AI MACROFY: 클라우드 및 온디바이스 멀티모달 LLM을 활용한 안드로이드 UI 에이전트 설계 및 성능 비교 연구

최세진, 오찬영*

국립공주대학교

marin6670@smail.kongju.ac.kr, *cyoh@kongju.ac.kr

A Study on the Design and Performance Comparison of Android UI Agents using Cloud and On-device Multimodal LLMs

Sejin Choi, Chanyoung Oh*

Kongju National Univ.

요 약

본 논문에서는 클라우드 및 온디바이스 LLM을 활용하여 안드로이드 기기를 자연어로 제어하는 UI 에이전트 애플리케이션 'AI Macrofy'를 설계하고 구현하였다. 기존 연구에서 다중 에이전트나 복잡한 파이프라인을 요구했던 것과 달리, 본 시스템은 모바일 기기의 제한된 연산 자원을 고려하여 단일 LLM 에이전트가 'Observation-Analysis-Action'의 구조화된 DSL출력을 따르도록 제약하여 파이프라인 복잡도를 낮추고 실용성을 확보하였다. 에이전트의 기반 모델을 클라우드 모델(Gemini 3 Pro, GPT-5.2)과 온디바이스 모델(Gemma 3n E4B)로 선정하고 실제 스마트 기기 환경에서 Android World 벤치마크의 10개 작업을 선정하여 비교 평가한 결과, 클라우드 모델의 경우 복합적인 작업에서도 의미 있는 작업 성공률을 달성하였다. 그러나 온디바이스 환경에서는 추론 지연과 문맥 유지 한계로 인해 실시간 상호작용성이 제한 되며 복잡한 작업에서 낮은 성공률을 보이는 것을 확인하였다.

Ⅰ. 서 론

최근 LLM 기반의 AI 에이전트 연구가 활발하게 진행되며 모바일 환경에서도 자연어 명령으로 스마트폰을 제어하려는 시도가 이어지고 있다. AutoDroid는 안드로이드 환경에서 LLM과 앱 특화 지식, UI 트리를 결합하여 작업 자동화가 가능함[1]을 보였고 AppAgent는 스크린샷을 인식하는 멀티모달 LLM 기반 에이전트를 통해 추론의 정확도, 성공률을 높일 수 있음[2]을 보였다. 한편, 기존 연구에서 대용량 모델 대비 온디바이스 소형 모델이 복잡한 추론과 다단계 작업 수행에서 현저한 성능 저하가 지적되고 있다. Haque et al.는, 3B 파라미터 이하의 SLLM은 단일 명령 수행에서는 70% 이상의 성공률을 보였으나, 문맥 유지가 필수적인 다중 턴(Multi-turn) 시나리오에서는 성공률이 20% 미만으로 급격히 저하됨[3]을 보고하였다.

본 논문에서는 Google 사에서 제공하는 Gemini 3 Pro 모델, Open AI의 GPT-5.2 모델과 온디바이스 추론 모델로 Google의 Gemma 3n E4B[5] 모델을 활용하여 안드로이드 환경에서 동작하는 AI 에이전트 애플리케이션을 구현한다. 성능 평가는 모바일 에이전트 분야에서 널리 사용되는 Android World 벤치마크[4]의 116개의 작업 요청 중 10개의 작업을 선별하여 수행하였으며, 클라우드 모델과 온디바이스 모델의 성능을 비교하여 온디바이스 추론성능의 한계를 확인하고자 한다. 이를 통해, 향후 상용 수준의 온디바이스 에이전트를 구현하기 위해 해결해야 할 엔지니어링 과제를 제시한다.

Ⅱ. 본론

2.1 AI Macrofy 시스템 개요

본 논문의 Agent 시스템은 LLM의 비정형 텍스트 출력을 안드로이드 애플리케이션이 해석할 수 있도록 변환하기 위해 도메인 특화언어(DSL-Domain-Specific-Language)를 정의하여 사용한다.

전체 시스템은 그림1과 같이 크게 Input Context Builder, AI Interface, Agent Response Parser 그리고 Action Execution 모듈로 구성된다.

본 시스템은 사용자의 자연어 요청을 입력받아 확률적 추론 능력을 갖춘 LLM 모델에 휴대폰의 제어 권한을 위탁함으로써, 사용자의 직접적인 개입 없이 모바일 기기를 자동 제어하는 기능을 수행한다.

