

• 특집 • AI 및 데이터 기반 적층제조 공정 최적화와 응용(AI and Data-driven Optimization and Applications in Additive Manufacturing Processes)

적층제조에서 작업자 개입이 대형 언어 모델 기반 공정 제어에 미치는 영향 분석

Experimental Study on the Effect of Human-in-the-loop Integration on Large Language Model-based Process Control in Additive Manufacturing

전성윤¹, 김태환², 김남훈^{1,2,#}Seongyoon Jeon¹, Taehwan Kim², and Namhun Kim^{1,2,#}¹ UNIST 인공지능대학원 (Graduate School of Artificial Intelligence, UNIST)² UNIST 기계공학과 (Department of Mechanical Engineering, UNIST)

Corresponding Author / E-mail: nhkim@unist.ac.kr, TEL: +82-52-217-2715

ORCID: 0000-0003-4429-2191

KEYWORDS: Fused deposition modeling (용융 적층 모델링), Large language model (대형 언어 모델), Human-in-the-loop (인간 참여 루프), Process parameter control (공정 변수 제어)

This study proposes a human-in-the-loop framework that integrates operator observations into a large language model (LLM) to control process parameters for defect handling in fused deposition modeling (FDM) 3D printing. Fully autonomous LLM-based control handles ambiguous sensor data poorly and cannot detect abnormal conditions that lie beyond the installed sensors. Operator observations may compensate for these limitations, but their actual impact on LLM decision-making has not been sufficiently validated. We therefore implemented the proposed framework and defined experimental scenarios involving erroneous parameter injection and environmental disturbances. The framework was evaluated in terms of LLM response quality and print quality. The LLM achieved over 80% response quality on the defined evaluation metrics and generated appropriate parameter adjustments, improving print quality by more than 55% on average. Comparative experiments further revealed that, without operator observations, the LLM sometimes failed to recognize defects. These findings demonstrate the effectiveness of human-LLM collaboration and provide a practical foundation for intelligent FDM process control.

Manuscript received: April 30, 2026 / Revised: June 1, 2026 / Accepted: June 7, 2026

1. 서론

적층제조는 복잡한 형상의 맞춤형 제품을 최소한의 재료 낭비로 생산할 수 있어 항공우주, 의료 등 다양한 산업 분야에서 빠르게 확산되고 있다[1]. 그중 Fused Deposition Modeling (FDM)은 열가소성 필라멘트를 노즐을 통해 압출·적층하는 방식으로, 낮은 장비 비용과 높은 재료 호환성을 바탕으로 널리 활용되는 적층제조 공정 중 하나이다[2]. 그러나 FDM 공정의 출력 품질은 베드 온도,

프린팅 속도 등 다양한 공정 변수와 주변 환경 조건의 복합적 상호작용에 의해 결정되며, 각 공정 변수가 최적 값으로 설정되지 않을 경우 Warping, Over-extrusion과 같은 결함이 빈번히 발생한다[3,4].

출력 품질을 안정적으로 유지하기 위해서는 공정 중 결함을 신속히 감지하고 대응하는 것이 중요하다. 최근에는 이를 위한 접근으로 Large Language Model (LLM)의 적층제조 공정 적용이 주목받고 있다. LLM은 방대한 데이터로 학습된 범용 모델로서, 다양한 도메인 지식과 추론 능력을 갖추고 있다. 이러한 능력을 바

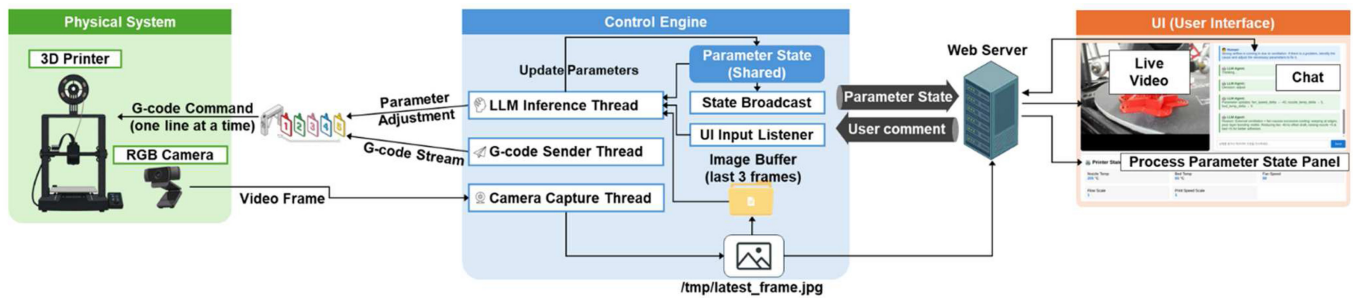


Fig. 1 Overview of a real-time FDM process parameter control framework integrating human operator and large language model (LLM) collaboration. The framework consists of three main components: (a) Physical system (3D printer and RGB camera), (b) Control engine (LLM inference, G-code sender, camera capture, UI input listener, and state broadcast threads), and (c) User interface for real-time monitoring and operator interaction

탕으로 공정 상태를 해석하고 판단 근거와 대응 방안을 자연어로 설명하며, 변화하는 공정 환경에 유연하게 대응할 수 있는 잠재력을 지닌다. Chandrasekhar 등은 LLM과 적층제조 도메인 지식베이스를 결합한 AMGPT를 통해 적층제조 질의 응답 시스템을 제안하였으며[5], Jadhav 등은 LLM-3D Print를 제안하며 서로 다른 종류의 필라멘트와 프린터 장비에서도 자율 제어를 통한 실시간 공정 변수 제어가 가능함을 보였다[6]. 이러한 연구들은 적층제조 분야에서 LLM의 유용성과 범용성을 보여주었다. 다만, 노이즈가 포함된 입력 데이터나 설치된 센서로 감지하기 어려운 이상 상황에서는 공정 상태 파악이 제한되므로, 이에 대응하기 위한 방안은 여전히 필요하다.

센서 기반 입력 한계를 보완하기 위해서는, 보다 다양한 센서로의 확장이나[7], 노이즈가 포함된 데이터에서도 공정 상태를 신뢰성 있게 파악할 수 있는 센서 외 추가 정보원 도입이 필요하다. 이러한 맥락에서 작업자의 관찰 정보를 시스템에 통합하는 방안을 고려할 수 있다. 작업자는 냄새, 소리, 진동 등 다양한 감각 정보를 통해 현장 상태를 종합적으로 인지하고 자연어로 표현할 수 있다. 반면, 해당 능력을 센서로 대체하기 위해서는 높은 경제적 비용과 시스템 복잡성이 요구된다. 이를 고려할 때, 작업자의 관찰 정보를 LLM 기반 공정 모니터링 시스템에 통합하는 접근은 센서 한계를 보완하기 위한 실용적 대안이 될 수 있다.

본 연구는 FDM 공정에서 센서 기반 입력의 한계를 보완하기 위한 접근으로, 작업자의 관찰 정보를 LLM의 추론 과정에 통합하는 Human-in-the-loop 기반 프레임워크를 제안한다. 제안된 프레임워크의 유효성을 검증하기 위해 이상 시나리오를 구성하고, 결함 인지 및 적절한 공정 변수 조정 성능을 평가하였다. 또한 출력물 품질에 미치는 영향을 정량적으로 평가하였다. 더불어 작업자 관찰 정보 제공 여부에 따른 응답 차이를 비교함으로써 작업자 개입이 LLM의 추론에 미치는 영향을 분석하였다. 본 연구의 기여는 다음과 같다:

- (1) 작업자의 관찰 정보를 자연어 형태로 입력 받아 LLM의 추론 과정에 결합하는 Human-in-the-loop 기반 실시간 FDM 공정 변수 제어 프레임워크를 설계하고 구현하였다.
- (2) 제안한 프레임워크를 적용한 실험 결과를 통해 시나리오

별 응답 품질과 출력물 품질을 정량적으로 평가하고, 관찰 정보 제공 여부에 따른 응답 차이를 비교·분석하였다.

2. 제안 프레임워크

본 프레임워크는 FDM 공정 중 작업자의 자연어 입력과 LLM의 추론을 결합하여 공정 변수를 실시간으로 제어하도록 설계되었다. 전체 프레임워크 구조는 Fig. 1에 제시하였다.

2.1 LLM 기반 의사결정 프로세스 설계

본 연구에서는 LLM을 공정 제어를 위한 의사결정 엔진으로 활용한다. 공정 중 작업자가 관찰 정보를 입력하면, LLM은 주어진 정보를 종합적으로 분석하여 공정 변수 조정 여부와 목표 값을 결정하고, 그에 대한 판단 근거를 함께 출력한다(Fig. 2). LLM에 전달되는 프롬프트는 시스템 프롬프트와 사용자 프롬프트로 구성되며, 각각 2.1.1절과 2.1.2절에 상세히 설명한다.

2.1.1 시스템 프롬프트 설계

범용 LLM은 별도의 지침 없이 특정 도메인의 제어 시스템에 적용될 경우, 일반되고 안전한 의사결정을 보장하기 어렵다. 이를 보완하기 위해 매 추론 시점마다 고정적으로 전달되는 역할 및 규칙 정의 텍스트인 시스템 프롬프트를 제공한다. 본 프레임워크에서는 LLM에 명확한 역할을 부여하고, 도메인 지식 부족, 부정확한 사용자 입력에 대한 무비판적 수용, 다중 결함 발생 시 공정 변수 조정 전략 간 충돌, 비정형 텍스트 출력 등의 문제를 완화하기 위해 시스템 프롬프트를 여섯 가지 요소로 구성하였으며, 각 구성 요소는 다음과 같다(Fig. A1).

(1) 역할 정의(Role Definition)

LLM에게 특정 전문가 역할을 부여한다. 예를 들어, “너는 FDM 공정 결함 감지 및 공정 변수 제어 전문가이다”와 같이 명시함으로써, 의사결정이 해당 도메인 전문가의 관점에 기반하여 이루어지도록 유도한다.

