

석사학위논문

초거대 AI 언어 모델을 활용한
생산설비 관리체계 향상방안 실증 연구

- 자동차부품 제조기업 중심으로 -

2024년

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합기술컨설팅전공

김 현 진

석사학위논문
지도교수 원종혁

초거대 AI 언어 모델을 활용한 생산설비 관리체계 향상방안 실증 연구

- 자동차부품 제조기업 중심으로 -

Empirical Study on How to Improve Production
Machine Management System Using Large AI
Language Model

- Focusing on automotive parts manufacturing -

2024년 6월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합기술컨설팅전공

김 현 진

석사학위논문
지도교수 원종혁

초거대 AI 언어 모델을 활용한 생산설비 관리체계 향상방안 실증 연구

- 자동차부품 제조기업 중심으로 -

Empirical Study on How to Improve Production
Machine Management System Using Large AI
Language Model

- Focusing on automotive parts manufacturing -

위 논문을 컨설팅학 석사학위 논문으로 제출함

2024년 6월 일

한성대학교 지식서비스&컨설팅대학원

스마트융합컨설팅학과

스마트융합기술컨설팅전공

김 현 진

김현진의 컨설팅학 석사학위 논문을 인준함

2024년 6월 일

심사위원장 정 진택 (인)

심 사 위 원 주 형근 (인)

심 사 위 원 원 종혁 (인)

국 문 초 록

초거대 AI 언어 모델을 활용한 생산설비 관리체계 향상방안 실증 연구

- 자동차부품 제조기업 중심으로 -

한성대학교 지식서비스&컨설팅대학원

스 마 트 융 합 컨 설 팅 학 과

스 마 트 융 합 기 술 컨 설 팅 전 공

김 현 진

현대사회에서 제조기업의 생산성 향상을 위한 생산설비에 대한 보존, 운영 및 유지보수의 효율성이 점점 더 중요해지고 있다. 생산설비의 특성상 연중무휴 동작하기 때문에 주요 부품의 마모, 훼손, 노후화로 인한 오동작 등과 같은 이상 증상이 비정기적으로 발생하게 된다. 이러한 이상 증상이 발생하게 되면 생산설비 가동이 중단되어 생산량 감소로 직결되기 때문에 중소기업 경영에 심각한 문제가 발생하게 된다. 이러한 문제를 해결하기 위해 스마트 공장 도입, 예지보전의 적용과 같은 다양한 개선 조치가 추진되고 있다. 그러나 이러한 시스템과 관련된 높은 비용은 실제 적용에 큰 어려움을 안겨준다.

이처럼 생산설비의 고장으로 인해 생산이 중단되는 문제를 해결하기 위한 연구인 사후보전 조치 방안이나 설비관리 성과지표 개선 방안에 대한 연구가 많이 이루어지지 않고 있다. 그 이유로는 생산설비의 고장뿐만 아니라 성능 수준에 미달하는 규격의 설비로 인해 대체 설비로 전환하거나 복구하는 등

예상치 못한 특수한 상황들로 인해 연구의 어려움이 있다. 여기에는 고장 원인 파악, 필요한 부품 조달, 전문 엔지니어 배치, 설비의 비가동 시간 동안 연계된 시설이나 설비에 미치는 영향 관리 등의 복잡한 문제가 있다. 실제로 사후보전 활동은 체계적으로 관리되기보다는 유지보수 전담 인력의 개인 업무 역량에 많이 의존하고 있다. 설비종합효율(OEE, Overall Equipment Effectiveness)을 평가할 때 예방보전이 강조되지만 실제로는 대부분의 시간이 사후보전 활동에 투자되고 있다. 결과적으로 많은 제조업체는 생산설비의 생존 가능성을 높이기 위해 높은 수준의 유지관리 매뉴얼을 구축하게 되었다. 이러한 복잡성으로 인해 사후보전 관행을 기반으로 생산성이나 생존 가능성을 향상시키기 위한 연구를 수행하는 것이 어려운 실정이다.

본 연구에서는 AI 기술을 활용하여 중소기업의 설비보전 활동 중 사후보전의 효율성을 향상시키는 방안을 제시하고 있다. 본 연구는 고장 원인, 고장 진단, 환경 요인 및 대응 시간을 포함한 설비보전 이력 데이터를 기반으로 중소기업의 지식 공유 문제를 해결하기 위해 초거대 언어 모델(LLM, Large Language Model)을 활용하였다. 이를 통해 설비종합효율을 향상시키고, 중소기업이 직면한 실질적인 문제에 대한 효과적인 해법을 제공하여 제조 산업의 생산성과 경쟁력을 높이는 것을 목표로 한다.

연구 방법으로는 피실험 기업 생산설비의 사후보전 데이터를 수집하고 빈도분석 및 표준통계 평가를 수행했다. 예비 실험에서는 스마트폰 매뉴얼의 특정 내용을 질의하고 관련 응답을 받는 방식으로 초거대 언어 모델 기반 시스템의 타당성을 검증했다. 이번 연구 실험에서는 각 고장 구성 요소에 대한 장비 고장 증상, 고장 유형, 가동 중지 시간을 종합적으로 조사하여 설비종합효율에 미치는 영향을 분석했다.

