한국정밀공학회지 제 43 권 제 3 호 pp. 237-245 March 2026 / 237
J. Korean Soc. Precis. Eng., Vol. 43, No. 3, pp. 237-245 http://doi.org/10.7736/JKSPE.025.00031
ISSN 1225-9071 (Print) / 2287-8769 (Online)

• 특집 • 정밀 설계와 재료 기술을 활용한 제조 혁신

생성형 AI를 활용한 제조 지능화: LLM 활용방안과 정보모델링 연계

Generative AI–enabled Intelligent Manufacturing: LLM Utilization Strategies and Information Modeling Integration

이예진[1], 김동찬[1,#]
Ye Jin Lee[1] and Dong Chan Kim[1,#]

1 한국공학대학교 기계공학과 (Department of Mechanical Engineering, Tech University of Korea)
Corresponding Author / E-mail: dckim@tukorea.ac.kr, TEL: +82-31-8041-0418
ORCID: 0000-0001-7180-4012

KEYWORDS: Generative AI (생성형 인공지능), Large language models (대규모 언어모델), Intelligent manufacturing (지능형제조), Information modeling (정보 모델링), Retrieval-augmented generation (검색 증강 생성), Ontology (온톨로지)

This paper examines the role of generative AI and large language models (LLMs) in advancing intelligent manufacturing as we transition from Industry 4.0 to Industry 5.0. We begin by analyzing the current limitations of rule-based and manufacturing data systems in facilitating flexible, human-centric production. Next, we categorize LLM utilization strategies into three methodological axes: fine-tuning domain-specific models, employing general-purpose models through prompt engineering, and utilizing retrieval-augmented generation (RAG), which includes multimodal RAG that integrates sensor and text data. For each strategy, we present representative case studies across key application areas such as asset management, maintenance intelligence, quality control, process optimization, and knowledge- and document-centric support systems. Concurrently, we explore how information modeling and ontology-based knowledge graphs can be integrated with LLMs to enhance structured manufacturing semantics, improve source traceability, and minimize hallucinations. Finally, we summarize the advantages and limitations of each approach and propose future research directions for human-centric manufacturing, including the development of trustworthy LLM pipelines, standardized data schemas, and closer integration between digital twins and LLM-based decision support systems.

Manuscript received: November 25, 2025 / Revised: December 8, 2025 / Accepted: December 11, 2025

1. 서론

I4.0 (Industry 4.0)의 디지털 트윈 기술은 다양한 제조 산업 과제에 대응할 수 있는 방법에 대한 기반을 제공하고 있다. 그러나 디지털 트윈 기반 자동화 기술이 고도화되어 있음에도 불구하고, 유연성, 맞춤형 생산, 고유성이 요구되는 공정에서는 완전한 자동화를 구현하기 어렵다. 특히, 제조 산업은 스마트 제조, 대량 맞춤생산 등의 요구에 충분히 대응하지 못한 채 여러 문제에 직면하고 있다[1]. 전통적인 로봇은 특정 작업에 대해 하드코딩된 사전 프로그래밍 방식으로 작동하며, 동적 생산 요구와 맞춤화에 적응하기가 힘들고[2,3], 이를 해결하기 위한 디지털 트윈, DL (Deep Learning), ML (Machine Learning), IoT (Internet of Things) 기술들이 존재하지만, 높은 구현 비용, 인력 기술 격차 등의 장벽이 존재한다[4]. 이러한 문제점을 해결하기 위해 로봇과 AI (Artificial Intelligence)를 활용한 공정 시스템이 구현되고 있지만[2], 수집된 공정 데이터를 기술자와 AI 서비스가 융합하여 사용자 친화적 기술로 발전하기에는 여전히 부족하다. 또한, 산업 현장에서 현장 전문가들이 단순한 작업자에 머무르지 않고 디지털 기술을 이해하고 활용할 수 있는 고도의 전문 지식을 요구하고 있다[5].

이와 같이, I4.0의 고도화로 인해 공정에서 인간이 수행하는 육체적 작업은 줄어드는 추세이지만, 미래 산업에서는 사람(사회적 요소), 사이버 공간(정보 및 디지털 기술), 물리적 세계(생산 설비 및 공정)가 유기적으로 결합하여 작동하는 고도화된 제조 시스템인 Sociocyber-physical Manufacturing System [6]의 관점이 자리 잡히고 있으며, 작업자가 복잡한 시스템을 이해하고 응용할 수 있는 능력이 필요하다. 더하여, 전문가들이 더 많은 의사결정 및 문제 해결 업무를 담당하는 미래가 예상되지만, 인구 감소 및 숙련 인구의 축소로 인해 생기는 전문가의 지식 전승 단절이 우려되는 상황이다. 해당 산업은 이러한 문제점을 해결하고자 다가오는 I5.0 (Industry 5.0)에 인간-AI (Human-AI) 협업 프로세스 설계 및 혁신 방법론의 개발과 구현을 목표로 하고 있다[1,5,7]. 그 중에서도 5차 산업의 핵심 목표는 사용자 친화적인 인터페이스와 신뢰 가능하고 설명 가능한 AI 지식 서비스를 바탕으로 공장 특화 협업 플랫폼을 구성하고, 이를 즉시 사용자에게 제공하여, 협업할 수 있는 사용자 제어형 협업 플랫폼 혁신 방법론을 제공하고자 한다[8].

그러나, 복잡하고 다양한 데이터가 공존하는 제조 현장 환경에서 생성형 AI가 맥락을 정확히 이해하고 안전하게 활용되기에 여러가지의 한계를 직면하고 있다[9,10]. 이러한 문제를 보완하기 위해 제조 산업에서는 크게 세 가지 LLM (Large Language Model) 활용 접근법이 연구되고 있다[11].

한편, 제조 산업에서 LLM을 적용하려는 연구는 응용 관점에서도 서로 다른 카테고리로 분화되고 있다. 대표적으로, 설비 상태 모니터링, 예측 유지보수, 공정 조건 추천, 공정 계획 최적화, QA 및 운영 지원을 다루는 문서 중심 지원 시스템 등이 있다. 현장 데이터는 직무별로 다양한 특성과 요구사항을 가지기 때문에 어떤 방법론이 더 적합한지 분석하는 것이 중요해지고 있다.