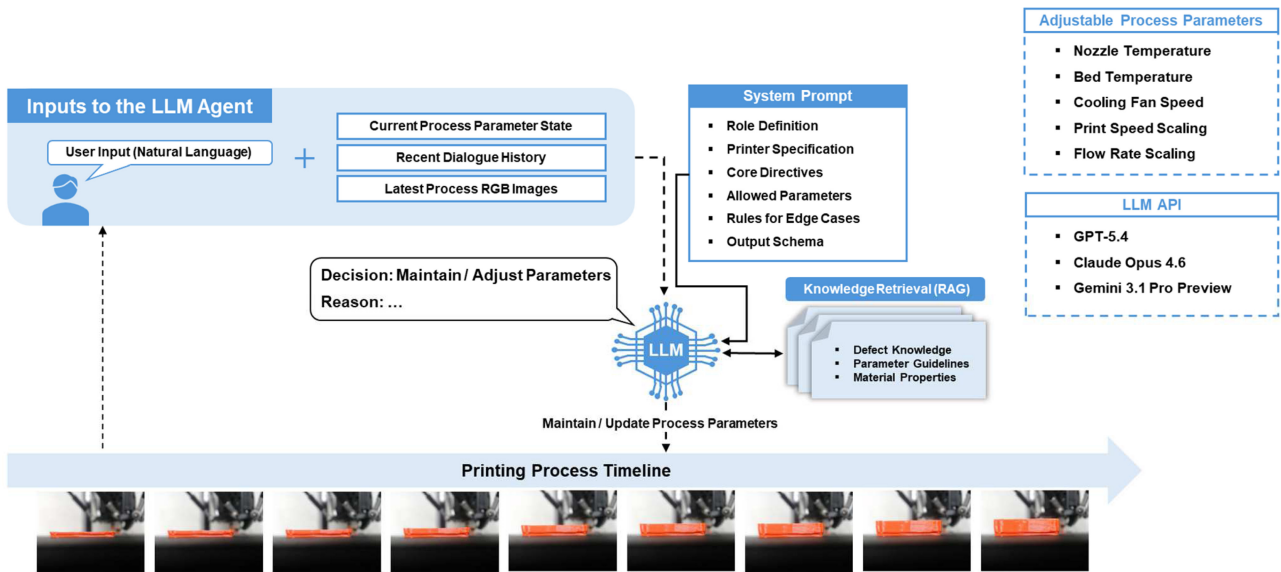


Fig. 2 LLM-based decision process design. During the FDM process, the LLM receives user prompts (user input, current process parameter state, latest process images, recent dialogue history) and a system prompt (role definition, printer specification, core directives, allowed parameters, rules for edge cases, output schema) to decide whether to maintain or adjust process parameters. If additional information is required, the LLM retrieves relevant data from an external knowledge base via RAG (retrieval-augmented generation)

(2) 프린터 사양(Printer Specification)

3D 프린터의 기종, 노즐 직경, 필라멘트 종류를 제공하여, 물리적 제약 조건을 반영한 판단이 이루어지도록 한다. 범용 LLM은 특정 하드웨어 특성에 최적화되어 학습된 모델이 아니므로, 이러한 정보를 사전에 제공함으로써 하드웨어 제약을 고려한 의사결정을 유도한다.

(3) 의사결정 기준(Core Directives)

LLM이 의사결정을 수행할 때, 입력 이미지, 공정 상태, 사용자 입력을 종합적으로 고려하도록 한다. 특히, 사용자 입력은 불완전하거나 부정확할 수 있으므로, 단독 근거로 사용하지 않도록 제한하였다. 이는 불필요한 공정 변수 조정을 방지하기 위한 것으로, 증거가 불충분한 경우 공정 변수를 조정하지 않는 원칙을 포함하였다.

(4) 제어 가능 공정 변수(Allowed Parameters)

사전에 정의되지 않은 공정 변수 생성이나, 허용 범위를 벗어난 출력이 발생하지 않도록 제약 조건을 명시하였다. 또한 초기 레이어에서는 베드 접착력 확보를 위해 쿨링 팬 속도 조정을 금지하는 공정 특화 규칙을 포함하였다.

(5) 예외 상황 처리 규칙(Rules for Edge Cases)

복수의 결함이 상충하는 경우 구조적 결함을 우선 처리하도록 정의하였다. 공정 제어와 무관한 질의에 대해서는 공정 변수 조정 없이 응답만 생성하도록 하였으며, 공정 관련 지식이 부족한 경우 외부 지식 베이스를 참조하도록 지시하였다. 이러한

규칙은 실제 운용 환경에서 발생할 수 있는 다양한 상황에 안정적으로 대응하기 위해 필요하다.

(6) 출력 형식(Output Schema)

LLM 응답에서 필요한 정보를 안정적으로 추출하기 위해 JSON 구조화 출력 형식을 강제한다. 출력은 조정 여부, 조정된 공정 변수와 조정 값, 판단 근거로 구성된다. LLM은 기본적으로 자유 형식의 텍스트를 생성하므로, 출력 형식을 명시하지 않을 경우 공정 변수 값을 안정적으로 추출하기 어렵다. 공정 변수 조정 값은 실제 제어 명령으로 변환되어 3D 프린터에 적용되므로, 정확한 추출을 위해 정형화된 출력 형식이 필수적이다.

2.1.2 사용자 프롬프트 설계

사용자 프롬프트는 매 추론 시점마다 갱신되는 동적 정보를 포함하며, LLM이 현재 공정 상황을 정확히 파악할 수 있도록 설계하였다. 이를 위해 사용자 프롬프트를 다음 네 가지 요소로 구성하였다.

(1) 사용자 자연어 입력

작업자가 현장에서 관찰한 내용을 자연어로 입력한 정보이다. 이는 기존 입력만으로 파악하기 어려운 상황 정보를 보완하는 역할을 한다.

(2) 공정 변수 상태

현재 베드 온도, 프린팅 속도 등의 설정 값을 포함하며, 이를 통해 LLM이 현재 공정 조건을 파악하고 공정 변수 조정의 필요성과 방향을 판단할 수 있도록 한다.

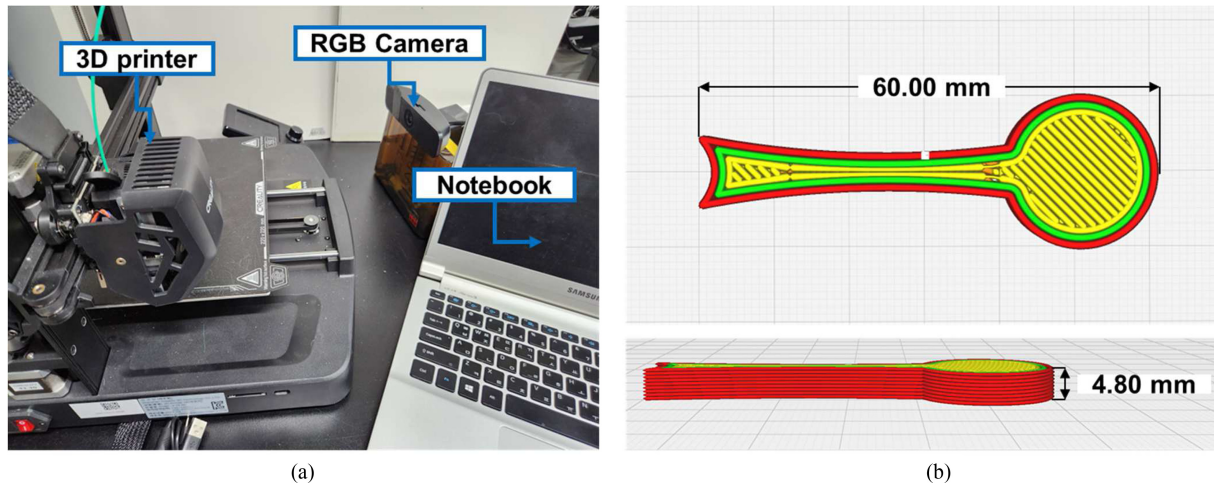


Fig. 3 Experiment setup and specimens' geometry. (a) Hardware setup, (b) Specimens' geometry

(3) 대화 이력

최근 5회의 대화 이력을 포함하여 LLM이 이전 조정 내역과 맥락을 유지한 상태에서 일관된 판단을 수행할 수 있도록 한다. 대화 이력이 없을 경우, 매 추론이 독립적으로 이루어져 불필요한 공정 변수 반복 조정이나 비효율적인 변동이 발생할 수 있다.

(4) RGB 카메라 입력

RGB 카메라로부터 획득된 최근 이미지 3장을 LLM 입력 형식에 맞추기 위해 Base64 형식으로 변환하여 제공한다. 서로 다른 시점에서 수집된 연속 이미지를 활용함으로써, LLM이 공정 상태의 시간적 변화를 파악하고, 단일 이미지 기반 판단의 한계를 보완할 수 있도록 한다.

2.1.3 RAG (Retrieval-augmented Generation) Module

범용 LLM은 FDM 공정에 특화된 세부 지식, 예를 들어 재료별 물성이나 공정 변수별 공정 영향 등에 대한 충분한 지식을 갖추지 못할 수 있다. 이를 보완하기 위해 외부 지식 베이스를 자율적으로 검색하는 RAG 모듈을 설계하였다. RAG 모듈의 전체 동작 흐름은 Fig. A2에 제시하였다.

RAG 모듈은 도메인 특화 지식의 필요 여부에 따라 검색 수행 여부를 결정한다. LLM 자체 지식만으로 판단이 가능한 경우에는 검색 없이 즉시 출력을 생성하고, 상세한 도메인 지식이 필요한 경우에는 외부 지식 베이스를 검색한 후 그 결과를 반영하여 최종 출력을 생성하도록 설계하였다. 이를 위해 시스템 프롬프트에 필요 시 외부 지식 베이스 참조하도록 지시하였다. 검색어는 LLM이 입력 정보를 바탕으로 생성한다. 예를 들어 LLM이 Stringing 결함을 인지하였음에도 적절한 공정 변수 수정 방안을 도출하지 못한 경우, “Stringing defect FDM PLA”와 같은 검색어를 생성하여 외부 지식 베이스에서 관련 문서를 검색한다. 검색된 문서는 기존 시스템 프롬프트 및 사용자 프롬프트와 함께 LLM의 입력 컨텍스트에 포함되며, 이를 기반으로 재추론을 수행하여 최종 출력을 생성한다.

관련 문서 검색은 텍스트 간 의미적 유사성을 정량화하는 임베딩 기반 방식을 사용한다. 지식 베이스의 각 문서와 검색어는 OpenAI의 text-embedding-3-small [8] 모델을 통해 동일한 벡터 공간으로 변환되며, 변환된 문서 벡터는 ChromaDB [9] 벡터 데이터베이스에 저장된다. 검색 시에는 검색어 벡터와 문서 벡터 간의 코사인 유사도를 계산하여 관련성이 높은 상위 3개의 관련 문서를 반환한다. 코사인 유사도는 두 벡터 간의 방향 유사성을 측정하는 지표로 식(1)과 같이 정의된다.

$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (1)$$

여기서 A와 B는 각각 입력 텍스트와 문서의 임베딩 벡터를 의미하며, 값이 1에 가까울수록 두 텍스트의 의미적 유사성이 높음을 의미한다.

