호처리]

[음향 및 신호처리]

SC / 시스템 및 제어

[스마트팩토리]

IE / 산업전자

[신호처리 및 시스템]

[컴퓨터 응용]

Journal of The Institute of Electronics and Information Engineers Vol.63, NO.8, August 2026 https://doi.org/10.5573/ieie.2026.63.8.116
ISSN 2287-5026(Print) / ISSN 2288-159X(Online)

논문 2026-63-8-12

SCADA 운영 보조를 위한 유한상태제어기 적용 로컬 소형 LLM 기반 AI agent 시스템

(Local Small LLM-based AI Agent System with Finite State Controller for SCADA Operation Assistance)

김 정 윤*, 이 승 호**

(Jeong-Yoon Kim and Seung-Ho Lee©)

요 약

본 논문은 실제 지역난방 운영 현장에서 수집된 SCADA 데이터를 바탕으로 구성된 XAI4HEAT를 이용하여, 지역난방 운영 질의에 응답하는 로컬 소형 LLM 기반 AI agent 시스템을 제안한다. 지역난방 SCADA 질의응답은 운영자가 특정 시점 값, 기간 평균, 설비 간 비교 결과를 신속히 확인하는 실무와 직접 연결되므로, 자연스러운 문장 생성보다 정확한 절차 수행과 안전한 비응답이 더 중요하다. 기존의 종단 간(end-to-end) AI agent는 다음 행동 선택과 인자 생성을 모두 자유 생성에 의존하므로, 소형 모델 환경에서 무효 JSON 출력, 성급한 최종 응답, 부적절한 재질문 남용이 빈번하게 발생한다. 이를 완화하기 위해 본 연구는 원본 질의-행동 기록(trace)을 한국어 운영 환경에 맞는 질의응답 데이터셋으로 재구성하였다. 이를 위해 데이터를 샘플 단위 층화 분할 방식으로 나누고, 행동 이름을 표준화한 뒤, 다음 행동을 정하는 단계와 도구 호출 인자를 만드는 단계로 분리하였다. 또한 재질문(clarification)과 응답 보류(abstention)는 고정된 템플릿으로 처리하도록 하였다. 아울러 실제 서비스 경로에서는 학습된 정책을 유한상태제어기(Finite-State Controller, FSC)로 대체하여, 소형 모델이 명시된 상태 전이와 절차에 따라 제한된 방식으로만 도구를 사용하도록 설계하였다. 이는 데이터 재가공에 더해 실제 추론 경로의 제어 논리를 외부 상태 기계로 구현하였다는 점에서 본 연구의 차별성을 이룬다. 이에 따라 Qwen3-0.6B 및 Llama-3.2-1B-Instruct는 LoRA로 1 epoch 미세조정된 payload 생성기 역할만 로컬에서 수행한다. 단일 RTX 5060 Ti(16GB) 환경의 offline replay 평가에서 learned pipeline은 과업 완수(task accomplishment) 0건/33건 (0%)에 그친 반면, 제안 방식인 FSC+Qwen는 과업 완수 8건/33건 (24.2%), safe non-answer 14건/17건 (82.4%), 무효 출력 1건을 기록하였다. 이는 실제 산업 현장의 데이터와 제한된 연산 자원 환경에서도 운영자 확인을 전제로 한 지역난방 SCADA 보조 시스템이 실질적인 운영 지원 수단으로 구현 및 활용될 수 있음을 보여준다.

Abstract

This paper proposes a local small-scale LLM-based AI agent for answering district heating operation queries using XAI4HEAT, a dataset built from real SCADA data collected at district heating sites. In this domain, correctly executing procedures and safely abstaining are more important than fluent language generation, because operators must verify point values, period averages, and cross-facility comparisons quickly and reliably. Conventional end-to-end agents depend on free-form generation for both action selection and parameter creation, which often leads to invalid JSON, premature final answers, and excessive clarification in small-model settings. To address this, we reconstruct query-action traces into a Korean query-answering dataset, standardize action names, separate next-action prediction from tool-argument generation, and handle clarification and abstention with fixed templates. We further replace the learned policy in deployment with a finite-state controller (FSC), constraining tool use through explicit state transitions and procedures. Qwen3-0.6B and Llama-3.2-1B-Instruct are fine-tuned locally with LoRA for one epoch as payload generators. In offline replay evaluation on a single RTX 5060 Ti (16GB), the learned pipeline achieved 0/33 (0%) task accomplishment, while FSC+Qwen achieved 8/33 (24.2%), with 14/17 (82.4%) safe non-answers and 1 invalid output. These results show the district heating SCADA assistance system based on operator verification can be implemented and utilized as a practical operational support tool, even with actual industrial site data and environments with limited computational resources.

Keywords : District heating SCADA, Small LLM agent, Finite-state controller, LoRA fine-tuning, Operator support

*학생회원, **평생회원, 국립한밭대학교 전자공학과(Dept. Electronic Engineering, Hanbat National University)
© Corresponding Author(E-mail : shyolee@hanbat.ac.kr)
※ 본 연구는 한국연구재단이 지원하고 국립한밭대학교가 수행하는 이공계 연구생활장려금 지원사업의 지원을 받아 수행되었습니다. (과제번호 : RS-2025-03992968)
Received ; April 24, 2026 Revised ; June 5, 2026 Accepted ; June 5, 2026

Ⅰ. 서 론

SCADA(Supervisory Control and Data Acquisition)는 산업 설비의 상태를 원격으로 감시하고 제어하기 위한 시스템으로, 지역난방 현장에서는 다수의 서브스테이션으로부터 장기간 시계열 데이터를 지속적으로 수집한다. 여기서 지역난방 서브스테이션은 열원에서 공급된 열을 건물 또는 구역 단위로 전달하는 현장 설비로서, 온도, 유량, 압력과 같은 운전 정보를 계측하고 조정하는 역할을 담당한다. 운영자는 이러한 데이터를 바탕으로 특정 시점 값 조회, 기간 평균 산출, 서브스테이션 간 비교, 데이터 가용성 점검과 같은 반복 질의를 수행한다. 최근 공개된 XAI4HEAT 데이터셋은 실제 지역난방 운영 현장에서 수집된 SCADA 데이터를 제공함으로써 이러한 운영 질의를 재현 가능한 형태로 연구할 수 있는 기반을 마련하였다[1]. 이는 본 연구가 합성 예시가 아니라 실제 산업 데이터에서 나타나는 질의 유형과 제약을 다룬다는 점에서 의미가 있다.