위와 같이 제조 산업 전반에 LLM을 융합하기 위한 다양한 연구가 진행되고 있으나, 기존 문헌들은 개별 응용 사례 소개에 머무르거나 특정 방법론에 편중된 경향이 있다. 따라서 본 논문에서는 FTLM (Fine-tuning Language Model), PELM (Prompt Engineering Language Model), RAG LLM (Retrieval-Augmented Generation Large Language Model)의 세 가지 LLM 활용 방안과 유지보수, 품질 및 공정 최적화, 도메인 QA 시스템의 응용 등 3가지의 제조 산업 대표 사례를 정리하고, 각 접근의 도메인 적합성, 신뢰성과 설명가능성, 데이터 구축 및 운영 비용 측면의 차별점을 분석함으로써 제조 산업에서 어떤 상황에 어떤 LLM 활용 전략이 적합한지 평가할 수 있는 실무적 기준을 제시하는 데에 기여를 하고자 한다. 또한 리뷰 과정의 타당성을 확보하기 위하여, 본 논문에서 검토한 주요 문헌은 제조 도메인 데이터를 다루면서 LLM 또는 생성형 AI를 핵심 구성요소로 활용한 연구를 중심으로 선정하였으며, 각 응용 카테고리별 사례 소개 시에는 이러한 선정 기준을 바탕으로 대표 연구를 선별하였다.

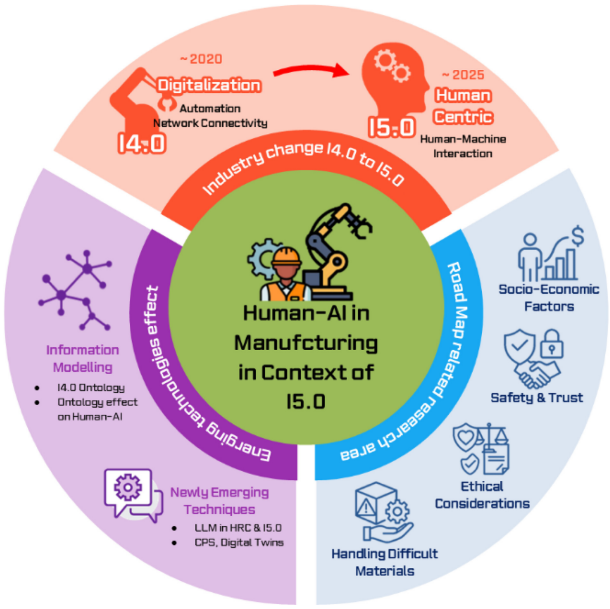

Fig. 1 Overview of human-AI collaboration in manufacturing in the context of I5.0, illustrating the transition from digitalization-focused I4.0 to human-centric I5.0 and summarizing key research areas: ontology-based information modelling, emerging techniques such as LLMs and digital twins, and roadmap issues including handling difficult materials, ethical considerations, safety and trust, and socio-economic factors

2. LLM 모델의 이해

LLM은 방대한 양의 텍스트 데이터로 학습된 딥러닝 기반 인공지능 모델로서, 자연어 처리 및 생성 작업에서 유능한 AI 모델이다[12]. NLP (Natural Language Processing)와 AI 모델의 발전은 1990년대 중반 규칙 기반 시스템에서 시작한다. 1990년대 후반에는 통계 모델로, 그리고 2000년대 초반에는 신경망 모델로 옮겨왔다[13]. RNN (Recurrent Neural Network) 기반의 'Self-attention' 및 'Transformer 기반' 신경망 구조의 도입으로부터 시작하여[14,15], 2010년대 후반에 FTLM의 확산에 크게 기여했다. 이러한 FTLM은 방대한 데이터로부터 인간의 개입 없이 보편적인 언어 표현을 학습할 수 있으며, 최근 몇 년간, LLM의 개발과 발전은 GPT-3 (OpenAI), PaLM (Google), LLaMA (Meta), Megatron-turing NLG (NVIDIA) 등을 대표로, 눈에 띄게 발전해오고 있다. 특히, LLM은 효율적인 문서 분석과 창의적인 아이디어 생성을 돕고, 복잡한 데이터 해석을 지원이 가능하다. 이러한 장점 덕분에 연구를 가속화하고, 발견 과정의 효율을 높이며, 학제 간 협력을 촉진함으로써 과학 분야와 사회과학 모두를 혁신할 수 있는 잠재력을 지니고 있다[12]. LLM 능력은 단순한 ChatBot (e.g. OpenAI 2023)을 넘어 제조 산업에서 인간과 기계 간의 대화형 게이트웨이 역할을 수행하며,

방대한 제조 데이터를 해석하여 정보에 기반한 의사결정을 촉진하고, 생산 및 관리 영역에 자연어 사용의 미래를 열어가고 있다[16-18]. 또한, 제조 분야에서 LLM의 배포는 제품 설계 및 개발부터 품질 관리, 공급망 최적화, 인력 관리에 이르기까지 제조의 다양한 측면을 자동화하고 향상시킬 수 있는 잠재력을 지니고 있다[19].

3. LLM 활용 방법론

3.1 FTLM 기반 방법론

FTLM은 사전 학습된 LLM을 특정 도메인이나 작업에 맞게 추가 학습시키는 전이학습 기법이다. 해당 방법론은 모든 파라미터를 업데이트하여 가장 높은 성능을 제공하지만, 방대한 계산 자원이 필요하다. 이를 극복하기 위해 아래와 같이 다양한 효율적인 파라미터를 개발한 Parameter-efficient Fine-tuning 기법들이 연구되었다.

- LoRA (Low-rank Adaptation) [20]: Hu et al. (2021) 이 제안한 방식으로, 모델의 가중치 행렬을 저차원 분해하여 소수의 파라미터만 학습한다. 원본 모델은 동결하고 추가된 저차원 행렬만 업데이트하여 학습 가능한 파라미터를 90% 이상 줄이면서도 전체 Fine-tuning에 근접한 성능을 달성하였다.
- QLoRA [21]: Dettmers et al. (2023) 의 방법으로, 모델을 4비트로 양자화하여 메모리 사용량을 추가로 감소시켜 단일 GPU에서도 대규모 모델 Fine-tuning이 가능하다.
- Adapter Layers [13]: Houlsby et al. (2019)이 제안한 방식으로, 기존 Transformer 레이어 사이에 작은 신경망 모듈을 삽입하여 여러 작업에 대해 각각의 어댑터를 학습하고 교체 사용할 수 있다.