외부 지식 베이스는 LLM이 결함 진단과 공정 변수 조정 전략 도출에 필요한 도메인 지식을 참조할 수 있도록 결함 지식, 재료 지식, 공정 변수 지식의 세 범주로 구성하였다. 결함 지식 문서는 Layer Shifting, Over-extrusion, Poor Bed Adhesion, Stringing, Under-extrusion, Warping의 여섯 가지 결함에 대해 시각적 패턴, 물리적 발생 원인, 권장 공정 변수 조정 전략을 포함한다(Fig. A3(a)). 재료 지식 문서는 PLA의 물성 및 열적 거동을 포함하여 재료 특성을 반영한 판단을 지원한다(Fig. A3(b)). 공정 변수 지식 문서는 Fan Speed, Flow Rate, Print Speed, Temperature의 정의와 물리적 영향을 포함하여 공정 변수 조정에 대한 이해를 지원한다(Fig. A3(c)).

2.2 Control Engine Module

Control Engine Module은 LLM Inference, G-code Sender, RGB Capture, UI Input Listener, State Broadcast의 다섯 가지 기능으로 구성된다. 각 기능은 공정 중 개입 주기와 처리 시간이 상이하며, 동시에 수행될 필요가 있으므로, Multi-thread [10] 방식을 사용하여 병렬적으로 동작하도록 설계하였다.

모듈의 동작 흐름은 다음과 같다. G-code Sender Thread는 슬라이싱된 G-code를 한 줄씩 시리얼 통신을 통해 지속적으로 프린터에 전송하며, 동시에 Camera Capture Thread는 1.5초 간격으로 프레임을 캡처하여 공정 상태를 시각적으로 수집한다. 사용자 입력은 UI Input Listener Thread를 통해 웹 서버로부터 수신되어 LLM 입력 대기열에 저장되며, LLM Inference Thread는 대기열의 데이터를 순차적으로 처리하며 공정 변수 조정 여부를 판단한다. 이때 LLM이 공정 변수 조정을 결정하면, 해당 명령은 G-code 스트림 중간에 삽입되어 프린터로 전달된다. 이 과정에서 G-code Sender Thread와 LLM Inference Thread가 동일한 시리얼 포트에 동시에 접근할 수 있다. 따라서 명령 충돌 및 누락을 방지하기 위해, 한 Thread가 시리얼 포트를 점유하는 동안 다른 Thread의 접근은 Lock 메커니즘을 통해 차단되도록 구현하였다[10]. 조정된 공정 변수는 주기적으로 웹 서버에 공정 변수 상태를 전송하는 State Broadcast Thread를 통해 UI 상에 실시간으로 반영된다.

3. 실험

3.1 실험 구성

본 연구에서는 Creality Ender-3 v3 SE 3D 프린터를 실험 플랫폼으로 사용하였으며, 공정 상태 모니터링을 위해 1080p 해상도의 Logitech C925e RGB 카메라를 노트북에 연결하여 운용하였다. 직경 1.75 mm PLA 필라멘트를 사용하였으며, 결함이 보다 명확하게 나타나도록 표준 사양인 직경 0.4 mm 노즐 대신 1.0 mm 노즐을 적용하였다. 출력 전 슬라이싱은 Ultimaker Cura [11]를 사용하여 수행하였다. 출력물은 warping 검증에 적합한 형상으로 설계하였으며, 실험 환경 및 출력물 형상은 Fig. 3에 제시하였다. 공정 중 RGB 카메라로 캡처된 프레임은 UI의 실시간 영상 스트리밍에도 함께 활용된다. 웹 서버는 Python 기반 웹 프레임워크인 Flask [12]를 사용하여 구현하였으며, UI는 HTML, CSS, JavaScript로 구성하였다. 웹 서버와 UI 간 통신은 Socket.IO [13]를 통해 LLM 응답 및 공정 변수 상태를 실시간으로 송수신하고, HTTP MJPEG 스트리밍을 통해 카메라 영상을 전달한다. 이를 통해 작업자는 공정 상태 모니터링과 LLM과의 상호작용을 하나의 화면에서 수행할 수 있다.

LLM은 상용 API로 널리 활용되는 OpenAI, Anthropic, Google의 최상위 모델인 GPT-5.4 [14], Claude Opus 4.6 [15], Gemini 3.1 Pro Preview [16]를 사용하였다. 세 API는 LangChain [17] 프레임워크를 기반으로 통합하였다. LangChain은 서로 다른 LLM API 호출 방식을 단일화하여 동일한 코드 구조로 운용할 수 있도록 지원하며, 이를 통해 세 모델에 동일한 입력 형식과 호출 구조를 적용하여 실험 조건을 통일하였다. 또한 모든 모델의 Temperature 변수를 0으로 설정하였다. Temperature는 LLM의 출력의 다양성을 조절하는 변수로, 값이 0일 경우 출력의 변동성을 최소화하여 일관된 의사결정이

Table 1 Initial values and operating ranges of controllable process parameters

Parameter	Initial value	Operating range
Print speed scale	1.0	0.7–1.3
Flow rate scale	1.0	0.8–1.2
Nozzle temperature	200	180–230
Bed temperature	60	50–80
Cooling fan speed	0 (layer 1) 200 (layer≥2)	0–255

이루어지도록 한다.

제어 대상인 공정 변수는 노즐 온도, 베드 온도, 쿨링 팬 속도, 프린팅 속도, 압출량으로 구성된다. 각 공정 변수는 M-code 명령(M104, M140, M106, M220, M221)을 시리얼 통신을 통해 출력 중간에 삽입함으로써 실시간으로 조정할 수 있도록 구현하였다. 프린팅 속도와 압출량은 슬라이서에서 설정된 기준 값에 대한 스케일 계수로 조정되며, 온도와 팬 속도는 현재 설정 값에 대한 변화량으로 조정된다. Table 1은 각 공정 변수의 초기 값과 허용 범위를 제시한다. 모든 공정 변수는 사전에 정의된 범위 내에서만 동작하도록 제한하였으며, 이를 초과하는 값은 안전을 위해 허용 범위의 경계 값으로 제한하여 적용하였다.

3.2 실험 시나리오

프레임워크의 성능 평가를 위해 두 유형의 Case를 설계하였으며, 전체 시나리오 구성은 Table A1에 제시하였다.

Case 1은 Incorrect Process Parameter Settings로, 출력 첫 번째 레이어 시작 시점에 비정상 공정 변수를 주입하여 결함을 유도한 경우이다. Fast Print Speed 시나리오에서는 프린팅 속도를 정상 대비 2배로 설정하여 형상 왜곡을 유도하였다. Low Flow Rate 시나리오에서는 압출량을 정상 값의 0.7배로 설정하여 Under-extrusion 조건을 구성하였다. Low Bed Temperature 시나리오에서는 베드 온도를 30°C로 설정하여 Warping이 발생하기 쉬운 조건을 형성하였다.

Case 2는 Environmental Disturbances로, 공정 중 물리적 환경 교란을 가한 경우이다. Wind Disturbance 시나리오는 실내 환기 상황에서 외부 기류가 프린터 주변으로 유입되는 상황을 가정하였으며, 출력 중 송풍 장치를 이용해 출력 필라멘트의 급속 냉각을 유도하였다. 이러한 조건은 Warping 결함을 유발할 수 있다. Vibration Disturbance 시나리오는 공장 환경에서 주변 장비의 작동에 의해 3D 프린터에 진동이 전달되는 상황을 가정하였으며, 외부 진동원을 3D 프린터 프레임에 접촉시켜 이를 재현하였다. 해당 조건은 Layer Shifting과 같은 문제를 유발할 수 있다.

각 시나리오에서 공정 중 작업자 관찰 정보를 채팅창에 자연어로 입력하여 LLM의 응답 품질 및 출력물의 품질 개선 여부를 검증한다. 작업자 입력은 특정 결함을 직접적으로 명시하지 않고 작업자의 관찰 정보만을 포함하는 형태로 사전에 정의하였다. 이는 실제 공정 환경에서 작업자가 제공 가능한 정보 수준을 반영하는

동시에, 명시적인 결함 정보 없이도 LLM이 상황을 추론하고 적절한 공정 변수를 결정할 수 있는지를 확인하기 위함이다. 시나리오별 상세 사용자 입력은 Table A2에 제시하였다.

3.3 평가 지표

본 연구의 정량적 성능 평가는 LLM 응답 품질과 출력물 품질의 두 측면에서 이루어진다.

LLM 응답 품질은 Table A3에 정의된 네 가지 지표로 평가한다. Detection Rate는 LLM이 이상 상황을 인식하고 조정 명령을 생성한 비율을 의미한다. Target Adjustment Rate는 각 시나리오에서 필수적으로 조정되어야 하는 핵심 공정 변수가 실제로 조정된 비율을 나타낸다. Direction Accuracy는 핵심 공정 변수가 조정된 경우, 해당 변수가 올바른 방향(증가 또는 감소)으로 조정된 비율로 정의된다. 시나리오별 핵심 공정 변수와 올바른 조정 방향에 대한 기준은 Table A4에 제시하였다.

Response Latency는 사용자 입력 시점부터 LLM 응답이 반환될 때까지 소요된 시간(초)으로 정의된다. 다만 해당 지표는 네트워크 상태, API 서버 부하, 시스템 자원 등 외부 요인의 영향을 크게 받기 때문에 정밀한 성능 지표로 활용하는 데 한계가 있다. 따라서 본 연구에서는 이를 모델 간 성능 비교가 아닌, 실제 FDM 공정에서 LLM API의 실시간 운용 가능성을 판단하기 위한 지표로 활용하였다.

출력물 품질은 Table A5에 제시된 기준에 따라 실제 출력물을 기반으로 물리적 측정 및 이미지 분석을 통해 평가하였으며, 프레임워크 적용 전후를 비교하였다. 구체적으로 Low Flow Rate 시나리오에서는 Surface Void를 평가지표로 사용하였다. Surface Void는 출력물을 RGB 카메라로 촬영한 후 배경을 제거하고, 물체 영역 내에서 검정 픽셀이 차지하는 비율로 산출하였다. Low Bed Temperature 및 Wind Disturbance 시나리오에서는 Warping Height를 평가지표로 사용하였다. 출력물을 클램프로 고정한 상태에서 들뜬 모서리와 바닥 사이의 거리를 디지털 캘리퍼스로 측정하였다. Fast Print Speed 시나리오에서는 Height Deviation을 평가지표로 정의하였다. 이는 디지털 캘리퍼스로 측정한 출력물 높이와 슬라이서 프로그램에서 설정된 높이 4.8 mm 간의 차이로 계산하였다. Vibration Disturbance 시나리오에서는 Layer shift distance를 평가지표로 사용하였다. 해당 지표는 출력물을 자와 함께 촬영한 후, 수평 방향 기준으로 최대 돌출 지점과 최소 돌출 지점 간의 거리를 측정하여 산출하였다.

3.4 실험 결과 및 분석

본 절에서는 제안된 프레임워크의 성능을 LLM 응답 품질과 출력물 품질 측면에서 정량적으로 분석한다. 실험은 동일한 형상의 출력물을 대상으로 3종의 LLM API와 Table A1에 제시된 5가지 시나리오에 대해 각 3회씩 반복 수행하여 총 45회 출력을 진행하였다.