그러나 실제 산업 현장에서는 단순 조회 기능만으로 충분하지 않다. 질문에 시간 정보가 빠져 있거나 설비명이 불명확한 경우에는 clarification, 즉 시스템이 부족한 정보를 다시 묻는 재질문이 필요하다. 반대로 존재하지 않는 설비나 지원되지 않는 지표가 입력된 경우에는 abstention, 즉 답을 지어내지 않고 응답을 보류하는 안전한 비응답이 요구된다. 특히 운영 환경에서는 잘못된 자동 응답이 단순 오답을 넘어 현장 판단을 왜곡할 수 있으므로, 높은 정답률만큼이나 안전한 비응답 전략이 중요하다.

이와 같은 문제를 해결하기 위해 최근에는 사용자의 질의를 해석하고 필요한 도구를 호출한 뒤 결과를 설명하는 AI agent 구성이 주목받고 있다[2]. 그 대표적 접근인 ReAct는 추론(reasoning)과 행동(acting)을 번갈아 수행하면서 도구 호출과 언어 생성의 결합을 시도한다[3]. 그러나 제한된 GPU 자원에서 구동 가능한 1B 이하 또는 1B급 소형 모델에 동일한 자유 생성 방식을 적용하면, 다음 절차를 정하는 action selection 단계와 도구 호출 인자를 만드는 payload generation 단계 모두에서 오류가 빈번해진다. 이는 소형 모델에 action selection과 payload formatting을 동시에 위임하는 방식이 현장 적용 안정성을 저하시킬 수 있음을 의미한다.

최근 LLM 기반 agent의 안정성을 높이기 위해 constrained generation[4], grammar-guided decoding[5], structured tool-calling 연구가 활발히 진행되고 있다. Constrained generation 및 grammar- guided decoding 계열 연구는 JSON, code, API call과 같이 정해진 형식을 갖는 출력을 생성할 때 decoding 과정에서 허용 가능한 token 또는 grammar를 제한함으로써 형식 오류를 줄이는 데 초점을 둔다 . 이러한 접근은 출력이 특정 schema나 context-free grammar를 만족하도록 만드는 데 효과적이지만, 산업 운영 질의에서 요구되는 절차적 의사결정, 예를 들어 질의 완전성 점검, 지원 가능 여부 판단, clarification과 abstention의 분기, 도구 호출 순서 제어까지 직접적으로 보장하는 것은 아니다.

Structured tool-calling 연구 역시 본 연구와 관련된다. ToolBench, Gorilla, API-Bank와 같은 연구는 LLM이 외부 API나 도구를 적절히 선택하고 호출하도록 학습하거나 평가하기 위한 benchmark와 framework를 제안하였다[6~8]. 이들 연구는 범용 도구 사용 능력과 API 호출 정확도를 향상시키는 데 중요한 기반을 제공한다. 그러나 대부분의 연구는 다양한 공개 API 또는 범용 tool-use scenario를 대상으로 하며, 지역난방 SCADA와 같이 제한된 도메인 지식, 엄격한 지원 가능 지표, 누락 정보에 대한 재질문, 존재하지 않는 설비에 대한 안전한 비응답이 동시에 요구되는 산업 운영 환경을 직접적으로 다루지는 않는다.

본 연구의 FSC는 이러한 선행 연구와 달리 단순히 출력 형식을 제약하는 grammar나 schema가 아니라, 운영 질의 처리 절차 자체를 상태 기계로 외부화한다. 즉, FSC는 모델이 생성한 자유 형식 응답을 사후 검증하는 모듈이 아니라, 질의 분석, 누락 정보 판단, 지원 가능성 확인, 도구 호출, 최종 응답 또는 비응답으로 이어지는 전체 추론 경로를 명시적으로 제한하는 제어기이다. 따라서 본 연구는 기존 constrained generation 및 structured tool-calling 연구와 문제의식은 공유하지만, 산업 SCADA 질의응답에서 필요한 절차 안전성과 운영자 보조 관점의 안전한 비응답을 중심으로 소형 로컬 LLM을 제어한다는 점에서 차별화된다.

한편 LoRA(Low-Rank Adaptation)는 전체 파라미터를 모두 다시 학습하지 않고 작은 저랭크 행렬만 추가로 조정하는 적응 기법으로, 제한된 연산 자원에서도 로컬 미세조정을 가능하게 한다[9]. Qwen3-0.6B와 Llama-3.2-1B-Instruct와 같은 소형 기반 모델은 대화형 추론과 도구 사용의 잠재력을 제공하지만[10, 11], 이러한 일반 능력이 곧바로 지역난방 SCADA 질의응답의 안정성으로 이어지지는 않는다. 실제 배치를 위해서는

소형 모델의 자유도를 줄이고, 정해진 절차 안에서만 도구를 사용하도록 만드는 외부 제어 구조가 필요하다.