제조 산업 적용 시에는 지속적 사전 학습과 지도 미세조정을 순차적으로 적용하는 전략이 효과적이다. Lu et al. (2025) [22]은 재료과학 도메인 적응 연구에서 CPT, SFT, DPO 등 다양한 전략을 비교하며, 여러 Fine-tuning된 모델의 병합이 개별 모델을 초과하는 시너지 능력을 창출할 수 있음을 확인한 반면, Rafailov et al. (2023)의 DPO (Direct Preference Optimization) [23]는 별도의 보상 모델 없이 선호도 데이터로 직접 학습하여 RLHF보다 간단하고 효율적이다.

위의 연구들을 통해 FTLM 방법론이 도메인 전문성 깊이 학습, 높은 정확도, 일관된 출력 스타일 제어가 가능하다는 장점이 존재하나, 대량의 레이블된 데이터 요구, 높은 계산 비용, 지식 업데이트 시 장시간 학습 등과 같은 한계점이 존재한다. 해당 방법론은 최적화하는 특정 작업, 복잡한 도메인 지식이 필요한 현장, 높은 정확도 요구하는 상황에는 적합하지만, 실시간으로 대응이 필요한 제조 산업에서의 도입은 어려움이 보이고 있다.

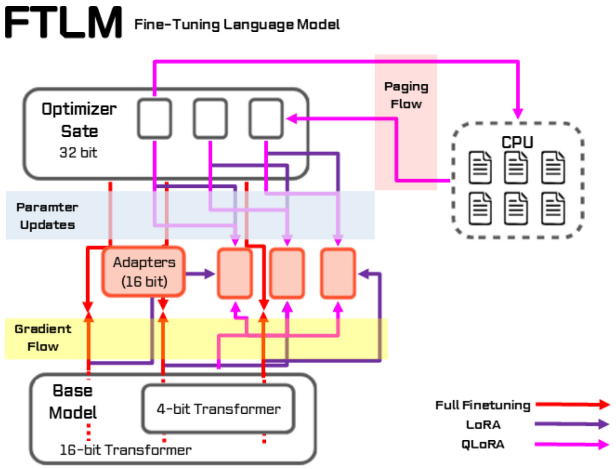

Fig. 2 Schematic of the FTLM fine-tuning framework, comparing full fine-tuning, LoRA, and QLoRA by showing how gradient flow, parameter updates, low-bit adapters, and optimizer state are organized across the 16-bit base model, 4-bit transformer, and CPU paging [23] (Adapted from Ref. 23 on the basis of OA)

3.2 PELM 기반 방법론

PELM은 LLM에 제공하는 입력 지시사항을 전략적으로 설계하고 최적화하여 정확하고 유용한 출력을 얻는 기법이다. 이 방법론은 자연어로 작성된 지시문이나 질문을 어떻게 만들고 전달하는지에 따라, LLM의 응답 품질을 최적화되기 때문에 아래의 학습 방법들로 연구가 진행되었다.

- Zero-shot Prompting [24]: 사전 예시 없이 작업 설명만으로 LLM이 응답하도록 하는 가장 기본적인 방식이다.
- Few-shot Prompting [25]: 몇 가지 예시를 제공하여 LLM이 패턴을 학습하고 유사한 작업을 수행하도록 유도하며, 컨텍스트 내 학습(In-Context Learning)을 활용한다.
- CoT (Chain-of-Thought) Prompting [26]: Wei et al. (2022)이 제안한 방식으로, LLM이 단계별 추론 과정을 거치도록 유도한다. Microsoft 연구(2023)에 따르면 CoT는 다단계 작업에서 LLM의 추론 정확도를 최대 40%까지 향상시켰다.
- ToT (Tree-of-Thought) Prompting [27]: Yao et al. (2023)의 방법으로, CoT의 선형적 추론을 확장하여 여러 추론 경로를 병렬로 탐색하고 최적의 옵션을 선택하는 방식이다.

Ouyang et al. (2022) [28]은 인간 피드백을 통해 지시사항을 따르도록 언어 모델을 학습시키는 방법을 제시했으며, 이는 Prompt Engineering의 효과를 높이는 기반이 되었다. 제조 산업에서는 역할 기반의 Role-based Prompting을 통해 '당신은 20년 경력의 제조 공정 엔지니어입니다'와 같이 설정된 전문가 페르소나가 맥락에 맞는 응답을 유도한다.

다양한 연구를 통해 PELM은 실시간 적용 가능, 효율적인 비용, 높은 유연성 등의 장점을 보였으나, 복잡한 도메인 지식이 필요한 작업에서는 제한적인 성능, Prompt 설계에 따라 가변적인

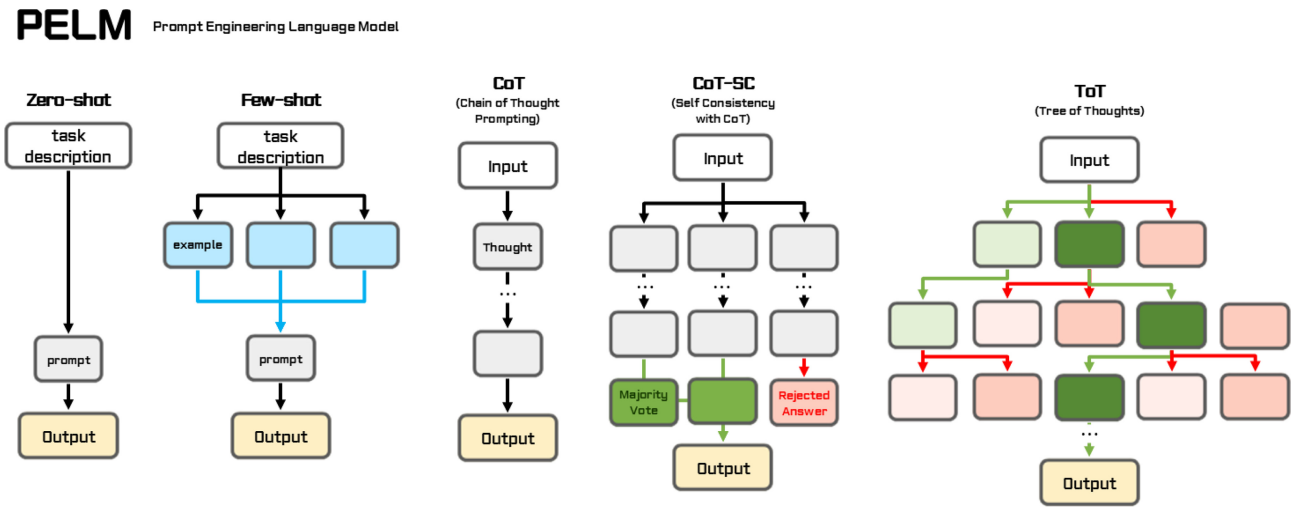

Fig. 3 Overview of prompt engineering strategies in PELM, contrasting zero-shot, few-shot, CoT, CoT-SC (self-consistent CoT), and ToT (tree-of-thoughts) prompting and illustrating how different prompting and reasoning structures lead to the final model output

결과에 대해 한계가 보인다. 다만, 빠른 프로토타이핑이 필요한 산업이나, 간단한 작업, 범용 활용이 필요한 상황에서는 필요한 방법론이다.