프레임워크 적용 결과와의 비교를 위해, 동일한 시나리오 조건에서 LLM 기반 의사결정과 작업자 관찰 정보 개입을 모두 제외한

베이스라인 조건을 구성하고, 각 시나리오별 1회씩 추가 실험을 수행하였다. 또한, 작업자 관찰 정보가 LLM의 의사결정 결과에 미치는 영향을 분석하기 위해, 관찰 정보 제공 여부에 따른 응답 비교 실험을 각 시나리오별로 1회씩 수행하였다.

3.4.1 LLM 응답 품질 평가 결과 및 분석

각 시나리오에서 LLM의 결함 감지 여부, 사전에 정의된 핵심 공정 변수에 대한 조정 수행 여부, 그리고 해당 변수의 조정 방향(증가 또는 감소)이 기대된 방향과 일치하는지를 기준으로 응답 품질을 정량적으로 평가하였다. 평가 결과는 Table A6에 제시되어 있다. GPT-5.4와 Claude Opus 4.6은 모든 시나리오에서 Detection Rate, Target Adjustment Rate, Direction Accuracy 모두 100% (15/15)를 기록하였다. 반면, Gemini 3.1 Pro Preview는 Wind Disturbance 시나리오에서 33.3% (1/3)의 성능을 보여 전체적으로 86.7% (13/15)의 성능을 보였다. 제안된 프레임워크는 대부분의 시나리오에서 높은 응답 품질을 보였으나, 일부 시나리오에서는 모델에 따라 성능 차이가 나타났다.

모델 간 성능 차이의 원인을 파악하기 위해, 성능 차이가 나타난 Wind Disturbance 시나리오의 추론 로그를 분석하였다. Fig. A4는 동일한 시나리오 조건에서 GPT-5.4와 Gemini 3.1 Pro Preview 모델에 입력된 사용자 입력과 이미지에 대한 응답 사례를 보여준다. GPT-5.4는 선풍 불균일과 접촉력 저하를 감지하고 공정 변수 조정을 수행한 반면, Gemini 3.1 Pro Preview는 결함을 감지하지 못하고 공정 변수를 조정하지 않았다.

이는 동일한 시나리오 조건에서도 모델에 따라 이미지상 결함 감지 결과가 상이할 수 있음을 보여준다. 또한, GPT-5.4의 결함 감지가 모델 자체의 추론에 의한 것인지, 혹은 작업자 관찰 정보의 영향에 따른 것인지 구분하여 분석할 필요가 있다. 관찰 정보 제공 여부에 따른 GPT-5.4의 시나리오별 응답 결과와 분석은 3.4.3절에 제시한다.

Response Latency 결과는 GPT-5.4가 평균 3.58 ± 0.47 초로 세 모델 중 가장 낮은 Response Latency를 보였으며, Claude Opus 4.6과 Gemini 3.1 Pro Preview는 각각 평균 7.06 ± 0.78 초와 9.42 ± 1.02 초로 상대적으로 높은 Response Latency를 나타냈다. GPT-5.4의 경우 Response Latency가 5초 이내로, 한 레이저 적층 완료 이전에 파라미터 변경이 가능한 시간이며, 실제 공정 환경에 적용 가능성을 시사한다.

응답 품질 측면에서 GPT-5.4와 Claude Opus 4.6은 동등하게 우수한 성능을 나타내었으나, GPT-5.4는 더 낮은 API 비용(Table A7)과 Response Latency를 보였다. 이러한 점을 종합할 때, 본 연구에서는 GPT-5.4를 활용한 프레임워크 구성이 가장 합리적인 선택임을 확인하였다.

3.4.2 출력물 품질 평가 결과 및 분석

본 절에서는 3.4.1절의 분석 결과를 바탕으로 GPT-5.4에 기반한 공정 변수 조정이 실제 출력물 품질에 미친 영향을 정량적으로 평가하고 그 결과를 분석한다. 출력물 품질 개선 결과는

Table 2 Print quality comparison before and after system application

Case	Scenario	Metric	Without system	With system		Improvement rate [%]
				Trial	Mean $\pm$ SD	
1	Fast print speed	Height deviation [mm]	1.93	0.51	0.61 $\pm$ 0.08	69.95
				0.65		
				0.66		
	Low flow rate	Surface void [%]	16.50	7.66	7.58 $\pm$ 1.26	54.06
				8.79		
				6.28		
2	Low bed temperature	Warping height [mm]	1.29	0.37	0.33 $\pm$ 0.05	74.42
				0.28		
				0.33		
	Wind disturbance	Warping height [mm]	1.83	0.75	1.03 $\pm$ 0.40	43.72
				0.84		
				1.49		
	Vibration disturbance	Layer shift distance [mm]	1.0	0.70	0.65 $\pm$ 0.05	35.00
				0.60		
				0.70		

Table 2에 제시되어 있다. 개선율은 프레임워크 적용 전 대비 적용 후 측정값의 변화 비율을 백분율로 나타낸 것이다.

Fast Print Speed 시나리오에서는 LLM에게 출력물의 적층 상태가 불안정해 보인다는 관찰 정보를 제공하였다. LLM은 해당 관찰 정보를 포함한 입력 정보를 종합적으로 분석하여 Print Speed 감소를 제안하였다. 그 결과 Height Deviation 이 1.93에서 평균 0.61 mm로 감소하여 69.95%의 개선율을 보였다. Low Flow Rate 시나리오에서는 필라멘트 압출량이 일정하지 않아 보인다는 작업자 관찰 정보를 제공하였다. LLM은 입력 정보를 종합적으로 분석하여 Flow Rate 증가를 제안하였다. 그 결과 Surface Void가 16.50%에서 평균 7.58%로 감소하여 54.06% 개선되었으며, 이에 대한 이미지 분석 결과는 Fig. 4에 제시하였다. Low Bed Temperature 시나리오에서는 베드가 충분히 따뜻하지 않은 것 같다는 작업자 관찰 정보를 제공하였다. LLM은 Bed Temperature 증가를 제안하였으며, 그 결과 Warping Height가 1.29에서 평균 0.33 mm로 감소하여 74.42%의 가장 높은 개선율을 기록하였다. Wind Disturbance 시나리오에서는 환기로 인해 강한 외부 바람이 유입되고 있다는 작업자 관찰 정보를 제공하였다. LLM은 이를 외부 냉각에 의한 출력물 접착력 저하 가능성과 연관 지어 해석하고 Bed Temperature 증가를 제안하였다. 그 결과 Warping Height가 1.83에서 평균 1.03 mm로 감소하여 43.72% 개선되었다. Vibration Disturbance 시나리오에서는 주변 장비로 인해 프린터기에 진동이 발생하고 있다는 작업자 관찰 정보를 제공하였다. LLM은 이를 진동에 의한 Layer Shift 발생 가능성과 연관 지어 해석하고 Print Speed 감소를 제안하였다. 그 결과 Layer Shift Distance가 1.0에서 평균 0.65 mm로 감소하여 35.00%의 가장 낮은 개선율을 보였다. 전체 시나리오에서 품질이 향상되었으며, 평균 약 55%의 개선율을 나타내었다.

분석 결과 Case 1 (Incorrect Process Parameter Settings)의

시나리오가 Case 2 (Environmental Disturbances)보다 높은 개선율을 나타냈다. 이는 LLM이 Case 2에서는 이미지와 작업자 관찰 정보에 의존하여 판단하는 반면, Case 1에서는 이상 상태가 공정 변수에도 반영되어 있어 원인 추정이 상대적으로 용이하기 때문으로 해석된다.

Low Bed Temperature 시나리오에서 가장 높은 개선율이 나타난 이유는 Warping이 명확하게 드러나는 출력물 구조와 관련된 것으로 보인다. 해당 조건에서는 결함의 이미지에 뚜렷하게 나타나므로, LLM이 이를 기반으로 공정 변수 조정을 수행하기 용이했을 가능성이 있다.

반면, Vibration Disturbance 시나리오에서 상대적으로 낮은 개선율이 나타난 원인은 이미지의 수집 조건에 기인한 것으로 판단된다. 본 실험에서는 단일 RGB 카메라로 1.5초 간격으로 촬영된 3장의 이미지만을 입력으로 사용하였다. 진동은 기준점을 중심으로 주기적으로 반복되는 운동 현상으로, 이러한 특성을 포착하기 위해서는 충분한 샘플 수가 요구된다. 그러나 주어진 3장의 이미지만으로는 진동여부를 명확히 판별하지 못해, 해당 시나리오에서의 성능 향상이 제한된 것으로 분석된다.

3.4.3 작업자 관찰 정보가 LLM 의사결정에 미치는 영향

본 절에서는 작업자 관찰 정보 제공 여부가 LLM의 의사결정에 미치는 영향을 분석한다. Fig. A9는 각 시나리오에서 동일한 입력 이미지에 대해 관찰 정보 제공 여부에 따른 GPT-5.4 응답 결과이다. Fig. A9(a) Fast Print Speed 시나리오에서는 관찰 정보 제공 여부에 관계없이 이미지상에서 결함을 감지 못하였다. 그러나, 관찰 정보가 제공된 경우에는 높은 Print Speed Scale 결함을 감지하고 공정 변수 조정을 수행한다. Fig. A9(b) Low Flow Rate 시나리오에서는 관찰 정보 제공 여부에 관계없이