이에 본 논문은 지역난방 SCADA 운영 보조를 위한 유한상태제어기(finite-state controller, FSC) 적용 로컬 소형 LLM 기반 AI agent 시스템을 제안한다. 본 연구의 기여는 다음과 같다. 첫째, 실제 산업 현장 데이터에 기반한 XAI4HEAT 질의-행동 trace를 한국어 운영 환경에 맞는 데이터셋으로 재가공하는 규칙을 정의하였다. 둘째, action selection과 payload generation을 분리한 뒤 실제 추론 경로에서는 학습된 정책 대신 FSC를 주 제어기로 채택하였다. 이는 기존 constrained generation 연구처럼 출력 문법만을 제한하는 방식이 아니라, 질의 완전성 점검, 지원 가능성 판단, 도구 호출 순서, clarification 및 abstention 전이를 명시적 상태 기계로 외부화한 설계라는 점에서 차별화된다. 셋째, 범용 API 호출 능력을 평가하는 기존 structured tool-calling benchmark와 달리, 본 연구는 지역난방 SCADA 운영 질의에서 요구되는 안전한 비응답과 절차 준수성을 offline replay 방식으로 평가하였다. 넷째, 단일 RTX 5060 Ti(16GB) 환경에서 LoRA 1 epoch만으로 구동 가능한 로컬 시스템을 구현하고, 규칙 기반 제어기와 소형 모델의 결합이 운영자 보조 관점에서 의미 있는 개선을 제공함을 확인하였다. 이러한 결과는 산업전자 분야에서 AI agent 기술을 적용할 때, 완전 자율화보다 운영자 확인을 전제로 한 보조 계층부터 도입하는 접근이 현실적임을 시사한다.

Ⅱ. 본 론

1. 제안하는 연구의 개요

제안 시스템은 사용자 자연어 질의를 입력받아 질의 파서, FSC, 도구 실행기, payload 생성기 (소형 LLM, e.g. Qwen3-0.6B, Llama-3.2-1B-Instruct), 최종 응답 템플릿으로 이어지는 파이프라인을 수행한다. 먼저 질의 파서는 질의문에서 서브스테이션, 지표, 시간 표현, 비교 여부를 추출한다. 이후 FSC는 현재 상태에서 CALL_TOOL, CLARIFY, ABSTAIN, ANSWER 중 어떤 절차를 선택할지를 결정한다. 여기서 action selection은 다음에 무엇을 할지를 정하는 단계이며, 제안 시스템에서는 이 결정을 작은 모델에 맡기지 않고 외부 제어기에서 수행한다. CALL_TOOL이 선택되면 payload 생성기가 호출된다. payload generation은 선택된 action_code에 대해 도구 실행에 필요한 인자만 JSON 형태로 만드는 단계이다. 예를 들어 특정 시점 값을 조회할 때에는 서브스테이션, 지표, 시각만 생성하고, 평균값 응답을 구성할 때에는 시작 시각, 종료 시각, 계산 결과만 생성하도록 제한한다. 생성된 payload로 도구가 실행되면 결과가 history에 추가되고, FSC는 이를 바탕으로 다음 상태로 전이한다. 마지막으로 최종 사용자 문장은 모델이 직접 자유 생성하지 않고, 시스템이 payload와 도구 결과를 받아 규칙 기반 한국어 문장으로 합성한다. 이러한 절차는 작은 모델의 역할을 좁혀 출력 불안정을 줄이는 동시에, 왜 특정 응답이 나왔는지 설명하기 쉬운 구조를 제공한다. 그림 1은 제안 시스템의 전체 구조도를 나타낸다.

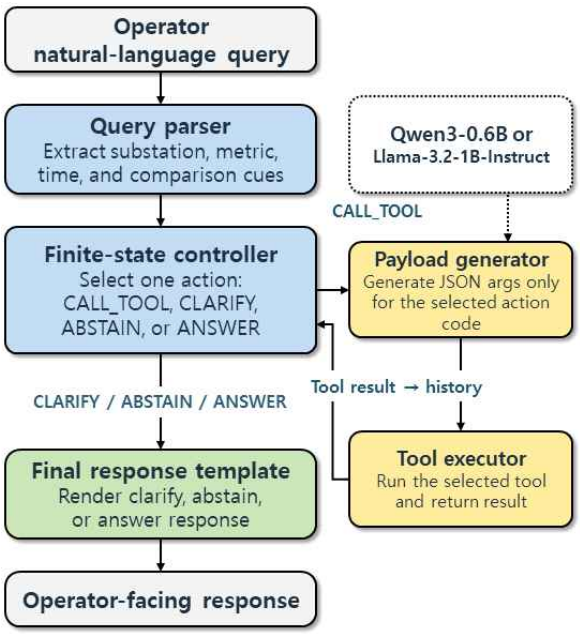

그림 1. 제안 시스템의 전체 구조도
Fig. 1. Overall structure of the proposed system.

2. 시스템 구조

제안 시스템의 핵심은 작은 모델에게 전체 trace를 자유 생성시키지 않고, 제어는 규칙으로 고정하며 생성은 payload로 한정하는 데 있다. 특히 본 연구는 데이터셋 재가공뿐만 아니라, 실제 서비스 경로의 주 제어 로직을 FSC으로 명시화한다. 이를 위해 시스템을 FSC, 로컬 payload 생성기, 재가공 데이터셋, 안전 정책 및 응답 생성의 네 부분으로 구성하였다. 그림 2는 예시 질의와 정답 행동 추적을 나타낸다. FSC는 action들의 순서를 고정하고 payload generator는 각 action에서 JSON 슬롯만을 채운다. 이러한 구조는 비교적 저성능인 소형 AI agent에서도 SCADA 운영 보조를 가능케 한다.