3.3 RAG LLM 기반 방법론

Lewis et al. (2020) [11]이 제안한 검색 증강 생성 기법인 RAG은 외부 지식 베이스에서 관련 정보를 검색하여 LLM의 응답 생성에 통합하는 하이브리드 AI 기법이다. RAG LLM 시스템은 세 가지 핵심 구성요소로 이루어진다: (1) 검색 시스템(Retriever)이 벡터 DB에서 관련 문서를 찾고, (2) 컨텍스트 통합(Context Integration)이 검색된 문서를 LLM의 Prompt에 통합하며, (3) 생성 시스템(Generator)이 검색된 정보를 기반으로 응답을 생성한다. 해당 RAG LLM 방법론은 3가지 파이프라인으로 분류된다.

- Naive RAG: 문서 전처리 단계에서 PDF, Word, 웹페이지 등 다양한 형식의 문서를 수집하고 텍스트를 추출한다. 긴 문서를 의미 있는 작은 조각으로 분할하는 과정인 Chunking 방법을 통해 의미 있는 정보를 200-500 토큰 크기로 분할한다. 분할된 각 Chunking를 고차원 벡터로 변환하여 임베딩을 생성하고, 원하는 정보를 DB를 통해 검색을 한다. Gao et al. (2023) [29]의 RAG Survey는 다양한 검색 전략과 생성 기법을 포괄적으로 분석했으며, Karpukhin et al. (2020) [30]의 DPR (Dense Passage Retrieval)은 질문과 문서를 동일한 임베딩 공간에 표현하여 효율적인 검색을 가능하게 했다.
- Advanced RAG: Asai et al. (2024) [31]의 Self-RAG가 자체 반성을 통해 검색, 생성, 비평을 학습하며, Yan et al. (2024) [32]의 Corrective RAG는 검색 결과를 평가하고 필요 시 재검색을 수행한다. Guu et al. (2020) [33]의 REALM, Izacard & Grave (2021) [34]의 FiD, Borgeaud et al. (2022) [35]의 RETRO 등 고급 RAG 아키텍처를 제시한다[35-38].
- Ontology-based RAG: 온톨로지는 해당 도메인의 개념, 속성, 관계를 정의한 형식적 지식 모델로서, 서로 다른 시스템 간 정보의 의미적 통합과 상호 작용이 가능하다. 지식그래프는 이러한 온톨로지 등에 기초하여 실제 데이터를 그래프 구조로 표현한 지식베이스로, 제조 기업에서는 제품 구조, 공정 흐름, 자원 정보, 설비 상태 등 다양한 정보를 KG로 관리하기 시작했다[39]. LLM은 비정형 텍스트에서 개념과 관계를 자동 추출하는 뛰어난 특성과 KG DB를 융합하여 효율성을 입증하는 연구가 늘어나고 있다. Xu et al. [40]은 LLM을 통해 초기 텍스트 엔티티 레이블링 작업을 자동화하고 전문가 검증을 최소화함으로써, 기존 대비 절반 수준의 시간·노력으로 정밀한 제조 지식그래프를 구축했다. 또한, Wang, P., et al. (2024) [41]은 온톨로지 통합 항공기 유지보수 연구를 통해 항공기 부품 계층 구조를 온톨로지로 표현하고 유지보수 로그와 결합하여 결함 부품 식별 정확도 30% 향상과 인공지능의 환각을 줄이는 결과를 확인하였다.

이러한 다양한 연구를 통해 RAG LLM는 모델 재학습 없이 최신 정보 활용, 출처 추적 가능, 비용 효율적이라는 장점이 있으나, 검색 품질에 의존, 응답 지연 시간 증가, 복잡한 추론 한계가 존재한다. 방대한 문서 활용, 지식 업데이트 빈번, 규제 준수 필요한 산업에서 유리하다.

4. 제조 산업에서의 LLM 연구 사례

LLM 기술은 제조 산업의 여러 영역에서 다양한 방식으로 응용되고 있다. 본 장에서는 제조 분야를 세 가지 카테고리인 설비 유지보수, 품질 관리 및 공정 최적화, 도메인 QA 시스템으로 구분하여, 어떤 방식으로 제조 산업 분야에 LLM이 도입되고

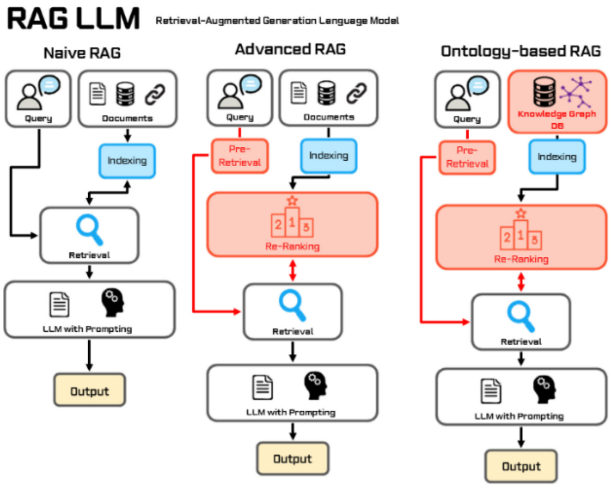

Fig. 4 Comparison of RAG architectures for LLMs, contrasting naive RAG, advanced RAG with pre-retrieval re-ranking, and ontology-based RAG that leverages a knowledge graph database to improve retrieval quality and downstream LLM outputs

있는지, 성과 및 한계를 제시하고자 한다. 더하여, 각 분야에서 보고된 최신 LLM 활용 연구 사례들과 적용 효과를 살펴보고자 한다.