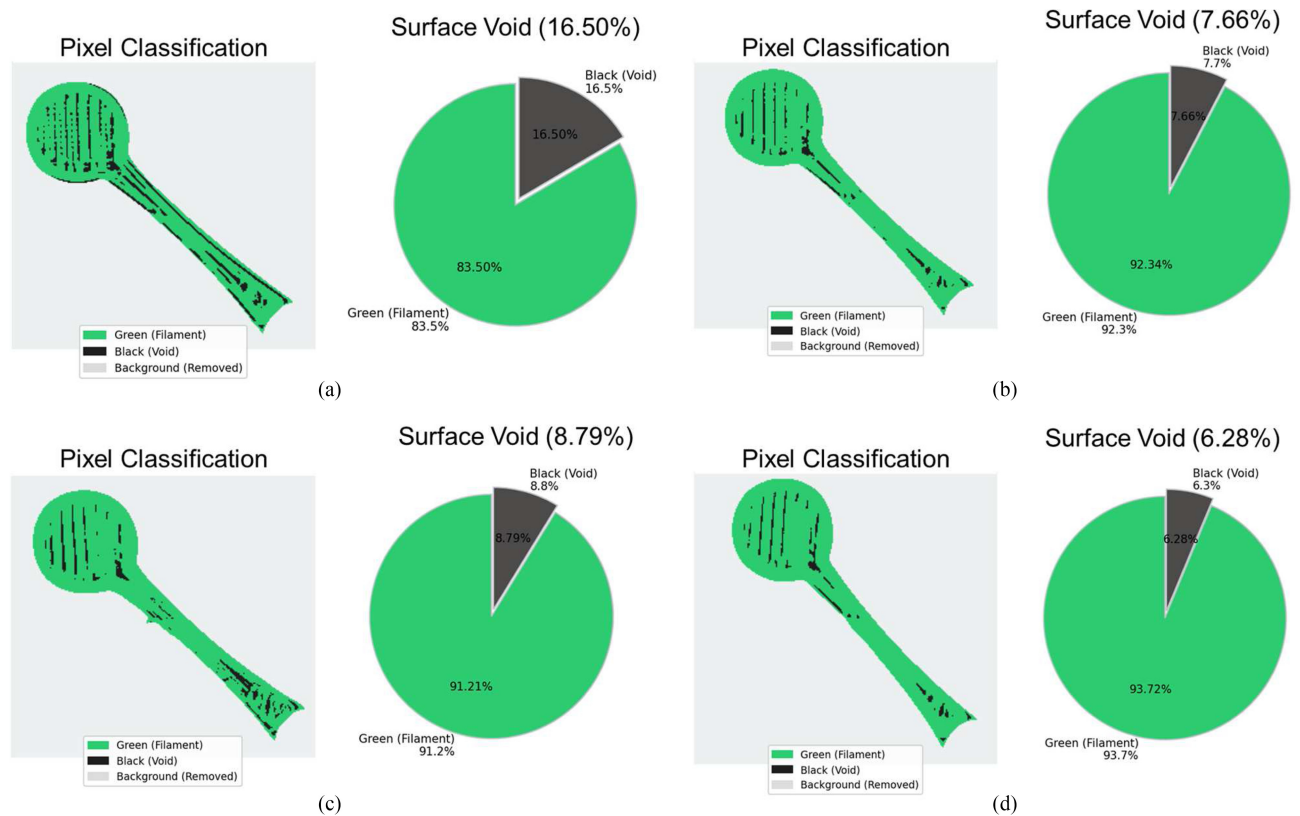


Fig. 4 Surface void of 3D-printed parts in the Low Flow Rate scenario. (a) Without system (16.50%) and (b)-(d) With system (7.66, 8.79, and 6.28%)

under-extrusion과 낮은 Flow Rate 결함을 감지했다. 이에 따라 Flow Rate 값을 증가시키는 동일한 제어 전략을 도출했다. Fig. A9(c) Low Bed Temperature 시나리오에서는 관찰 정보가 제공되지 않은 경우 이미지 상에서 결함을 감지 못하였다. 반면, “베드가 미지근하다”는 관찰 정보가 제공된 경우, 낮은 Bed Temperature 결함을 감지하고 베드 온도를 높이는 조정을 진행하였다. Fig. A9(d) Wind Disturbance 시나리오에서는 관찰 정보가 없는 경우, 결함을 감지하지 못하였다. 반면, “환기로 인한 기름이 유입되고 있다”라는 관찰 정보가 제공된 경우, 출력물의 부착 불안정 및 비드 형상 이상을 감지하였다. 결함 원인으로, 외풍 또는 과도한 냉각을 예상하고 Fan Speed 감속을 수행했다. Fig. A9(e) Vibration Disturbance 시나리오에서는 관찰 정보가 없는 경우 출력물의 형상 이상을 감지하였다. 원인으로서는 Over-extrusion 또는 냉각 부족을 예상하였다. “프린터가 주변 기계로 인해 진동하고 있다”는 관찰 정보가 제공된 경우, 형상 이상을 감지하고 원인으로 움직임 교란을 예상하였다.

이러한 결과를 종합하면 다음과 같이 분석할 수 있다. 첫째, Fig. A9(b) 사례의 입력 이미지를 살펴보면 Surface Void 결함이 뚜렷하게 나타나며, 관찰 정보 제공 없이도 Flow Rate 공정 변수를 조정한다. 이는 이미지상 결함이 명확하게 나타난 경우, LLM이 추가적인 정보 없이도 결함을 인지하고 적절한 공정 변수 조정 전략을 도출할 수 있음을 시사한다. 둘째, Figs. A9(a)와

A9(c) 사례에서는 관찰 정보가 제공된 경우에만 공정 변수 상태 이상이 감지되었다. 이는 관찰 정보가 LLM의 추론 과정에서 공정 변수 상태를 확인하도록 유도하는 보조 정보로 활용되었을 가능성을 시사한다. 셋째, Fig. A9(d) 사례에서는 이미지상 시편에 미세한 들림이 나타났음에도 불구하고, 관찰 정보가 제공되지 않은 경우에는 LLM이 결함을 감지하지 못하였다. 반면, 관찰 정보가 제공된 경우에는 접착력 저하를 감지하고 공정 변수 조정을 수행하였다. 이 사례는 작업자 관찰 정보의 제공 여부가 결함 감지 결과와 연관될 가능성을 시사한다. 넷째, Fig. A9(e) 사례에서는 진동에 대한 관찰 정보가 제공된 경우 해당 정보와 관련된 결함이 감지된 반면, 관찰 정보가 제공되지 않은 경우에는 진동과의 연관성이 낮은 결함이 감지되었다. 이 사례는 작업자 관찰 정보 제공 여부가 결함 해석 결과와 연관될 가능성을 시사한다.

결과적으로, 이미지상 결함이 명확한 사례에서는 작업자 관찰 정보 없이도 LLM이 결함을 감지하고 공정 변수 조정을 수행하였다. 반면, 이미지상 결함이 뚜렷하지 않은 사례에서는 작업자 관찰 정보 제공 여부에 따라 LLM의 결함 감지와 해석 결과에 차이가 나타났다. 특히 관찰 정보가 제공된 경우에는 공정 변수 상태를 확인하여 결함 원인에 대한 설명이 보다 구체적으로 이루어지는 응답이 관찰되었다. 이러한 결과는 작업자 관찰 정보 제공 여부가 결함 감지 및 해석 과정과 연관될 가능성을 시사한다. 또한 본 사례들에서는 작업자 관찰 정보가 제공되지

않은 경우 결함을 감지하지 못한 사례가 일부 존재한 반면, 관찰 정보가 제공된 경우 모든 사례에서 결함을 감지하고 공정 변수를 적절히 조정하였다. 이는 작업자 관찰 정보가 LLM의 결함 감지 성능 향상에 긍정적으로 기여할 가능성을 보여준다. 다만, 본 실험은 제한된 시나리오를 대상으로 수행되었으며, 관찰 정보 제공 여부 외의 잠재적 교란 변수를 충분히 통제하지 못하였다. 따라서 작업자 관찰 정보가 응답 차이에 미친 영향을 직접적으로 입증하기에는 한계가 있다. 향후에는 반복 실험과 체계적인 Ablation Study를 통해 작업자 관찰 정보의 실제 기여도를 정량적으로 검증할 필요가 있다.

4. 결론

본 연구에서는 설치된 센서로 결함 감지가 어려운 이상 상황에서, 작업자와 LLM의 협력을 통해 FDM 공정 중 공정 변수를 실시간으로 조정하는 Human-in-the-loop 기반 프레임워크를 제안하였다. 제안된 프레임워크의 성능과 유효성은 응답 품질 및 출력물 품질에 대한 정량 평가와 작업자 관찰 정보 영향을 분석한 정성 평가를 통해 확인하였다.

Case 1 (Incorrect Process Parameter Settings)과 Case 2 (Environmental Disturbances)의 시나리오 실험 결과, 제안된 프레임워크는 정의된 응답 품질 지표에서 80% 이상의 성능을 보였으며, 출력물 품질 또한 전체 시나리오에서 평균 약 55%의 개선율을 나타내 유의미한 성능 향상을 확인하였다. 작업자 관찰 정보 유효성 실험에서는 이미지상 결함이 명확한 사례의 경우 관찰 정보 제공 없이도 결함이 감지되었다. 반면, 이미지상 결함이 뚜렷하지 않은 사례에서는 관찰 정보 제공 여부에 따라 공정 변수 상태를 확인하고 결함 감지 및 해석 결과에 차이가 나타났다. 특히 관찰 정보가 제공된 경우에는 공정 변수 상태를 확인하여 결함 원인에 대한 해석이 보다 구체적으로 이루어지는 응답이 관찰되었다. 이러한 결과는 작업자 관찰 정보가 LLM의 결함 감지 및 해석 과정에 긍정적으로 기여할 가능성을 보여준다.

본 연구는 FDM 공정의 완전 자율 제어로의 이행 과정에서 발생할 수 있는 한계를 보완하기 위한 접근 방향을 제시한다는 점에서 의미를 갖는다. 또한, 보다 발전된 제조 환경 구축을 위한 참고 사례로 활용될 수 있기를 기대한다.

다만 본 연구는 제한된 수의 시나리오를 기반으로 수행되었으며, 작업자 관찰 정보 또한 실험 간 일관성을 확보하기 위해 사전에 정의된 문장을 사용하였다. 따라서 실제 제조 현장에서 나타날 수 있는 보다 다양한 표현 방식과 모호한 설명에 대해서는 충분히 검증되지 않았다는 한계가 있다. 향후에는 실험 환경의 습도 및 온도 변화 등 다양한 조건을 반영한 확장 실험과 함께 다양한 형태의 작업자 입력에 대한 강건성을 평가하여 프레임워크의 일반화 가능성을 검증할 필요가 있다.

ACKNOWLEDGEMENT

This work was supported by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Climate, Energy & Environment (MCEE) of the Republic of Korea (No. RS-2023-00240918).

APPENDIX

A1. 시스템 프롬프트

Fig. A1은 LLM의 일관된 의사결정을 유도하기 위해 설계된 시스템 프롬프트의 구성 요소를 나타낸다. 각 구성 요소는 역할 정의, 프린터 사양, 핵심 지침, 허용 파라미터, 예외 상황 처리 규칙, 출력 형식으로 이루어지며, LLM이 공정 제어 맥락을 명확히 이해하고 안정적인 응답을 생성하도록 한다.

A2. RAG 모듈 흐름도

Fig. A2는 제안된 RAG (Retrieval-augmented Generation) 모듈의 전체 흐름을 나타낸다. 상단은 추론 과정에서 외부 지식을 활용하는 LLM Inference 흐름을, 하단은 지식베이스 구축 과정을 나타내며, 이를 통해 공정 관련 지식을 동적으로 참조할 수 있도록 한다.

A3. 외부 지식베이스 문서 예시

Fig. A3은 RAG 모듈에서 활용되는 외부 지식베이스 문서의 예시를 제시한다. 결함, 재료, 공정 변수에 대한 문서가 포함되며, LLM이 다양한 공정 상황에 대해 보다 정확한 판단을 내릴 수 있도록 지원한다.