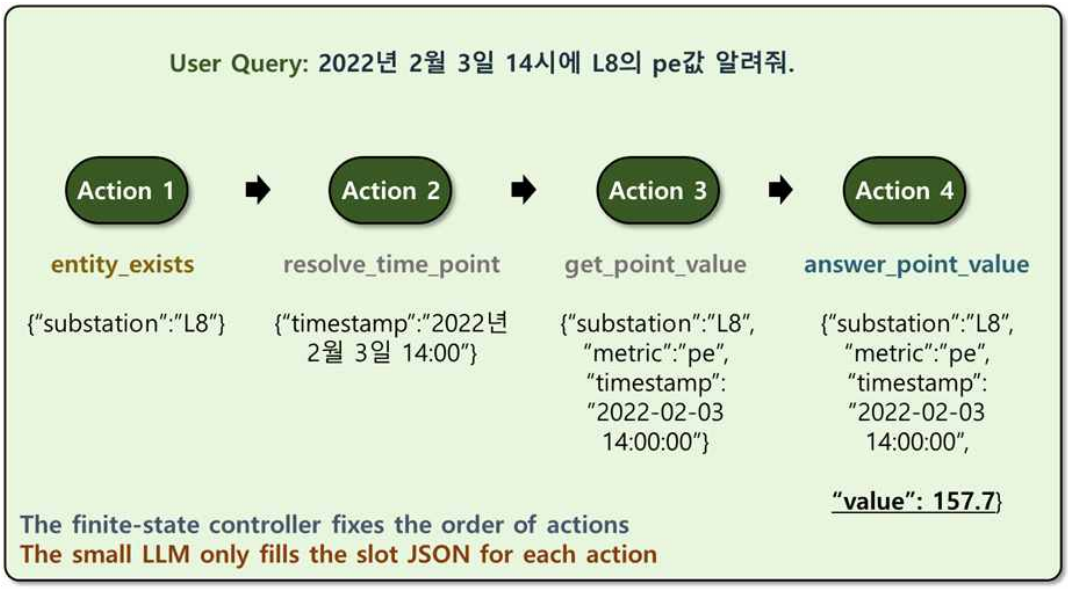

그림 2. 제안 시스템의 예시 질의와 정답 행동 추적
Fig. 2. Example of the overall flow of the proposed system.

제안 시스템의 동작을 구체적으로 설명하기 위해 특정 시점 값 조회 질의를 예로 들 수 있다. 사용자가 "특정 서브스테이션의 특정 시점 공급온도를 알려줘"와 같은 질의를 입력하면, 질의 파서는 먼저 질의문에서 station, metric, time slot을 추출한다. 이후 FSC는 세 slot이 모두 채워져 있는지 확인하고, 정보가 충분한 경우에만 entity exists, resolve time point, get point value, ANSWER 순서로 상태를 전이시킨다. 이때 소형 LLM은 다음 action을 자유롭게 선택하지 않고, FSC가 이미 결정한 action에 필요한 JSON payload만 생성한다. 예를 들어 get point value 단계에서는 station, metric, timestamp와 같은 필드만 생성하며, 최종 사용자 응답은 도구 실행 결과와 고정 템플릿을 이용해 생성된다. 반대로 사용자의 질의에 시간 정보가 누락된 경우, FSC는 get point value로 바로 전이하지 않고 clarify time 상태로 이동한다. 이 경우 시스템은 임의의 시간을 가정하지 않고 조회 시점을 다시 묻는다. 또한 존재하지 않는 서브스테이션이 입력된 경우에는 abstain nonexistent entity로, 지원되지 않는 지표가 입력된 경우에는 abstain unsupported metric으로 전이하여 답변을 생성하지 않는다. 이처럼 제안 시스템은 소형 LLM이 전체 trace를 자유 생성하도록 두지 않고, 질의 완전성 점검, 지원 가능성 판단, 도구 호출 순서, 재질문 및 응답 보류를 FSC의 명시적 상태 전이로 제어한다.

가. 유한상태제어기(FSC)

FSC는 소형 모델이 임의로 행동을 결정하지 못하도록 하고, 사전에 정의된 절차 안에서만 도구를 사용하게 하는 제어 계층으로 설계되었다. 이는 본 연구의 주요 차별화 요소로서, 잘못된 출력을 사후적으로 차단하는 보조 규칙에 머무르지 않고 질의 해석 이후 충족되어야 할 정보 조건, 재질문 및 응답 보류의 선택 기준, 그리고 도구 호출 순서를 명시적으로 규정하는 주 제어기이다. 시스템은 먼저 station, metric, time 정보가 충분히 명시되었는지와 실제 지원 가능한 질의인지 점검한다. 정보가 부족한 경우에는 clarify time, clarify station, clarify metric 상태로 전이하여 누락된 항목을 다시 묻는다. 존재하지 않는 설비, 지원되지 않는 지표, 관측 범위를 벗어난 시간에 대해서는 각각 abstain nonexistent entity, abstain unsupported metric, abstain out of range time으로 정규화하여 답변을 보류한다. 정보가 충분한 경우에만 entity exists, resolve time point, resolve time range, get point value, get timeseries, aggregate window, compare substations와 같은 도구 동작을 순차적으로 호출한다. 이와 같은 상태 전이는 learned policy에서 자주 나타났던 조기 ANSWER와 무효 action sequence를 억제하는 역할을 한다. 또한 clarification과 abstention 문장을 모델이 자유롭게 쓰지 않고 canonical action_code만 선택하게 한 뒤, 실제 사용자에게는 고정 템플릿 문장으로 응답하도록 함으로써 안전성과 설명 가능성을 동시에 확보하

였다. 따라서 본 연구의 FSC는 소형 모델을 산업 현장 절차에 맞게 제한하는 핵심 제어 계층이자, 자유 생성형 agent를 현장형 운영 보조 구조로 전환하는 주요 설계적 기여라 할 수 있다. 그림 3은 FSC의 내부 상태 전이 전체 구조를 나타낸다.

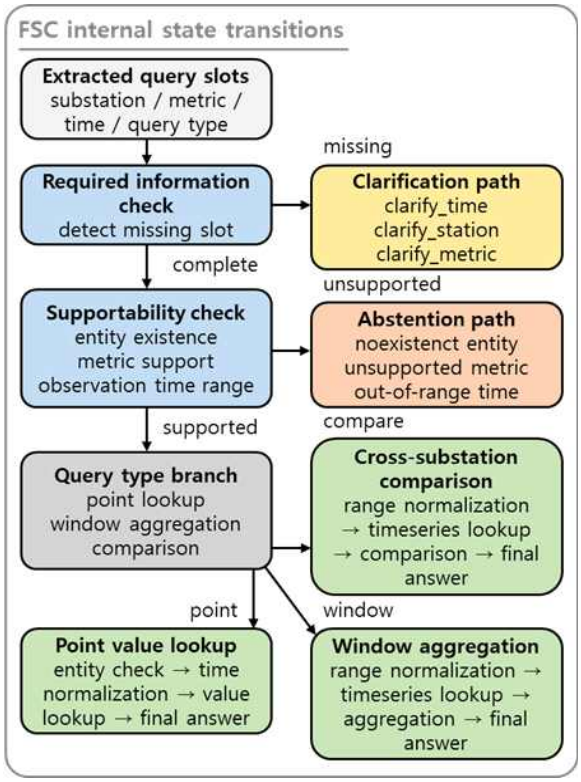

그림 3. FSC의 내부 상태 전이 전체 구조
Fig. 3. Overall structure of the FSC internal state transitions.