4.1 LLM을 활용한 유지보수 관리 응용 사례

설비 유지보수 분야에서는 예지보전과 고장 예측, 유지보수 기록 분석 등에 LLM을 접목하려는 시도가 나타나고 있다. 예측 유지보수인 경우, 기존 센서 데이터 기반으로 고장을 예측했다면, 이제는 LLM이 시계열 데이터 해석 능력과 추론 능력을 활용하여 기존보다 설명력 높은 예측과 조언을 얻을 수 있다. 예를 들어, Deng. et al. (2024) [42]은 LLM 에이전트가 기계 학습 모듈로부터 입력 받은 고장 예측 결과를 해석하고, 유지보수 의사결정을 지원하도록 하는 프레임워크를 제시하였다. 해당 연구는 GPT-4 모델에 RAG LLM과 Tool 사용 능력을 부여하여, 예측된 고장 원인에 따라 상황에 맞는 정비 조치를 제안하도록 하였다. 그 결과 추가 학습 없이도 기존 방법과 유사한 수준의 고장 탐지 성능을 보임과 동시에, 다양한 고장 분석과 조치 방안을 설명하고 제공할 수 있음을 검증하였다. 해당 연구는 RAG LLM이 단순 예측을 넘어 전문가의 역할까지 수행하며, 유지보수 업무에 대해 해결책을 제시하는 잠재력을 보여준다.

또 다른 접근으로, 도메인 특화 LLM을 활용한 사례도 있다. Wang & Li. et al. (2023) [43]의 연구에서는 제조 설비 유지보수 안내를 돕기 위해 유지보수 매뉴얼의 데이터를 LLM에게 Fine-tuning한 FTLM을 활용하였다. 일반 모델 대비 현장 질문에 대해 정확하고 구체적인 답변 능력을 보였으며, 이는 유지보수 분야에 전문용어와 맥락을 학습이 가능함과 동시에, 현장 적합성이 높음을 보여준다. 시스템에서는 현장 장비의 상태를 모니터링하면서, 고장이 감지되거나 작업자가 지식을 요청하면 FTLM이 적절한 조치 절차를 제시하고 AR로 표시하여 사람-AI 협업을 구현하였다. 이러한 연구들은 LLM이 제조 산업의 유지보수 관리 분야에 실시간 조언, 기록 분석가, 교육 도구 등으로 활용될 수 있음을 확인할 수 있다. 한편, Jones et al. (2025) [44]는 자산 디지털 트윈 환경에서 예지보전 계획 (PdM Planning)에 LLM을 도입하여, 과거 정비 계획 패키지와 관련 문서를 RAG로 검색한 뒤 LLM이 반복 작업의 정비 패키지를 자동 생성하거나 수정이 가능하도록 하는 워크플로를 제안하였다. 이러한 연구들은 유지보수 분야에서 RAG 기반 문서 검색, 도메인 온톨로지, FTLM을 결합하여 예지보전 의사결정과 정비 계획 수립을 지원하는 LLM 활용이 확산되고 있음을 보여준다.

4.2 LLM을 활용한 품질 및 공정 최적화 응용 사례

품질 관리 및 공정 최적화 영역에서는 LLM이 생산 공정 데이터 분석, 품질 이상 징후 감지, 공정 파라미터 최적화 등 다양한 역할을 구현할 수 있다. 특히, 기존 데이터에 존재하는 영상이나 센서 데이터 기반으로 이루어지던 품질 검사에 LLM의 텍스트 처리 및 추론 능력을 결합해 종합적인 품질 판단을 내리는 연구들이 나타나고 있다. 한 예로, Badini. et al. (2023) [45]은 적층 제조 공정에 ChatGPT를 활용하여 품질 문제를 해결하는 연구를 수행하였다. 3D 프린팅 공정 중에 발생하는 품질 문제 (e.g. 제품이 베드에서 떨어짐, Warping, 출력 중 실 끊김 등)에 대해 ChatGPT에게 질문하였고, 그 결과 LLM이 높은 정확도와 조직적인 문제 해결 방안을 제시하여 전문가 수준의 트러블슈팅을 해결하는 결과를 보였다. 이는 LLM이 공정 엔지니어의 QA 조언자로서 품질 이슈에 대한 지식기반 해결책을 제시할 수 있음을 제시한다. 또 다른 사례로, Fan. et al. (2024) [46]의 연구에서 ISF (Incremental Sheet Forming) 공정과 관련된 방대한 논문을 지식 추론 도구인 Ontology-based RAG LLM을 이용해 분석하여, 사용자가 질문을 하면, 공정 매개변수 조정에 대한 조언을 얻는 연구를 진행하였다. 해당 연구는 LLM이 문헌 내 실험 결과와 이론을 종합하고, 자동으로 문헌 리뷰를 하면서, 최적 공정 조건을 찾는 시간을 단축하는 결과를 보였다. Liu et al. (2025) [47]은 LLM을 기반으로 가상 보조원을 실제 제조 환경에 배치하고, 기술문서 기반 QA, 유지보수 절차 안내, 품질 기준 문서 응답 등 다양한 작업을 지원하는 RAG 아키텍처를 구현하였다. 해당 시스템은 검색기반 문서 필터링과 LLM 응답 생성을 결합하여 높은 정확도와 신뢰도를 달성했으며, 실제 사용자의 문서 응답 시간 단축과 정확도 향상에 기여하였다. 이처럼 Ontology-based RAG LLM을 통해 숨겨진 공정 지식 발굴이나 다분야 최적화 정보를 얻는 접근법은 작업자가 일일이 모든 정보를 읽지 않고 모델의 추론 능력을 활용해 창의적 해결법이 단시간에 모색 가능하다.

이와 같이 품질, 공정 부문에서는 대화형 문제 해결, 지식 기반 공정 개선, 공정 최적화 등 다양한 방향으로 LLM 활용이 진행되고 있다. PELM부터 FTLM까지 혼용되어 사용되고 있으며, 품질 이상을 사전에 예측하고 공정 변수를 동적으로 최적화하는 데 LLM의 추리력과 학습능력이 기여하고 있다[9]. 향후에는

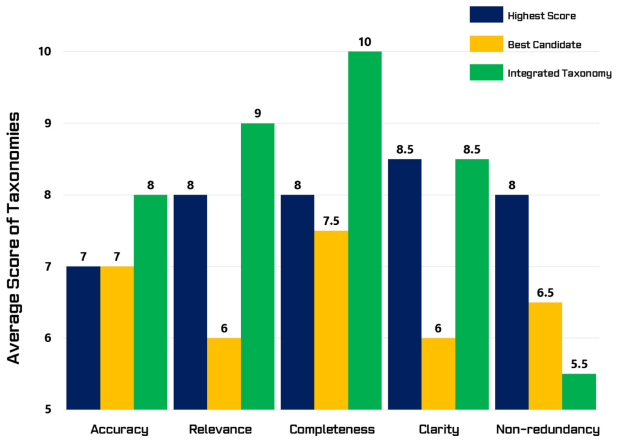

Fig. 5 Expert evaluation of taxonomy quality for the ontology-based RAG LLM, reporting average scores across five criteria (accuracy, relevance, completeness, clarity, and non-redundancy) for three settings: the highest-scoring baseline taxonomy, the best candidate, and the integrated taxonomy generated by the ontology-based RAG LLM, showing that the integrated taxonomy achieves the best overall performance [48] (Adapted from Ref. 48 on the basis of OA)

품질 데이터(e.g. 결함 보고서, 검사 결과)를 LLM이 실시간 분석하여 품질 경향을 모니터링하고, 공정 제어 시스템에 피드백을 주는 형태로 발전할 것이 예상된다.