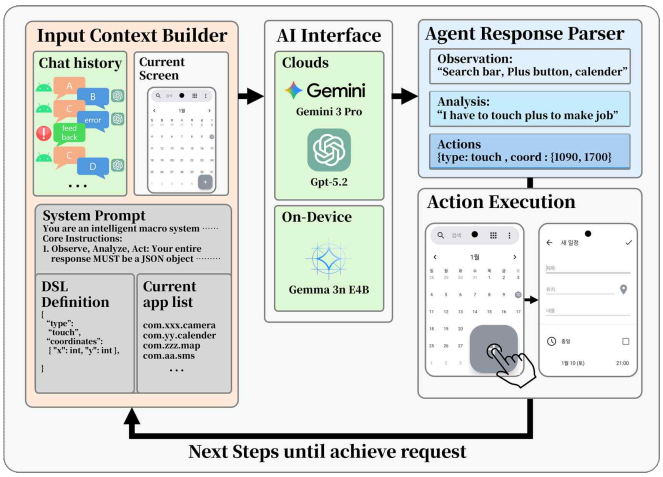

그림 1 AI Macrofy 시스템 구성도

2.1.1. Action Execution

Action Execution 모듈은 안드로이드의 접근성 권한으로 터치, 길게 누르기, 더블 탭, 텍스트 입력, 스와이프, 스크롤, 끌어놓기, 제스처, 대기하기, 앱 열기, 토스트 알림 띄우기 등 12개 명령을 수행하도록 구성하였다.

본 모듈은 정해진 규칙을 따르는 JSON 객체에 의존하여 UI 제어 명령을 수행한다. 따라서 JSON 형식을 지키지 않거나 잘못된 제어요청으로 오류가 발생하였을 때 이를 Chat history에 반영한다. 이는 AI Interface 모듈의 출력이 잘못되었었음을 다시 입력으로 제공하여 다른 출력이 유도되도록 하는 피드백 장치이다.

2.1.2. Input Context Builder

Input Context Builder 모듈은 LLM이 사용자 요청에 따른 정형화된 명령을 생성하도록 Chat history, 현재 화면, 시스템 프롬프트를 관리한다. Chat history는 AI가 매 스텝, 어떤 행동을 수행했었는지에 대한 맥락을 갖도록 대화를 기록한다. Action Execution 단계에서 AI의 잘못된 출력으로 오류가 발생하면 그림1에서 초록색 말풍선과 같이, 오류를 Chat history로 기록하여 피드백을 형성한다. 이를 통해 다음 출력시에는 동일 다른 출력이 유도된다.

시스템 프롬프트는 Action Execution 모듈에서 사용되는 모든 JSON 객체의 예시와 각종 규칙과 UI 기본 요소들을 해석하는 기반 지식이 제공되어, 멀티모달 에이전트가 휴대폰 화면에 대한 해석력을 더 가질 수 있도록 하였고, 현재 휴대폰에 설치되어 있는 앱의 리스트를 함께 제공하여 즉시 AI가 원하는 앱을 찾을 수 있도록 한다.

2.1.3. AI Interface & Agent Response Parser

AI Interface 모듈은 다양한 모델을 교체하기 용이하도록 LLM 접근을 추상화한 계층이다. 클라우드 서비스인 Gemini 3 Pro 모델과 GPT-5.2 모델을 포함하여, 클라우드/온디바이스 모델이 늘어나도, 일관성 있게 Input Context를 AI 모델에 전송하고 Agent Response를 얻을 수 있다. 온디바이스 추론은 MediaPipe와 LITERT-LM 형식의 모델을 지원한다.

Agent Response Parser 모듈은 Response로부터 발생할 수 있는 다양한 예외상황을 처리하여 내부에서 Action Execution 모듈에 전달한 유효한 실행 명령만을 추출하는 역할을 갖는다.

2.2 성능 평가

2.2.1 지표 및 데이터 셋

에이전트의 성능을 객관적으로 측정하고 평가하기 위해, 모바일 GUI 에이전트 연구의 표준 벤치마크로 널리 활용되는 Google Research의 AndroidWorld 데이터 셋[4]을 사용하였다. AndroidWorld는 AI 에이전트의 성능을 평가하기 위해 116개 작업으로 구성된 벤치마크이며 easy, medium, hard로 구분된 다양한 난이도의 작업들을 포함한다.

본 실험에서는 전체 데이터셋 중 난이도별로 easy 3개, medium 4개 hard 3개 총 10개의 대표 작업을 선정하여 각 모델당 1회씩 수행하였고, 성능 측정 지표는 다음과 같이 정의하였다.