나. 로컬 Payload 생성기

Payload 생성기는 FSC이 이미 선택한 action_code에 필요한 인자만 생성한다. 다시 말해, 이 모듈은 무슨 행동을 할지를 판단하지 않고 그 행동에 필요한 인자를 어떻게 채울지에만 집중한다. 본 연구에서는 Qwen3-0.6B와 Llama-3.2-1B-Instruct를 후보 모델로 사용하였다. 예를 들어 get point value에서는 substation, metric, time stamp를, answer window mean에서는 start time, end time, value를 출력한다. 최종 사용자 문장은 생성 모델이 직접 작성하지 않고 시스템이 payload를 받아 규칙 기반 템플릿으로 합성하므로, 작은 모델의 문장 생성 실수가 곧바로 사용자 응답 오류로 이어지는 문제를 줄일 수 있다. 모델 적응은 LoRA를 사용하여 수행하였다. 구체적으로 optimizer는 AdamW[12], learning rate는 0.0002, 스케줄러는 cosine, warmup ratio는 0.05로 설정하였다. 또한 gradient checkpointing을 사용한 상태에서 1 epoch만 학습하여 단일 GPU 환경에서의 실용 가능성을 점검하였다. Qwen 계열은 batch size 4와 gradient accumulation 4, Llama 계열은 batch size 2와 accumulation 8을 사용하였다. 이러한 설계는 소형 로컬 모델이 현장 배치를 위한 계산 제약 안에서 payload 생성 역할을 수행할 수 있는지 검증하기 위한 것이다.

다. 데이터셋 생성 규칙

원본 데이터는 500개의 XAI4HEAT 파생 샘플로 구성된다. 본 연구에서는 동일한 원본 샘플에서 파생된 질의가 학습 세트와 평가 세트에 동시에 포함되는 것을 방지하기 위해, 샘플 식별자 단위로 질의 유형의 비율을 유지하면서 데이터를 분할하였다. 평가 세트는 전체의 10%인 50개 샘플로 구성되며, 특정 시점 값 조회 9개, 기간 집계 15개, 서브스테이션 간 비교 9개, 재질문 12개, 응답 보류 5개가 포함된다. 또한 원본 질의-행동 기록은 한국어 질의응답 흐름에 맞추어 재가공하였으며, 행동 이름 체계를 표준화하여 정리하고, 다음에 수행할 행동을 고르는 학습 데이터와 선택된 행동에 필요한 도구 호출 인자를 작성하는 학습 데이터로 나누었다. 재질문, 응답 보류, 최종 응답에 사용되는 템플릿 역시 한국어 표현으로 정리하였다. 학습 데이터는 다음 행동 선택 역할에 대해 원본 분포 기준 1,495개 예제와 불균형 보정 후 1,865개 예제로, 도구 호출 인자 생성 역할에 대해 원본 분포 기준 1,342개 예제와 불균형 보정 후 1,730개 예제로 재구성하였다. 질의 유형별 예제 수가 한쪽으로 치우치는 문제를 완화하기 위해, 학습 세트에 한해서 부족한 유형의 예제를 일정 한도까지만 복제하여 보강하였다. 평가 세트는 실제 성능이 왜곡되지 않도록 원래 분포를 유지하였다. 이와 같은 분리의 목적은 소형 모델이 전체 JSON 형식의 행동 기록을 한 번에 생성하면서 발생하던 구조 오류를 줄이는 데 있다. 특히 재질문과 응답 보류는 도구 호출 인자 생성 대상에서 제외하고 외부 템플릿으로 처리함으로써 실제 시스템이 요구하는 안정적인 행동 호출 형식을 확보한다.

라. 안전 정책 및 응답 생성

최종 응답은 FSC 상태와 도구 실행 결과에 기반하여 생성된다. Answer-required 질의에 대해서는 숫자 payload와 시간 정보를 규칙 기반 한국어 문장으로 합성하고, clarification은 누락된 slot을 명시적으로 질문하도록 고정하였다. 이는 모델의 문장 생성 오류가 직

접적인 운영 지시로 전달되는 위험을 줄이기 위한 설계이다. 또한 본 시스템은 완전 자율 의사결정 시스템이 아니라 운영자 보조 시스템으로 정의된다. 따라서 unsupported metric, nonexistent entity, out-of-range time에 대해서는 답변을 강행하지 않고 중단하며, 성능이 낮게 나타난 서브스테이션 간 비교 질의는 운영자 확인을 우선하는 것이 바람직하다. 이러한 안전 정책은 답할 수 없는 상황에서는 답하지 않는다는 원칙을 명시적으로 구현한다는 점에서 의미가 있다. 특히 산업 현장에서는 정답처럼 보이는 잘못된 문장보다, 근거 부족을 이유로 한 재질문이나 응답 보류가 더 안전한 선택일 수 있다. 따라서 제안 시스템의 성능 평가는 단순한 정답률뿐 아니라, 얼마나 일관되게 안전한 비응답을 수행하는지도 함께 살펴보아야 한다. 그림 4는 제안 시스템의 안전 정책 및 응답 생성을 나타낸다.