A4. Framework 평가를 위한 실험 시나리오

Table A1은 프레임워크 성능 평가를 위해 설계된 실험 시나리오를 정리한 것이다. Case 1 (Incorrect Process Parameter Settings)과 Case 2 (Environmental Disturbances)로 구분되며, 각 시나리오는 특정 조건을 인위적으로 부여하여 다양한 이상 상황을 재현한다.

A5. 시나리오별 사용자 입력 목록

Table A2는 각 실험 시나리오에서 사용된 작업자 입력 문장을 정리한 것이다. 실험 간 일관성을 확보하기 위해 모든 입력은 사전에 정의된 문장을 사용하였으며, 이를 통해 LLM 응답 비교의 공정성을 확보하였다.

A6. LLM 응답 품질 평가 지표

Table A3은 LLM의 응답 품질을 정량적으로 평가하기 위한 지표를 정의한다. Detection Rate, Target Adjustment Rate,

Role Definition
You are an expert in FDM 3D printing process diagnosis and adaptive control.
Printer Specification
Printer specification: Printer Model: Creality Ender-3 v3 SE 3D Printer Nozzle Diameter: 1.0 mm Filament: PLA (1.75 mm)
Core Directives
Your decision MUST be based on: - Visible evidence in the images provided. - The provided current printer_state. - The human comment (Note: comments may be incomplete or incorrect. Use to guide closer image inspection, NOT as standalone evidence). * Exception: If the human explicitly requests optimization (e.g., faster printing, better quality), you may adjust parameters even if no defect is visible. * If evidence is insufficient, choose decision="continue" with empty param_updates.
Allowed Parameters
You are strictly limited to updating the following parameters within their ranges. IMPORTANT: - print_speed_scale and flow_scale are SCALE FACTORS applied to fixed BASE values defined before the printing process begins. - The final parameter value is computed as: final_value = base_value × scale. - nozzle_temp_delta, bed_temp_delta, and fan_speed_delta are RELATIVE adjustments added to the current machine settings. Parameters: - print_speed_scale (float): [0.70, 1.30] Scale factor applied to the base printing speed. - flow_scale (float): [0.80, 1.20] Scale factor applied to the base extrusion flow. - nozzle_temp_delta (int): [-5, 5] Adjustment added to the current nozzle temperature. - bed_temp_delta (int): [-5, 5] Adjustment added to the current bed temperature. - fan_speed_delta (int): [-40, 40] Adjustment added to the current fan speed. Do NOT adjust when current_layer ≤ 1; keep fan speed at 0 for bed adhesion. * SAFETY RULE: Do NOT invent new parameters under any circumstances.
Rules for Edge Cases
- Conflict Resolution: If multiple defects require conflicting parameter adjustments, prioritize the defect that affects structural integrity or is most urgent. - Question Handling: If the human comment is a general question or informational request (not requiring process control), treat it as a question. Set decision="continue", param_updates={}, and provide a concise answer in the "reason" field (max 200 chars). - Tool Usage: If you lack the specific 3D printing knowledge to diagnose the issue, use the 'search_3d_printing_knowledge' tool FIRST before generating your final response.
Output Schema
You must output STRICT JSON matching the schema below. You MUST wrap the JSON in a markdown code block (e.g., ```json\n...\n```). Do not include any other text inside the block. <pre> ```json { "decision": "continue" "adjust", "param_updates": { "param_name": <value> }, "reason": "Brief technical reasoning referencing images/comment/state/knowledge (max 200 chars)" } </pre>

Fig. A1 Components of the system prompt. From top to bottom, the sections represent (a) Role definition, (b) Printer specification, (c) Core directives, (d) Allowed parameters, (e) Rules for edge cases, (f) Output schema

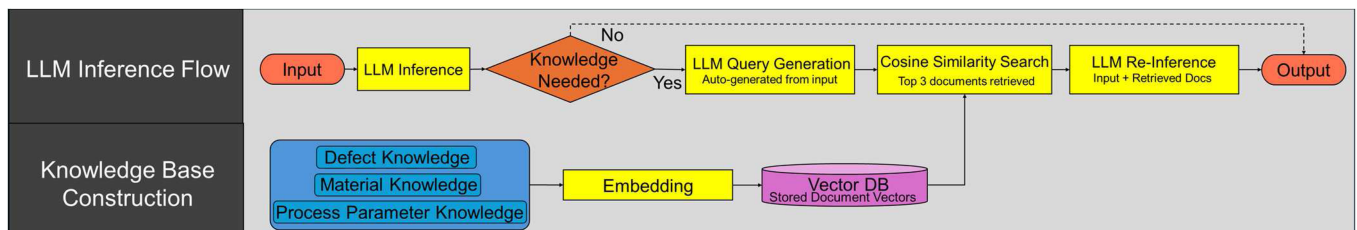


Fig. A2 Flowchart of the RAG module: LLM Inference flow (top) and knowledge base construction (bottom)

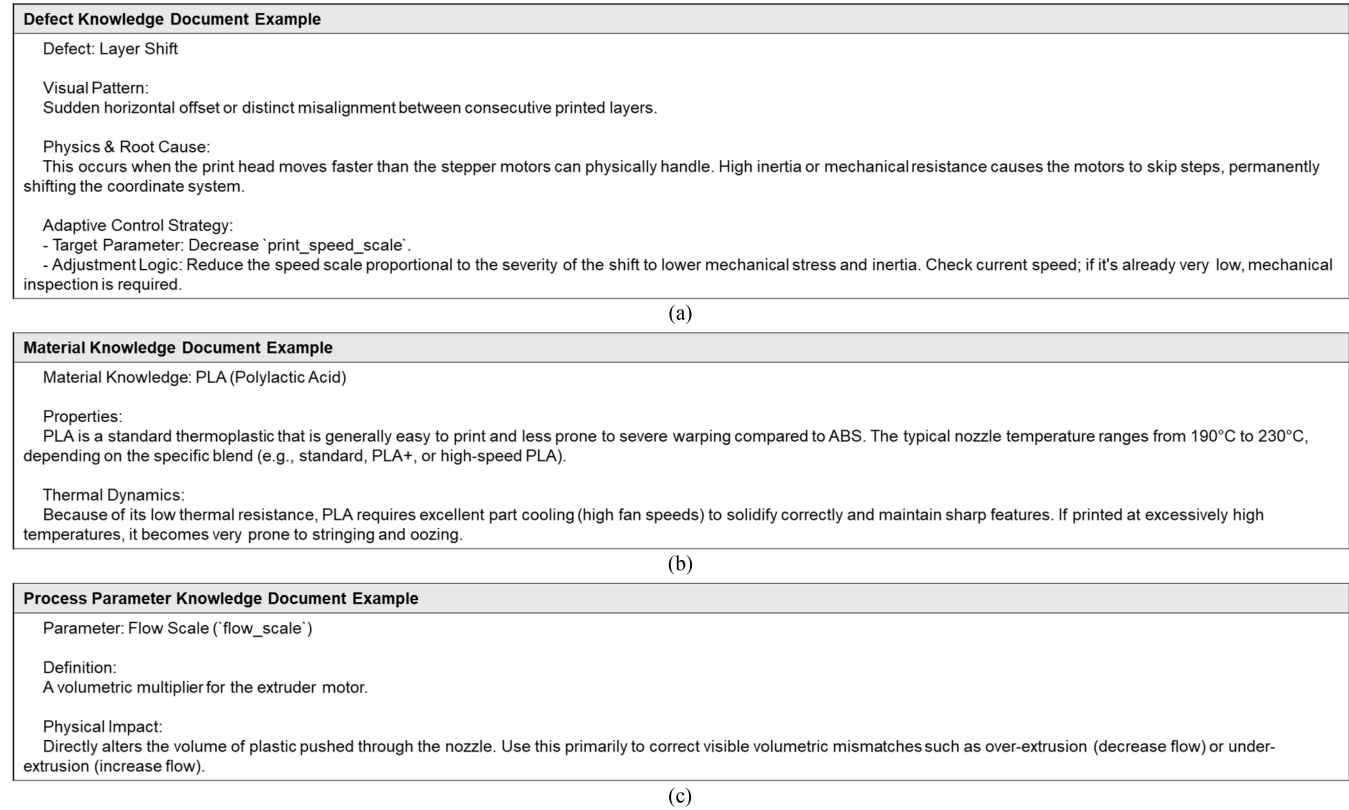


Table A3 LLM response quality evaluation metrics

Metric	Definition	Measurement
Detection rate	Proportion of cases where the LLM generates a parameter adjustment command	$\frac{\text{Number of cases with adjustment commands}}{\text{Total number of cases}}$
Target adjustment rate	Proportion of cases where the LLM adjusts the correct target parameter	$\frac{\text{Number of cases with correct target parameter adjustments}}{\text{Total number of cases}}$
Direction accuracy	Proportion of cases where the correct-direction target parameter adjustment is made	$\frac{\text{Number of correct direction adjustments}}{\text{Total number of cases}}$
Response latency	Time elapsed from user input to LLM response	Timestamp difference between user input and LLM response

Table A4 expected parameter adjustments per scenario for LLM response evaluation

Scenario	Target parameter	Adjustment direction
Fast print speed	Print speed scale	Decrease
Low flow rate	Flow scale	Increase
Low bed temperature	Bed temperature delta	Increase
Wind disturbance	Fan speed delta	Decrease
Vibration disturbance	Print speed scale	Decrease

Table A5 Print quality evaluation metrics

Metric	Applicable scenario	Measurement instrument
Surface void [%]	Low flow rate	Image analysis
Warping height [mm]	Low bed temperature, wind disturbance	Digital caliper
Height deviation [mm]	Fast print speed	Digital caliper
Layer shift distance [mm]	Vibration disturbance	Image analysis

A8. 출력물 품질 평가 지표

Table A5는 출력물 품질을 정량적으로 평가하기 위한 지표를 정리한 것이다. 각 시나리오에 따라 적절한 측정 지표를 적용하였으며, 이미지 분석 및 물리적 측정을 통해 품질 변화를 평가한다.

A9. LLM 응답 품질 평가 결과

Table A6은 각 LLM 모델의 응답 품질 평가 결과를 나타낸다. 대부분의 시나리오에서 높은 성능을 보였으나, Wind Disturbance 시나리오에서는 모델 간 성능 차이가 확인된다.

A10. LLM API 비용 비교

Table A7은 각 LLM API의 토큰당 비용을 비교한 것이다. 모델별 입력 및 출력 비용 차이를 통해 실제 시스템 적용 시 경제적 고려 요소를 제시한다.