- 작업 성공률: 에이전트가 목표 상태에 정확히 도달했는지를 백분율로 환산한 수치로, 에이전트의 수행 신뢰성을 나타낸다.
- 평균 소요 시간: 각 작업의 시작부터 작업 완료까지 소요된 전체 시간을 측정하여 실시간성을 평가한다.
- 평균 추론 시간: 각 추론이 시작되고 결과를 얻기까지 걸린 시간의 평균
- 사용 액션 수: 10개의 작업을 수행하는 동안 수행한 각각의 액션(터치, 스크롤, 입력) 의 횟수이다.

2.2.2 실험 환경

본 실험은 Samsung Galaxy Z Flip 7 기기에서 수행되었으며, 구체적인 하드웨어 사양은 엑시노스 2500 프로세서와 12GB LPDDR5X RAM을 포함하며, Android 16 (One UI 8.0) 환경에서 구동되었다. 온디바이스 모델은 오픈소스 모델 Gemma-3n-E4B-it-litert-lm을 사용하였다. 해당 모델은 에이전트 애플리케이션에서 LiteRT-LM 라이브러리를 통해 GPU 가속을 활성화하여 모바일 추론성능을 최적화하였다. Gemini 3 Pro와 GPT-5.2모델은 각각 Google과 OpenAI가 제공하는 API를 통해 클라우드에서 연산이 수행되었다.

2.2.3 결과 분석

모델	작업 성공률	평균 소요시간	평균 추론 시간	사용 액션 수
Gemini 3 Pro	80%	357s	20.7s (api)	총 364회
GPT-5.2	70%	170s	7.6s (api)	총 190회
Gemma 3n E4B	0%	n/a	27.3s (온디바이스)	n/a

표 1 모델별 작업 성공률, 평균 소요시간, 평균 추론 시간, 사용 액션 수

실험 결과, 작업 성공률이 높은 모델은 Gemini 3 Pro였지만, GPT-5.2는 더 짧은 추론 시간과 적은 액션으로 유사한 성과를 보였다. 반면, Gemma 3n E4B는 카메라 앱 실행 등의 기본 동작에 대해 DSL을 지키지 않는 출력으로 인하여 올바른 DSL 출력이 나올 때까지 계속 추론하며 결국 0%의 작업 성공률을 기록하였으며 추론 시간은 약 27초로 가장 파라미터 수가 적은 모델임에도 온디바이스 자원의 한계로 가장 긴 추론 시간을 기록하였다.

Ⅲ. 결론

본 논문에서는 안드로이드 환경에서 동작하는 단일 LLM 기반 UI 제어 에이전트 'AI Macrofy'를 설계 및 구현하고, 최신 클라우드 모델 과 온디바이스 모델의 성능을 비교 분석하였다. 실험 결과, 클라우드 모델은 평균 75% 정도의 작업 성공률을 기록하며 복합적인 작업 수행 능력을 입증하였다. 그러나 온디바이스 모델은 문맥 유지 능력의 한계와 DSL 정의를 벗어나는 출력으로 0%의 작업 성공률을 보였으며 추론 시간 또한 클라우드 모델 대비 최대 3.6배 가량 지연될 수 있음을 확인하였다.

본 연구를 통해, 온디바이스 모델이 개인정보 보호 및 AI 비용 측면에서 이점을 가짐에도 불구하고, 상용 가능한 수준의 성능을 확보하기 위해서 엄격한 DSL 준수 능력 확보와 연산 경량화를 통하여 추론 지연시간 단축이 선행되어야 함을 확인하였다.

ACKNOWLEDGMENT

본 연구는 산림청(한국임업진흥원) 산림과학기술 연구개발사업 '(RS-2025-25441817)'의 지원에 의하여 이루어진 것입니다.

참 고 문 헌

[1] H. Wen., "AutoDroid: LLM-powered Task Automation in Android," ACM MobiCom '24: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking , pp. 543-557.

[2] C. Zhang et al., "AppAgent: Multimodal Agents as Smartphone Users," in The Twelfth International Conference on Learning Representations (ICLR), 2024

[3] M. A. Haque et al., "TinyLLM: Evaluation and Optimization of Small Language Models for Agentic Tasks on Edge Devices," arXiv preprint arXiv:2511.22138, 2025.

[4] C. Rawles et al., "AndroidWorld: A Dynamic Benchmarking Environment for Autonomous Agents," arXiv preprint arXiv:2405.14573, 2024.

[5] Google DeepMind, "Gemma 3: Multimodal open models built on Gemini," Google DeepMind Technical Report, 2025.