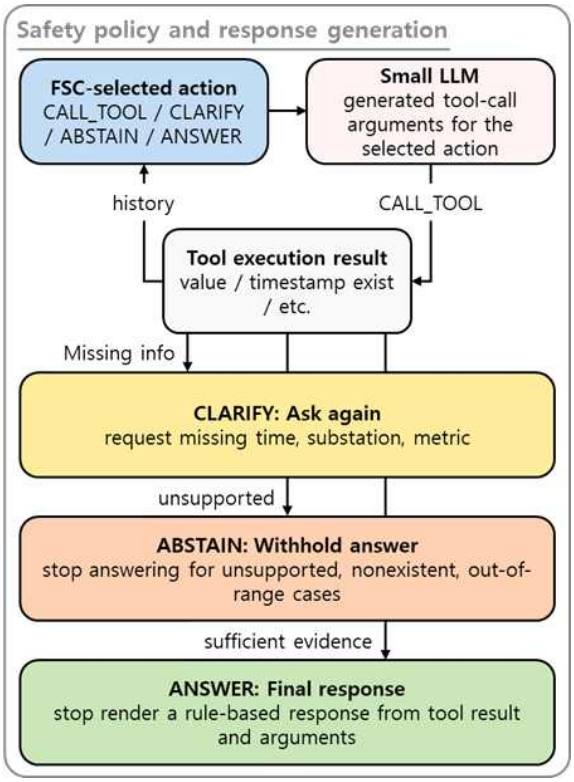

그림 4. 제안 시스템의 안전 정책 및 응답 생성
Fig. 4. Safety policy and response generation of proposed system.

Ⅲ. 실 험

1. 실험 환경

모든 실험은 단일 NVIDIA GeForce RTX 5060 Ti(16GB VRAM) 환경에서 수행하였다. AI agent는 모두 bfloat16과 gradient checkpointing을 사용하였으며, 1 epoch LoRA 미세조정만으로 로컬 학습 가능 여부를 검증하였다. 구현은 python 라이브러리인 transformers, datasets, peft, Trainer를 사용하였다[13~15].

2. 데이터 셋

평가에 사용한 데이터는 실제 지역난방 운영 현장에서 수집된 SCADA 데이터에 기반한 XAI4HEAT [1] 질의-행동 trace 500개를 한국어 운용 환경에 맞추어 재구성한 것이다. 전체 샘플은 train 450개와 eval 50개로 분리되며, eval set의 interaction outcome 분포는 answer 33개, clarify 12개, abstain 5개이다. 실제 시스템 평가는 이 50개 샘플에 대한 offline replay 방식으로 수행하였다. Offline replay는 기록된 질의와 도구 실행 순서를 실제 운영자 상호작용 없이 재생하여 각 단계의 의사결정을 평가하는 방식이다. 본 평가는 일반 대화 벤치마크가 아니라 산업 운전 맥락이 반영된 한글 질의를 대상으로 시스템의 절차 재현 안정성을 측정한다. 모든 tool action과 answer 계열 action만 학습 대상으로 포함하였으며, clarify 계열과 abstain 계열은 외부 템플릿으로 처리하였다. 이 구성은 소형 모델이 문장 생성에 소비하는 자유도를 줄이고 구조화된 JSON interface에 집중하도록 설계되었다. 이와 같은 분리는 소형 모델이 전체 JSON trace를 한 번에 생성하면서 발생하던 schema drift를 줄이는 데 목적이 있다. 특히 clarification과 abstention은 payload 생성 대상에서 제외하고 외부 템플릿으로 처리함으로써, 실제 시스템이 요구하는 안정적 tool call interface를 확보하였다.

3. 실험 및 고찰

본 연구의 성능 평가는 50개 eval 샘플 각각에 대해 gold trace와 시스템이 생성한 trace를 비교하는 offline replay 방식으로 수행하였다. 동일한 사용자 질의와 평가 조건에서 action 선택, payload 생성, 최종 응답 도달 여부를 단계별로 대조함으로써 시스템의 절차 준수 여부와 안전성을 점검하였다. 평가 지표는 과업 완수(task accomplishment, TA), 안전한 비응답(safe non-response), 최종 행동 일치(correct final action), exact trace, 무효 출력(invalid output)으로 구성하였다. 과업 완수는 answer-required 33개 질의에서 최종 ANSWER가 gold final payload와 일치하는 경우로 정의하였다. 이는 단순히 답변처럼 보이는 문장

을 생성했는지를 측정하는 지표가 아니라, 필요한 도구 호출과 값 해석을 거쳐 사용자가 요구한 결과를 끝까지 정확히 산출했는지를 평가하는 엄격한 지표이다. 안전한 비응답은 clarification 또는 abstention이 필요한 17개 질의에서 시스템이 같은 유형의 비응답을 선택한 경우로 정의하였다. 최종 행동 일치와 exact trace는 각각 마지막 단계의 의사결정 일치성과 전체 절차 일치성을 점검하기 위한 보조 지표이며, 무효 출력은 시스템 실행이 불가능한 형식 오류를 측정한다. 이러한 다층적 평가는 산업 현장에서 중요한 “정확히 답하기”와 “무리하게 답하지 않기”를 함께 검증하기 위해 필요하다. 비교 대상 시스템은 학습 및 제어 방식에 따라 구성하였다. Prompt-only는 별도의 추가 학습 없이 기본 LLM에 프롬프트만 제공하여 action과 payload를 생성하는 방식이다. Fine-tuned는 전체 모델 파라미터를 대상으로 학습 데이터에 맞게 supervised fine-tuning을 수행한 방식이며, LoRA는 저랭크 어댑터를 이용해 일부 추가 파라미터만 학습하는 parameter-efficient fine-tuning 방식이다. FSC는 본 논문에서 제안하는 finite state controller로, 현재 상태에서 허용 가능한 action 전이와 출력 형식을 제한함으로써 모델의 절차적 오류와 무효 출력을 줄이기 위해 사용하였다. 따라서 FSC+Prompt-only, FSC+Fine-tuned, FSC+LoRA는 각각의 LLM 생성 방식 위에 FSC를 결합한 시스템을 의미한다. 구성 요소 수준 평가에서는 learned policy의 안정성이 충분하지 않음이 확인되었다. Qwen의 decision accuracy와 action_code accuracy는 각각 53.3%와 31.5%였고, Llama의 decision accuracy와 action_code accuracy는 각각 60.6%와 32.1%였다. 이는 전체 시스템의 병목이 payload 생성 자체보다, 우선 어떤 절차를 선택할지 결정하는 action selection 단계에 더 크게 존재함을 시사한다. 표 1은 시스템 수준의 offline replay 평가 결과를 나타낸다. Prompt-only 설정에서는 Llama와 Qwen 모두 TA 0건/33건, Final Action 0건/50건, Safe Non-response 0건/17건을 기록하여, 프롬프트만으로는 본 데이터셋의 절차적 요구사항을 충족하기 어려웠다. 또한 Prompt-only Llama는 Invalid 27건, Prompt-only Qwen은 Invalid 13건으로 무효 출력도 많이 발생하였다. Fine-tuned 설정에서도 과업 완수는 개선되지 않았다. Fine-tuned Llama는 TA 0건/33건, Final Action 4건/50건, Safe Non-response 4건/17건, Invalid 23건을 기록하였고, Fine-tuned Qwen은 TA 0건/33건, Final Action 1건/50건, Safe Non-response 1건/17건, Invalid 9건을 기록하였다. LoRA 설정에서는 Qwen 기반 모델에서 TA 2건/33건이 나타났으나, 전체적으로는 여전히 낮은 과업 완수율을 보였다. FSC를 결합한 경우에는 전반적으로 Final Action, Safe Non-response, Invalid output에서 개선이 관찰되었다. 특히 FSC+LoRA Qwen은 TA 8건/33건 (24.2%), Final Action 22건/50건 (44.0%), Safe Non-response 14건/17건 (82.4%), Invalid 1건을 기록하여 전체 시스템 중 가장 우수한 성능을 보였다. FSC+LoRA Llama는