A11. Wind Disturbance 시나리오에서 Gemini 3.1 Pro Preview와 GPT-5.4 응답 비교

Fig. A4는 Wind Disturbance 시나리오에서 두 LLM의 응답 차이를 비교한 것이다. 동일한 상황에서도 모델에 따라 이상

상황 인지 및 공정 변수 조정 여부에 차이가 발생함을 보여준다.

A12. 이상 출력 시나리오별 3D 프린팅 품질 측정 실험 결과

Figs. A5-A8은 각 시나리오에서 시스템 적용 여부에 따른 출력물 품질 변화를 나타낸다. 시스템 적용 시 전반적으로 결함이 감소하고 품질이 개선되는 경향이 확인된다.

A13. 작업자 관찰 정보 제공 여부에 따른 GPT-5.4 응답 결과 비교

Fig. A9는 작업자 관찰 정보 제공 여부에 따른 LLM 응답 차이를 비교한 것이다. 관찰 정보가 제공된 경우 공정 변수 상태를 보다 적극적으로 반영하는 경향이 나타나며, 의사결정 과정에 영향을 미칠 수 있음을 보여준다.

Table A6 LLM response quality evaluation results

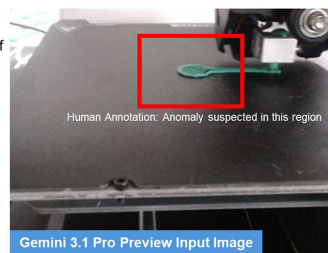
Model	Scenario	Detection rate	Target adjustment rate	Direction accuracy	Latency mean $\pm$ SD [s]
Claude Opus 4.6	Fast print speed	100% (3/3)	100% (3/3)	100% (3/3)	7.40 $\pm$ 2.36
	Low flow rate	100% (3/3)	100% (3/3)	100% (3/3)	4.72 $\pm$ 1.46
	Low bed temperature	100% (3/3)	100% (3/3)	100% (3/3)	4.14 $\pm$ 0.32
	Wind disturbance	100% (3/3)	100% (3/3)	100% (3/3)	10.41 $\pm$ 2.05
	Vibration disturbance	100% (3/3)	100% (3/3)	100% (3/3)	8.63 $\pm$ 1.35
	Total	100% (15/15)	100% (15/15)	100% (15/15)	7.06 $\pm$ 0.78
Gemini 3.1 Pro Preview	Fast print speed	100% (3/3)	100% (3/3)	100% (3/3)	10.35 $\pm$ 0.58
	Low flow rate	100% (3/3)	100% (3/3)	100% (3/3)	7.55 $\pm$ 0.05
	Low bed temperature	100% (3/3)	100% (3/3)	100% (3/3)	10.42 $\pm$ 2.40
	Wind disturbance	33.3% (1/3)	33.3% (1/3)	33.3% (1/3)	9.95 $\pm$ 2.24
	Vibration disturbance	100% (3/3)	100% (3/3)	100% (3/3)	8.87 $\pm$ 1.33
	Total	86.7% (13/15)	86.7% (13/15)	86.7% (13/15)	9.42 $\pm$ 1.02
GPT-5.4	Fast print speed	100% (3/3)	100% (3/3)	100% (3/3)	3.76 $\pm$ 0.65
	Low flow rate	100% (3/3)	100% (3/3)	100% (3/3)	3.55 $\pm$ 0.20
	Low bed temperature	100% (3/3)	100% (3/3)	100% (3/3)	3.71 $\pm$ 0.95
	Wind disturbance	100% (3/3)	100% (3/3)	100% (3/3)	3.22 $\pm$ 0.40
	Vibration disturbance	100% (3/3)	100% (3/3)	100% (3/3)	3.68 $\pm$ 1.39
	Total	100% (15/15)	100% (15/15)	100% (15/15)	3.58 $\pm$ 0.47

Table A7 LLM API cost comparison (Input/output pricing per 1M tokens)

Model	Input	Output
Gemini 3.1 Pro Preview	\$2.00 ($\leq$ 200k)	\$12.00 ($\leq$ 200k)
	\$4.00 ($>$ 200k)	\$18.00 ($>$ 200k)
Claude Opus 4.6	\$5.00	\$25.00
GPT-5.4	\$2.50	\$15.00

Human Instruction:
Strong airflow is entering due to ventilation. Please adjust parameters if there is a problem.

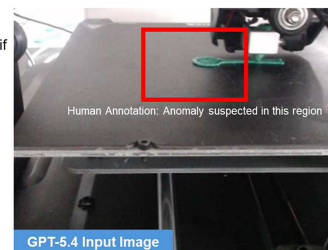
LLM Response — Gemini 3.1 Pro Preview:
"decision": "continue",
"param_updates": {},
"reason": "No visible defects currently despite the strong airflow. The print is adhering well. Continuing with current settings, but monitor for any signs of warping or under-extrusion."



(a)

Human Instruction:
Strong airflow is entering due to ventilation. Please adjust parameters if there is a problem.

LLM Response — GPT-5.4:
"decision": "adjust",
"param_updates": {"bed_temp_delta": 5, "fan_speed_delta": -40, "print_speed_scale": 0.9},
"reason": "Early layers show a thin line/weak adhesion risk; strong ventilation can cool PLA and lift edges. Raise bed temp, cut fan, and slow slightly to improve bonding."



(b)

Fig. A4 Comparison of LLM responses and human annotations under a wind disturbance condition. (a) Gemini 3.1 Pro Preview continues printing without parameter adjustment, despite the human annotator identifying an anomaly in the print region. (b) GPT-5.4 applies parameter adjustments (bed temp: +5, fan speed: -40, print speed: $\times 0.9$) citing weak adhesion risk due to strong ventilation

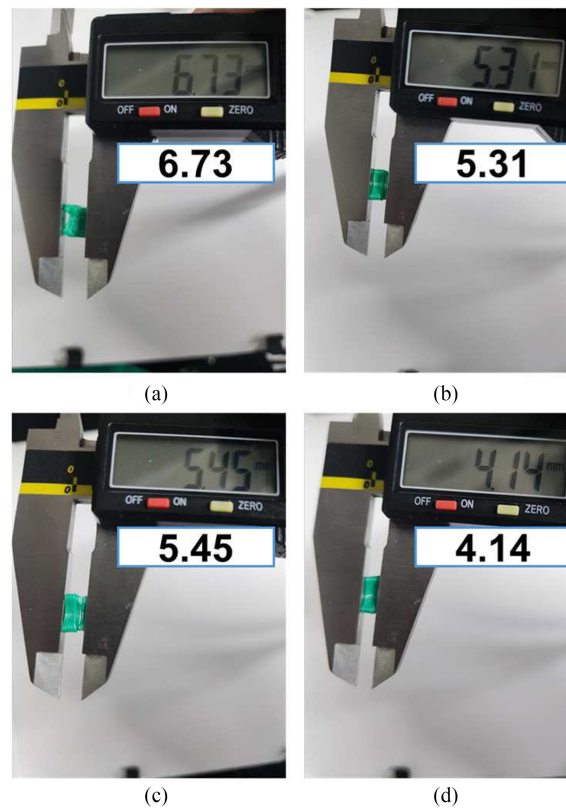


Fig. A5 Measured height of 3D-printed parts under fast print speed scenario: (a) Without system (6.73 mm) and (b)-(d) With system (5.31, 5.45, and 4.14 mm)

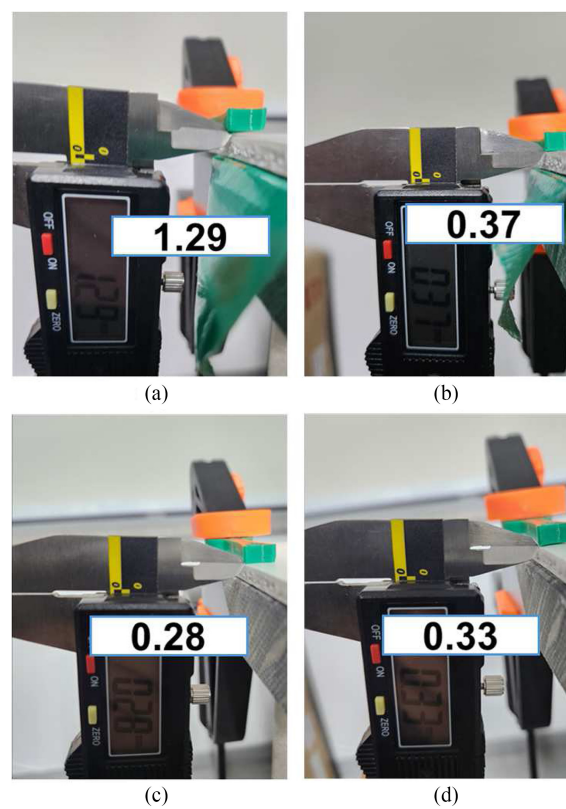


Fig. A6 Warping height of 3D-printed parts under low bed temperature scenario: (a) Without system (1.29 mm) and (b)-(d) With system (0.37, 0.28, and 0.33 mm)

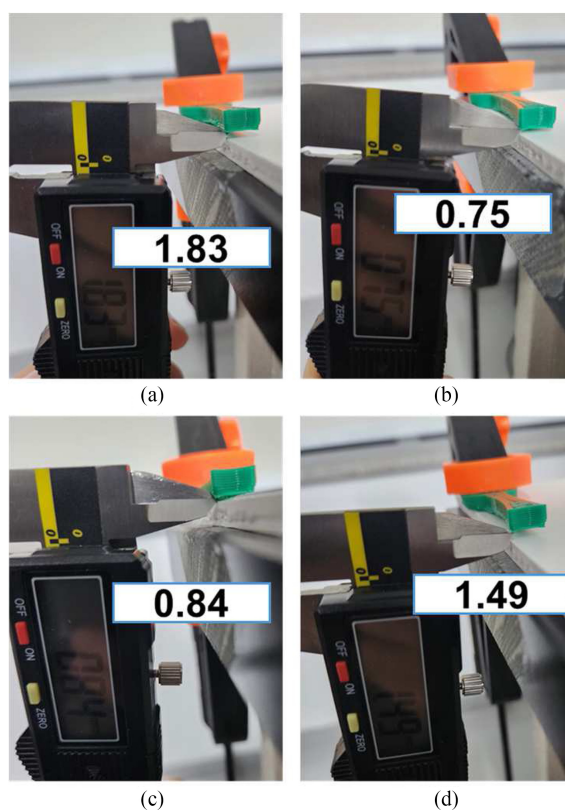


Fig. A7 Warping height of 3D-printed parts under wind disturbance scenario: (a) Without system (1.83 mm) and (b)-(d) With system (0.75, 0.84, and 1.49 mm)