4.3 LLM을 활용한 도메인 QA 시스템 응용 사례

제조 기업에서는 방대한 기술 문서, 설계 사양, 작업 지침, 유지보수 이력 등이 다양한 데이터가 생성이 되면서, 이를 효율적으로 활용하기 위한 지식 기반 지원 시스템에 LLM이 적극 도입되고 있다. LLM 기반 ChatBot이나 도메인 QA 시스템을 구축하여, 현장 작업자나 엔지니어의 질문에 대해 신속하고 정확한 답변을 제공하는 연구들이 진행 중이다. 예를 들어, Kernan Freire. et al. (2024) [48]의 연구에서 공장 내부의 설비 매뉴얼, 문제 해결 보고서 등 문서를 모두 통합한 LLM 응답 시스템을 개발하였다. 이 시스템은 운영자가 자연어로 질문하면 RAG LLM을 통해 답변을 생성하는 형태로, 실제 현장 테스트에서 정보 검색 속도 향상과 이슈해결에 대한 효율성이 증대하는 효과를 보였다. 그러나, 사용자들은 사람이 답하는 것에 익숙해 여전히 인간 전문가를 선호하는 경향이 나타났으며, 신뢰성과 논리 측면에서 시스템 개선이 필요하다는 한계가 존재한다. 다른 예시로, Sobhan. et al. (2025) [49]의 연구에서는 데이터가 디지털 데이터가 아닌 표와 이미지가 많고 스캔본이 활용되는 제조 공정의 기술 문서를 이용하여, 테이블에서 이미지까지 처리하는 RAG 파이프라인을 통해 기술 문서 질의응답의 정확도를 크게 향상시키는 연구를 진행하였다. 해당 연구는 벡터DB를 활용한 유사도 검색 후, FTLM을 활용하여, 유사 데이터를 선별해 답변의 사실 충실도를 95% 이상으로 끌어올리는 결과를 보였다. 이는 복잡한 제조 문서의 경우 단순 키워드 매칭이 아닌 LLM의 언어 이해력을 접목한 하이브리드 검색이 필요하며, LLM이 문맥을 재해석하여 관련 정보를 재구성하는 능력을 보여주는 연구이다.

이와 같이, 문서 지원 분야인 도메인 QA 시스템에서 LLM은 엔터프라이즈 검색의 고도화, 현장 작업자 교육, 문서 자동화 등에 응용되고 있다. 응용 연구들은 대량의 비정형 지식을 LLM으로 하여금 이해시키고 필요한 부분을 정확히 찾아 응답하게 하는 것이며, 이를 위해 이를 위해 프롬프트 최적화와 RAG 결합 전략이 널리 활용되고 있다[49]. 또 다른 사례로는 LLM을 활용하여 대량의 비정형 데이터를 자동으로 온톨로지 및 KG로 변환한 연구가 보고되고 있다. Trajanoska. et al. (2023) [50]의 연구에서 지속가능성 보고서의 방대한 텍스트 데이터를 GPT-4가 추출하여 도메인 온톨로지 개념에 맞게 분류하고 Triple을 생성함으로써 지식그래프를 활용한 Ontology-based RAG LLM 구축하였다. 반면, Wahid. et al. (2024) [51]는 산업용 지식그래프 (Industry 4.0 KG) 에 LLM을 연계하여, 사용자들이 복잡한 SPARQL 질의나 데이터 스키마를 몰라도 자연어 질문만으로 필요한 정보를 얻을 수 있는 프레임워크를 소개하였다. 예를 들어 작업 현장의 엔지니어가 “어제 라인2에서 발생한 품질 이상 원인은 무엇인가?”와 같이 물으면, LLM이 KG를 조회하여 관련 Triple을 가져오고 이유를 설명하고, 현장 전문가들이 손쉽게 지식에 접근하도록 하였다. 한편, 도메인 QA 기능을 공정 전반의 의사결정 지원으로 확장하려는 시도도 보고되고 있다. Liu et al. (2025) [52]는 항공우주 부품 제조 품질 문제의 원인 분석을 지원하기 위해, 논문, 보고서, 표준 문서를 통합한 산업 지식베이스 위에 RAG LLM을 결합한 Knowledge Enhanced QA 시스템을 제안하였다.

이로써 현장에서는 사람들이 일일이 매뉴얼을 찾는 대신 자연어로 묻고 답을 얻는 생산성 향상이 기대되지만, 동시에 잘못된 정보 제공에 대한 검증 절차와 모델의 신뢰 형성이 계속 중요한 과제로 남아있다.

5. 결론

본 연구는 제조 산업 내 네 가지 주요 응용 카테고리를 중심으로 LLM의 활용 사례를 비교하고, 각 접근법의 특성과 시사점을 제시하였다. 유지보수 관리 분야에서는 RAG LLM과 도메인 맞춤형 FTLM이 병행되는 경향이 두드러졌으며, 대형 언어 모델을 기반으로 한 예지보전 의사결정 보조 및 현장 특화 QA 시스템이 성과를 보였다. 품질 및 공정 최적화 영역에서는 범용 LLM을 활용한 공정 개선 및 품질 예측, 제조 데이터 기반 스케줄링 최적화 등 다양한 접근이 시도되었고, 현장 데이터 활용 여부에 따라 PELM 기반 접근과 Fine-tuning 방식이 병행되었다. 도메인 QA 시스템 분야에서는 최신 문서를 반영하고 출처 기반의 신뢰성 있는 응답이 중요하게 작용하면서, RAG LLM이 사실상 표준으로 자리 잡고 있다.