표 1. 시스템 수준 offline replay 평가 결과
Table 1. System-level offline replay evaluation results.

System	Main Performance	Safety / Validity
Prompt-only Llama	TA 0/33, Final Action 0/50	Safe Non-response 0/17, Invalid 27
Prompt-only Qwen	TA 0/33, Final Action 0/50	Safe Non-response 0/17, Invalid 13
Fine-tuned Llama	TA 0/33, Final Action 4/50	Safe Non-response 4/17, Invalid 23
Fine-tuned Qwen	TA 0/33, Final Action 1/50	Safe Non-response 1/17, Invalid 9
LoRA+Llama	TA 0/33, Final Action 4/50	Safe Non-response 9/17, Invalid 7
LoRA+Qwen	TA 2/33, Final Action 7/50	Safe Non-response 9/17, Invalid 6
FSC+ Prompt-only+Llama	TA 0/33, Final Action 7/50	Safe Non-response 4/17, Invalid 13
FSC+ Prompt-only+Qwen	TA 3/33, Final Action 9/50	Safe Non-response 4/17, Invalid 9
FSC+Fine-tuned Llama	TA 0/33, Final Action 6/50	Safe Non-response 13/17, Invalid 10
FSC+Fine-tuned Qwen	TA 4/33, Final Action 9/50	Safe Non-response 11/17, Invalid 6
FSC+LoRA Llama (Ours)	TA 0/33, Final Action 15/50	Safe Non-response 15/17, Invalid 5
FSC+LoRA Qwen (Ours)	TA 8/33, Final Action 22/50	Safe Non-response 14/17, Invalid 1

표 2. 제안 시스템(FSC+Qwen)의 answer-required 질의 유형별 결과
Table 2. Results by answer-required query type of the proposed system (FSC+Qwen).

Query Type	Number of Successful Cases	Notes
Point lookup	3 / 9	Some success on single-timestamp value lookup tasks
Temporal aggregation	5 / 15	Performed best on period-average queries
Cross-substation comparison	0 / 9	Comparison queries remain an unresolved challenge

TA는 0건/33건에 머물렀으나, Final Action 15건/50건 (30.0%), Safe Non-response 15건/17건 (88.2%), Invalid 5건을 기록하여 FSC가 안전한 비응답과 출력 형식 안정성 측면에서 기여함을 보여준다. 물론 FSC+LoRA Qwen의 TA 8건/33건 (24.2%)은 절대적으로 높은 성능이라고 보기는 어렵다. 따라서 제안 시스템을 완전 자율 운영 시스템으로 해석하기에는 한계가 있다. 다만 본 평가의 TA가 최종 payload의 정확한 일치까지 요구하는 엄격한 기준이라는 점을 고려하면, 제안 방식은 단순한 문장 생성 품질이 아니라 실제 절차 수행 관점에서 제한적이지만 의미 있는 개선을 보였다고 해석할 수 있다. 특히 Prompt- only Qwen의 Invalid 13건, Fine-tuned Qwen의 Invalid 9건, LoRA+Qwen의 Invalid 6건과 비교할 때, FSC+LoRA Qwen에서 Invalid가 1건으로 감소한 점은 중요하다. 이는 FSC가 LLM의 자유 생성 과정에서 발생하는 실행 불가능한 출력을 억제하는 데 효과가 있음을 보여준다. 또한 Safe Non-response가 14건/17건 (82.4%)으로 높게 유지된 점은, 산업 현장에서 운영자 확인을 전제로 사용하는 보조시스템의 관점에서 중요한 장점이다.