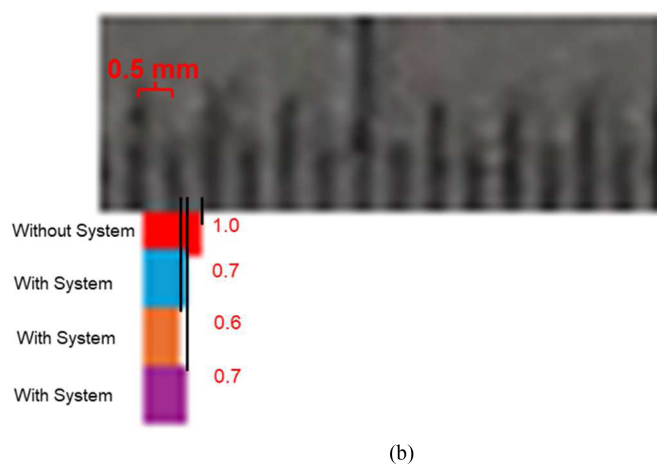
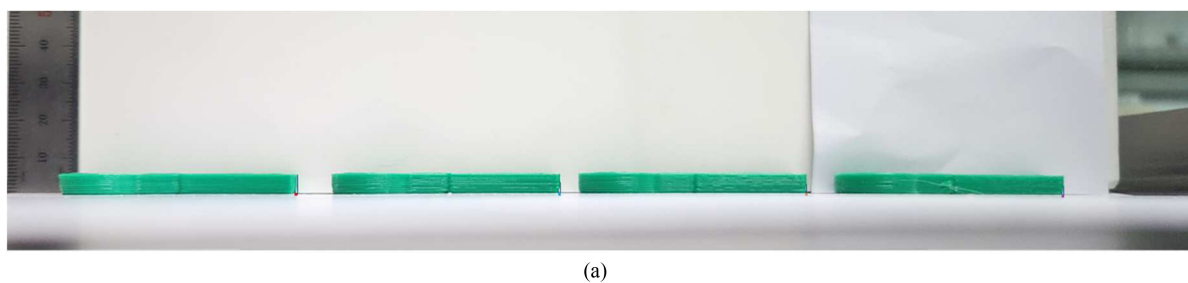
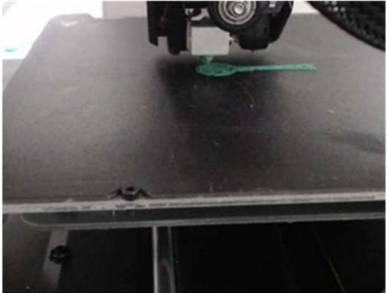


Fig. A8 Layer shift distance of 3D-printed parts under Vibration Disturbance scenario: (a) Overall view of printed samples, and (b) Measured layer shift distances. The top sample corresponds to the case without the system (1.0 mm), while the lower samples correspond to cases with the system (0.7, 0.6, and 0.7 mm)



Human Instruction:

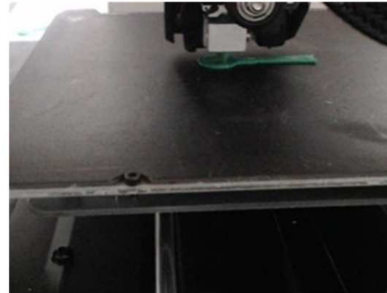
LLM Response — GPT-5.4:
 "decision": "continue",
 "param_updates": {},
 "reason": "Image shows small PLA print adhering to bed on layer 3 with no clear under/over-extrusion, lifting, or severe stringing. No visible defect justifies changing settings."



Human Instruction:
 Rattling noise is heard during movement. Please describe any defects and adjust parameters if there is a problem.

LLM Response — GPT-5.4:
 "decision": "adjust",
 "param_updates": {'print_speed_scale': 0.7, 'flow_scale': 0.9},
 "reason": "Image shows early layers printing without clear extrusion defect. Comment reports rattling during movement; current speed scale is excessive, so reduce speed to lower vibration/ringing risk."

(a)



Human Instruction:

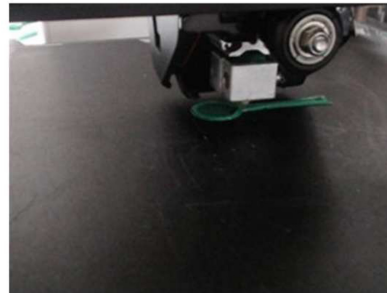
LLM Response — GPT-5.4:
 "decision": "adjust",
 "param_updates": {'flow_scale': 0.8, 'print_speed_scale': 0.9},
 "reason": "Image shows sparse/patchy extrusion on layer 3, indicating under-extrusion. Current flow_scale is very low at 0.7, raise flow and slow slightly to stabilize deposition."



Human Instruction:
 The amount of filament being extruded looks inconsistent. Please describe any defects and adjust parameters if there is a problem.

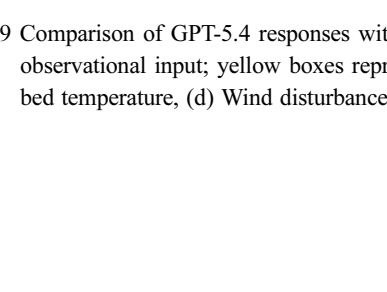
LLM Response — GPT-5.4:
 "decision": "adjust",
 "param_updates": {'flow_scale': 0.85, 'print_speed_scale': 0.9, 'nozzle_temp_delta': 5, 'fan_speed_delta': -40},
 "reason": "Image shows thin, patchy lines and poor top fill on layer 3, consistent with under-extrusion. Current flow 0.7 is very low; raise flow/temp, slow slightly, reduce cooling."

(b)



Human Instruction:

LLM Response — GPT-5.4:
 "decision": "continue",
 "param_updates": {},
 "reason": "Image shows small PLA patch on bed with no clear defect; adhesion appears adequate at layer 3. Evidence insufficient for a safe parameter change."



Human Instruction:
 I touched the bed and it feels barely warm. Please describe any defects and adjust parameters if there is a problem.

LLM Response — GPT-5.4:
 "decision": "adjust",
 "param_updates": {'bed_temp_delta': 5},
 "reason": "Image shows early layers still attached with no clear defect, but bed is only 30C at layer 3 and user reports barely warm bed; raise bed temp to improve PLA adhesion margin."

(c)

Human Instruction:

LLM Response — GPT-5.4:
 "decision": "continue",
 "param_updates": {},
 "reason": "Early layers look adhered and extrusion appears consistent. No clear defect visible in image/state; insufficient evidence for a safe parameter change."

Human Instruction:
 Airflow is entering due to ventilation. Please describe any defects and adjust parameters if there is a problem.

LLM Response — GPT-5.4:
 "decision": "adjust",
 "param_updates": {'fan_speed_delta': -40, 'bed_temp_delta': 5, 'nozzle_temp_delta': 5, 'print_speed_scale': 0.9},
 "reason": "Layer 3 shows narrow bead with weak first-track adhesion/lifting risk, consistent with draft cooling. Reduce fan, raise bed/nozzle temp, and slow slightly."

(d)

Human Instruction:

LLM Response — GPT-5.4:
 "decision": "adjust",
 "param_updates": {'print_speed_scale': 0.9, 'flow_scale': 0.9, 'fan_speed_delta': 20},
 "reason": "Image shows early-layer blob/flattened loop near nozzle, consistent with over-extrusion and poor cooling on layer 3; slightly reduce speed/flow and increase fan."

Human Instruction:
 The printer is vibrating due to nearby machinery. Please describe any defects and adjust parameters if there is a problem.

LLM Response — GPT-5.4:
 "decision": "adjust",
 "param_updates": {'print_speed_scale': 0.85},
 "reason": "Image shows smeared/offset early layers near nozzle, consistent with motion disturbance. Comment notes vibration; slowing print reduces inertial error and improves layer placement."

(e)

Fig. A9 Comparison of GPT-5.4 responses with and without user observational context for each scenario. blue boxes represent logs without observational input; yellow boxes represent logs with observational input provided. (a) Fast print speed, (b) Low flow rate, (c) Low bed temperature, (d) Wind disturbance, (e) Vibration disturbance

REFERENCES

1. Ngo, T. D., Kashani, A., Imbalzano, G., Nguyen, K. T., Hui, D., (2018), Additive manufacturing (3D printing): A review of materials, methods, applications and challenges, Composites Part B: Engineering, 143, 172-196.
2. N. Turner, B., Strong, R., A. Gold, S., (2014), A review of melt extrusion additive manufacturing processes: I. Process design and modeling, Rapid Prototyping Journal, 20(3), 192-204.
3. Tlegenov, Y., Hong, G. S., Lu, W. F., (2018), Nozzle condition monitoring in 3D printing, Robotics and Computer-Integrated Manufacturing, 54, 45-55.
4. Song, R., Telenko, C., (2017), Material and energy loss due to human and machine error in commercial FDM printers, Journal of Cleaner Production, 148, 895-904.
5. Chandrasekhar, A., Chan, J., Ogoke, F., Ajenifujah, O., Farimani, A. B., (2024), AMGPT: A large language model for contextual querying in additive manufacturing, Additive Manufacturing Letters, 11, 100232.
6. Jadhav, Y., Pak, P., Farimani, A. B., (2025), LLM-3D print: large language models to monitor and control 3D printing, Additive Manufacturing, 105027.
7. Kong, L., Peng, X., Chen, Y., Wang, P., Xu, M., (2020), Multi-sensor measurement and data fusion technology for manufacturing process monitoring: a literature review, International Journal of Extreme Manufacturing, 2(2), 022001.
8. OpenAI, New embedding models and api updates. <https://openai.com/blog/new-embedding-models-and-api-updates>
9. Chroma, Open-source search infrastructure for AI. <https://www.trychroma.com>
10. Tanenbaum, A. S., Bos, H., (2022). Modern operating systems, Pearson Education.
11. Ultimaker, Ultimaker cura. <https://ultimaker.com/software/ultimaker-cura>
12. Pallets Projects, Flask. <https://flask.palletsprojects.com>
13. Socket.IO, Introduction. <https://socket.io/docs>
14. SiSingh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., (2025), OpenAI GPT-5 system card, arXiv preprint arXiv:2601.03267.
15. Anthropic, System card: Claude Opus 4.6. <https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf>
16. Google DeepMind, Gemini 3.1 pro. <https://deepmind.google/models/model-cards/gemini-3-1-pro/>
17. Github, Langchain. <https://github.com/langchain-ai/langchain>

**Seongyeon Jeon**

M.S. candidate in the Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology. His research interest is Vision-Language Models.
E-mail: jeonsy639710@unist.ac.kr

**Taehwan Kim**

Ph.D. candidate in the Department of Mechanical Engineering, Ulsan National Institute of Science and Technology. His research interest is digital twin.
E-mail: wpemcbd2@unist.ac.kr

**Namhun Kim**

Professor in the Department of Mechanical Engineering and the Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology. His research interests are 3D printing, manufacturing system modeling and agent-based simulation.
E-mail: nhkim@unist.ac.kr