Table 1 LLM-based QA and monitoring systems: comparison of question types and generated responses

Ref	System	Question	Response
[48]	LLM issue analysis tool	Free-form questions from operators about production issues or documents. e.g., how to resolve a specific issue	Chat-style answer plus an issue-analysis screen that lists the relevant document sections supporting the response.
[49]	Doc RAG QA system	Technical questions from equipment manuals. e.g., “Why is AC hi-pot test used?”, “How to complete the TTR test?”	RAG pipeline generates natural-language answers that closely follow the ground-truth text extracted from the manuals.
[51]	KG LLM monitoring system	Query about machine sensor data. e.g., “What is the temperature of Machine_6 in the specified time period?”	System answers that there are no temperature observations for Machine_6 in that period, based on the underlying KG data.

그러나 이러한 활용에도 불구하고, 제조 분야에서의 LLM 적용은 도메인 지식 부족, 환각 현상, 데이터 보안, 설명가능성, 비용과 범용성 한계 등 다양한 기술적 과제에 직면해 있다. 이에 따라 향후 연구는 시계열 센서, CAD 데이터, 텍스트 등을 통합한 멀티모달 LLM, 온톨로지 및 지식그래프 연계를 통한 신뢰성 강화, 현장 피드백을 반영한 지속학습 프레임워크, 경량화 및 엣지 적용 가능한 모델 최적화, 책임성과 투명성을 고려한 윤리적 활용 체계 구축에 집중될 필요가 있다.

본 논문은 LLM의 적용이 제조 산업 내 다양한 문제 해결 방식에 어떤 방식으로 기여하는지를 구조적으로 정리하고, 각 기술의 장단점을 비교함으로써 향후 도입 전략과 연구 방향 기반을 제공하고자 하였다.

ACKNOWLEDGEMENT

This study was supported by research fund and Academic Promotion System from Tech University of Korea (2025), and the National Research Foundation of Korea (NRF) grant funded by the Ministry of Education (No. NRF-2017R1A6A1A03015562).

REFERENCES

1. Wilhelm, J., Petzoldt, C., Beinke, T., Freitag, M., (2021), Review of digital twin-based interaction in smart manufacturing: Enabling cyber-physical systems for human-machine interaction, International Journal of Computer Integrated Manufacturing, 34(10), 1031-1048.
2. Mukherjee, D., Gupta, K., Chang, L. H., Najjaran, H., (2022), A survey of robot learning strategies for human-robot collaboration in industrial settings, Robotics and Computer-Integrated Manufacturing, 73, 102231.
3. Shah, R., Doss, A. S. A., Lakshmaiya, N., (2025), Advancements in ai-enhanced collaborative robotics: Towards safer, smarter, and human-centric industrial automation, Results in Engineering, 105704.
4. Fekrisari, M., Kantola, J., (2024), Integrating industry 4.0 in manufacturing: Overcoming challenges and optimizing processes (case studies), The TQM Journal, 36(9), 347-370.
5. Monostori, L., Kádár, B., Bauernhansl, T., Kondoh, S., Kumara, S., Reinhart, G., Sauer, O., Schuh, G., Sihn, W., Ueda, K., (2016), Cyber-physical systems in manufacturing, Cirp Annals, 65(2), 621-641.
6. Frazzon, E. M., Agostino, Í. R. S., Broda, E., Freitag, M., (2020), Manufacturing networks in the era of digital production and operations: A socio-cyber-physical perspective, Annual Reviews in Control, 49, 288-294.
7. Dhanda, M., Rogers, B. A., Hall, S., Dekoninck, E., Dhokia, V., (2025), Reviewing human-robot collaboration in manufacturing: Opportunities and challenges in the context of industry 5.0, Robotics and Computer-Integrated Manufacturing, 93, 102937.
8. Tóth, A., Nagy, L., Kennedy, R., Bohuš, B., Abonyi, J., Ruppert, T., (2023), The human-centric industry 5.0 collaboration architecture, MethodsX, 11, 102260.
9. De Simone, V., Di Pasquale, V., Miranda, S., (2023), An overview on the use of AI/ML in manufacturing MSMEs: Solved issues, limits, and challenges, Procedia Computer Science, 217, 1820-1829.
10. Haridasan, P. K., Jawale, H., (2024), Generative Ai in manufacturing: A review of innovations, challenges and future prospects, Journal of Artificial Intelligence, Machine Learning and Data Science, 2(2), 1418-1424.
11. LLewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D., (2020), Retrieval-augmented generation for knowledge-intensive NLP tasks, Advances in Neural Information Processing Systems, 33, 9459-9474.
12. Kumar, P., (2024), Large language models (llms): Survey, technical frameworks, and future challenges, Artificial Intelligence Review, 57(10), 260.
13. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S., (2019), Parameter-efficient transfer learning for NLP, Proceedings of the International Conference on Machine Learning, 2790-2799.

14. Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, h., Zhang, X., Zhao, T., Panalkar, A., Mehta, D., Pasquali, S., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., White, C., Gu, Q., Pei, J., Yang, C., Zhao, L., (2025), Domain specialization as the key to make large language models disruptive: A comprehensive survey, ACM Computing Surveys, 58(3), 1-39.

15. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., Polosukhin, I., (2017), Attention is all you need, Advances in Neural Information Processing Systems, 30.

16. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G., (2023), Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971.

17. Smith, S., Patwary, M., Norick, B., LeGresley, P., Rajbhandari, S., Casper, J., Liu, Z., Prabhumoye, S., Zerveas, G., Korthikanti, V., Zhang, E., Child, R., Aminabadi, R., Y., Bernauer, J., Song, X., Shoeybi, M., He, Y., Houston, M., Tiwary, S., Catanzaro, B., (2022), Using deepspeed and megatron to train Megatron-Turing NLG 530B, a large-scale generative language model, arXiv preprint arXiv:2201.11990.

18. World Economic Forum, (2024), Why large language models (LLMs) are the future of manufacturing. https://www.weforum.org/stories/2024/04/why-large-language-models-are-so-important-for-the-future-of-the-manufacturing-industry/

19. Li, Y., Zhao, H., Jiang, H., Pan, Y., Liu, Z., Wu, Z., Shu, P., Tian, J., Yang, T., Xu, S., Lyu, Y., Blenk, P., Pence, J., Rupram, J., Banu, E., Liu, N., Wang, L., Song, W., Zhai, X., Song, K., Zhu, D., Li, B., Wang, X., Liu, T., (2024), Large language models for manufacturing, arXiv preprint arXiv:2410.21418.

20. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W., (2022), Lora: Low-rank adaptation of large language models, ICLR, 1(2), 3.

21. Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., (2023), Qlora: Efficient finetuning of quantized LLM, Advances in Neural Information Processing Systems, 36, 10088-10115.

22. Lu, W., Luu, R. K., Buehler, M. J., (2025), Fine-tuning large language models for domain adaptation: Exploration of training strategies, scaling, model merging and synergistic capabilities, npj Computational Materials, 11(1), 84.

23. Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., Finn, C., (2023), Direct preference optimization: Your language model is secretly a reward model, Advances in Neural Information Processing Systems, 36, 53728-53741.

24. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., Iwasawa, Y., (2022), Large language models are zero-shot reasoners, Advances in Neural Information Processing Systems, 35, 22199-22213.

25. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., (2020), Language models are few-shot learners, Advances in Neural Information Processing Systems, 33, 1877-1901.

26. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., (2022), Chain-of-thought prompting elicits reasoning in large language models, Advances in Neural Information Processing Systems, 35, 24824-24837.

27. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K., (2023), Tree of thoughts: Deliberate problem solving with large language models, Advances in Neural Information Processing Systems, 36, 11809-11822.

28. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R., (2022), Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems, 35, 27730-27744

29. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, YI., Sun, J., Wang, H., (2023), Retrieval-augmented generation for large language models: A survey, arXiv preprint arXiv:2312.10997.

30. Karpukhin, V., Oguz, B., Min, S., Lewis, P. S., Wu, L., Edunov, S., Chen, D., Yih, W.-t., (2020), Dense passage retrieval for open-domain question answering, arXiv:2004.04906.

31. Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H., (2024), Self-rag: Learning to retrieve, generate, and critique through self-reflection, arXiv:2310.11511.

32. Yan, S. Q., Gu, J. C., Zhu, Y., Ling, Z. H., (2024), Corrective retrieval augmented generation, arXiv:2401.15884.

33. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M., (2020), Retrieval augmented language model pre-training, Proceedings of the International Conference on Machine Learning, 3929-3938.

34. Izacard, G., Grave, E., (2021), Leveraging passage retrieval with generative models for open domain question answering, Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, 874-880.

35. Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., De Las Casas, D., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J., Elsen, E., Sifre, L., (2022), Improving language models by retrieving from trillions of tokens, Proceedings of the International Conference on Machine Learning, 2206-2240.

36. Wan, Y., Chen, Z., Liu, Y., Chen, C., Packianather, M., (2025), Empowering LLM by hybrid retrieval-augmented generation for domain-centric QA in smart manufacturing, Advanced Engineering Informatics, 65, 103212.

37. Buehler, M. J., (2024), Generative retrieval-augmented ontologic graph and multiagent strategies for interpretive large language model-based materials design, ACS Engineering Au, 4(2), 241-277.

38. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Metropolitansky, D., Ness, R., Larson, J., (2024), From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130.

39. De Santis, A., Balduini, M., De Santis, F., Proia, A., Leo, A., Brambilla, M., Della Valle, E., (2024), Integrating large language models and knowledge graphs for extraction and validation of textual test data, Proceedings of the International Semantic Web Conference, 304-323.

40. Xu, Y., He, S., Chen, J., Wang, Z., Song, Y., Tong, H., Liu, G., Liu, K., Zhao, J., (2024), Generate-on-graph: Treat llm as both agent and kg in incomplete knowledge graph question answering, arXiv preprint arXiv:2404.14741.

41. Wang, P., Karigiannis, J., Gao, R. X., (2024), Ontology-integrated tuning of large language model for intelligent maintenance, CIRP Annals, 73(1), 361-364.

42. Deng, H., Namoano, B., Zheng, B., Khan, S., Erkoyuncu, J. A., (2024), From prediction to prescription: Large language model agent for context-aware maintenance decision support, Proceedings of the PHM Society European Conference, 10-10.

43. Wang, H., Li, Y.-F., (2023), Large language model empowered by domain-specific knowledge base for industrial equipment operation and maintenance, Proceedings of the 5th International Conference on System Reliability and Safety Engineering, 474-479.

44. Jones, G., Williams, J., Berg, T., Stowe, A., Birt, L., Li, X., Generative large language models for predictive maintenance planning, Available at SSRN 5800812.

45. Badini, S., Regondi, S., Frontoni, E., Pugliese, R., (2023), Assessing the capabilities of ChatGPT to improve additive manufacturing troubleshooting, Advanced Industrial and Engineering Polymer Research, 6(3), 278-287.

46. Fan, H., Fuh, J., Lu, W. F., Kumar, A. S., Li, B., (2024), Unleashing the potential of large language models for knowledge augmentation: A practical experiment on incremental sheet forming, Procedia Computer Science, 232, 1269-1278.

47. Liu, X., Erkoyuncu, J. A., Fuh, J. Y. H., Lu, W. F., Li, B., (2025), Knowledge extraction for additive manufacturing process via named entity recognition with LLMs, Robotics and Computer-Integrated Manufacturing, 93, 102900.

48. Kernan Freire, S., Wang, C., Foosherian, M., Wellsandt, S., Ruiz-Arenas, S., Niforatos, E., (2024), Knowledge sharing in manufacturing using LLM-powered tools: User study and model benchmarking, Frontiers in Artificial Intelligence, 7, 1293084.

49. Sobhan, S., Haque, M. A., (2025), LLM-assisted question-answering on technical documents using structured data-aware retrieval augmented generation. arXiv preprint arXiv:2506.23136.

50. Trajanoska, M., Stojanov, R., Trajanov, D., (2023), Enhancing knowledge graph construction using large language models, arXiv preprint arXiv:2305.04676.

51. Wahid, A., Yahya, M., Zaman, F., Zhou, B., Breslin, J. G., Ali, M. I., Kharlamov, E., (2024), Integrating I4.0 knowledge graphs with large language models beyond SPARQL endpoints, Proceedings of the SOFLIM2KG-SemIIM.

52. Liu, R., Ren, H., Ren, H., Rui, W., Cui, W., Liang, X., Yang, C., Gui, W., (2025), Knowledgess, Engineering.

Ye Jin Lee
Ph.Dc. candidate in the Department of Mechanical Engineering Tech Univer-sity of Korea (TU Korea). Her research interest is robot engineering, digital twin and AI for smart manufacturing.
E-mail: yeajin719@tukorea.ac.kr

Dong Chan Kim
Assistant Professor in the Department of Mechanical Engineering Tech University of Korea (TU Korea). His research interest include intelligent manufacturing, digital twin, robotic machining processes, and AI based monitoring system.
E-mail: dckim@tukorea.ac.kr