세부 과업 유형별로는 특정 시점 값 조회에서 3건/9건 (33.3%), 기간 집계에서 5건/15건 (33.3%)의 과업 완수를 달성하였다. 반면 서브스테이션 간 비교는 0건/9건 (0%)로 매우 낮았다. 이는 여러 서브스테이션의 상태를 동시에 해석하고 비교 기준을 일관되게 적용하는 작업에서 현재 시스템의 한계가 큼을 나타낸다. 따라서 현 단계의 적용 가능 범위는 특정 시점 값 조회와 기간 집계 중심의 운영 보조 시나리오로 제한되며, 서브스테이션 간 비교에는 별도의 규칙 모듈, 비교 전용 planner, 또는 후속 학습 전략이 필요하다. 모호성 해소형 질의에는 비교적 안정적으로 대응하였다. 구체적으로 시간 정보 재질문은 5건/5건 (100%), 설비 정보 재질문은 3건/3건 (100%), 지표 정보 재질문은 4건/4건 (100%)를 기록하였다. 이는 누락 정보를 식별하고 재질문하는 과정에서 일정 수준의 일관성을 확보했음을 의미한다. 산업 현장에서는 이러한 재질문 능력 자체가 운영자와의 협업 품질을 좌우하므로, 낮은 과업 완수율과는 별개로 평가할 가치가 있다. 종합하면, 제안한 FSC+LoRA 구조는 아직 제한적인 과업 완수 성능을 보이지만, 출력 유효성, 최종 행동 선택, 안전한 비응답 측면에서 기존 Prompt-only 및 단순 학습 기반 접근보다 안정적인 절차 수행 가능성을 보였다.원했다는 점에서, 로컬 소형 LLM 기반 AI agent를 산업 현장 보조 시스템으로 적용하기 위한 현실적 출발점을 제시한다. 특히 특정 시점 값 조회와 기간 집계 중심의 반복 질의 보조, 그리고 모호한 질의에 대한 재질문 유도 측면에서는 산업 현장의 운영 지원 도구로 연결될 여지가 있다.

Ⅳ. 결 론

본 논문에서는 지역난방 SCADA 운영 질의응답을 위한 유한상태제어기 기반 로컬 소형 LLM 시스템을 제안하였다. XAI4HEAT 기반 질의-행동 trace를 재가공하여 데이터셋을 구성하고, 실제 추론 경로에서는 FSC가 action selection을 담당하고 LoRA로 미세조정한 소형 LLM이 JSON 슬롯만 채우는 payload generation만 수행하도록 설계하였다. 오프라인 replay 평가에서 end-to-end learned pipeline은 Qwen과 Llama 모두 과업 완수가 0건/33건으로 불안정했으나, 제안 방식인 FSC+Qwen은 8건/33건의 개선된 과업 완수 개수와 14건/17건의 안전 비응답을 달성하였다. 이러한 개선은 FSC가 action 선택과 도구 호출 순서를 명시적으로 제한하고, 재질문과 응답 보류를 안정적인 상태 전이로 처리한 결과로 해석된다. 즉, 본 연구의 핵심 기여는 성능 향상보다, 산업 현장에서 요구되는 절차성과 안전성을 외부 제어 구조로 보완한 데 있다. 본 연구의 산업적 의의는 AI agent를 이용한 완전 자동화가 아니라, 실제 현장 데이터, 로컬 연산 환경, 운영자 확인 절차를 함께 고려한 운영 보조 시스템을 제시했다는 데 있다. 특히 FSC는 소형 LLM 기반 AI agent를 현장 운영 지원 도구로 사용하기 위한 안전장치로 기능하며, 특정 시점 값 조회, 기간 집계, 모호한 질의에 대

한 재질문과 같은 반복적 운영 업무에서 운영자의 판단을 보조할 수 있다. 이러한 해석은 산업전자 분야에서 AI agent 기술을 적용할 때에도 중요하다.

향후에는 데이터셋 생성 규칙 보강, 최종 응답 정책의 완전한 템플릿화, multi-turn 온라인 사용자 평가, 그리고 더 다양한 산업 현장 데이터에 대한 검증을 통해 시스템의 적용 범위를 확대할 예정이다.

REFERENCES

[1] S. Cvetković, M. Zdravković, and M. Ignjatović, “Exploring district heating systems: A SCADA dataset for enhanced explainability,” Data in Brief, vol. 59, 111320, 2025.

[2] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J.-R. Wen, “A Survey on Large Language Model based Autonomous Agents,” Frontiers of Computer Science, vol. 18, no. 6, 186345, 2024.

[3] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models,” in Proc. International Conference on Learning Representations (ICLR), 2023.

[4] S. Geng, M. Josifoski, M. Peyrard, and R. West, “Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning,” arXiv preprint arXiv:2305.13971, 2023.

[5] S. Ugare, T. Suresh, H. Kang, S. Misailovic, and G. Singh, “SynCode: LLM Generation with Grammar Augmentation,” arXiv preprint arXiv:2403.01632, 2024.

[6] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, D. Li, Z. Liu, and M. Sun, “ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs,” in Proc. International Conference on Learning Representations (ICLR), 2024.

[7] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, “Gorilla: Large Language Model Connected with Massive APIs,” in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2024.

[8] M. Li, Y. Zhao, B. Yu, F. Song, H. Li, H. Yu, Z. Li, F. Huang, and Y. Li, “API-Bank: A Comprehensive Benchmark for Tool-Augmented LLMs,” in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3102–3116, 2023.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” arXiv preprint arXiv:2106.09685, 2021.

[10] A. Yang et al., “Qwen3 Technical Report,” arXiv preprint arXiv:2505.09388, 2025.

[11] Meta, “Llama 3.2 collection and Llama-3.2-1B-Instruct model card,” Hugging Face Model Card, Sept. 2024.

[12] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in Proc. International Conference on Learning Representations (ICLR), 2019.

[13] T. Wolf et al., “Transformers: State-of-the-Art Natural Language Processing,” in Proc. EMNLP: System Demonstrations, pp. 38–45, 2020.

[14] Q. Lhoest et al., “Datasets: A Community Library for Natural Language Processing,” in Proc. EMNLP: System Demonstrations, pp. 175–184, 2021.

[15] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, and M. Tietz, “PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods,” Hugging Face, 2022.

저 자 소 개

김 정 윤(학생회원)
2020년 국립한밭대학교 전자공학과 학사 졸업.
2022년 국립한밭대학교 전자공학과 석사 졸업.
2022년~현재 국립한밭대학교 전자공학과 박사과정.
<주관심분야 : 감정인식, 딥러닝>

이 승 호(평생회원)
1986년 한양대학교 전자공학과 학사 졸업.
1989년 한양대학교 전자공학과 석사 졸업.
1994년 한양대학교 전자공학과 박사 졸업.
1994년~현재 국립한밭대학교 전자공학과 교수
<주관심분야 : 영상신호처리, 딥러닝, AR>