본 연구는 AI 기반의 사후보전 효율성 향상 모델을 실증연구 결과를 통해 제시함으로써 제조기업의 생산성 및 경쟁력 향상에 기여하는 것을 목표로 한다. 이는 중소기업이 직면한 실질적인 과제를 해결하고 지속 가능한 산업 성장을 촉진하는데 중요한 역할을 할 것이다.

【주요어】 설비보전, 사후보전, 긴급보전, LLM, AI, RAG

목 차

I. 서 론	1
1.1 연구의 배경 및 목적	1
1.2 연구의 방법 및 구성	6
II. 이론적 배경	7
2.1 인공지능(AI) 정의	7
2.2 초거대 언어 모델(LLM) 정의	18
2.2.1 언어모델(LM, Language Model) 정의	18
2.2.2 초거대 언어 모델(LLM) 정의	26
2.2.3 검색 증강 생성(RAG) 정의	30
2.3 설비보전의 정의	34
2.3.1 설비보전의 역사	36
2.3.2 예방보전 정의	38
2.3.3 사후보전 정의	41
2.3.4 개량보전 정의	43
2.3.5 긴급보전 정의	45
2.4 설비종합효율(OEE) 정의	47
2.4.1 설비효율화	49
2.4.2 설비종합효율 측정	50
2.4.3 설비종합관리(TPM) 측정 방안	53
2.5 선행연구 분석	54
2.5.1 선행연구 분석결과	54
2.5.2 설비보전 개선 방안 연구	59
III. 연구설계 및 구현	65
3.1 실험 환경 및 구성	65
3.1.1 연구설계	65

3.1.2 실험 환경 및 구성	72
3.1.3 Llama 3	73
3.1.4 시스템의 구성	74
3.1.5 코드의 구현	75
3.1.6 실험 데이터 생성	78
IV. 연구 결과	79
4.1 실험 결과	79
4.2 결과 분석	79
V. 결 론	83
5.1 결론 및 시사점	83
5.2 연구의 한계점 및 향후 연구방향	85
참 고 문 헌	86
ABSTRACT	96

표 목 차

[표 2-1] 글로벌 수준의 설비종합효율 측정 기준	50
[표 2-2] 설비보전 선행연구 분석	54
[표 3-1] 2023년도 A사 사후보전 이력 현황	66
[표 3-2] 2023년도 유형별 고장 비율	67
[표 3-3] 2023년도 유형별 누적 비가동 시간 비율	68
[표 3-4] 동일한 문제에 대한 보전 이력	69
[표 3-5] 실험 시스템의 구성	72
[표 4-1] 시스템 적용 전과 적용 후 비교	80
[표 4-2] A사의 고장 유형별 평균 비가동 시간 증감률	80

그 립 목 차

[그림 1-1] 고장 증가 추세 비교	2
[그림 1-2] 돌발고장 증가 추세 비교	2
[그림 1-3] 설비관리 담당 인재 육성 과제 분석	3
[그림 1-4] 산업 구분 및 규모에 따른 연도별 미충원율	4
[그림 2-1] 인공지능 붐의 역사	7
[그림 2-2] 인공 신경망 뉴런의 다양한 연결 패턴	8
[그림 2-3] 퍼셉트론의 구성	9
[그림 2-4] 동기식 반복 네트워크와 이에 상응하는 계층화된 네트워크	10
[그림 2-5] CNN의 아키텍처	10
[그림 2-6] Atari의 게임 스페이스 인베이터의 상태에 대해 DQN이 예측한 상태 값	11
[그림 2-7] 트랜스포머 모델의 아키텍처	12
[그림 2-8] 언어 모델의 메타학습	13
[그림 2-9] BERT의 사전 학습 및 미세 조정 절차	13
[그림 2-10] HMM 단어 인식기 다이어그램	20
[그림 2-11] Feed-forward Neural Network의 구조	21
[그림 2-12] RNN 언어 모델의 구조	22
[그림 2-13] 메모리 셀의 아키텍처	22
[그림 2-14] 트랜스포머 모델의 구조	24
[그림 2-15] LLM & Generative AI Relation Diagram	26
[그림 2-16] LLM의 Fine-tuning 단계	28
[그림 2-17] RAG 연구의 기술 트리	30
[그림 2-18] LLM을 활용한 RAG 구조도	31
[그림 2-19] 질의 응답에 적용된 검색 증강 생성(RAG)의 대표적인 사례	32
[그림 2-20] 설비보전의 종류	35
[그림 2-21] 상태기준보전(CBM)의 실시 과정	39
[그림 2-22] 제조설비 보전업무 흐름도	42
[그림 2-23] 질의 및 RAG 내부 검색 결과	60
[그림 2-24] 초거대 언어 모델을 통한 답변 처리	61

[그림 2-25] 매뉴얼의 실제 내용	62
[그림 3-1] A사의 사후보전 처리 현황	65
[그림 3-2] 2023년도 유형별 고장 비율	67
[그림 3-3] 2023년도 유형별 누적 비가동 시간 비율	68
[그림 3-4] RAG기반 생성형 AI 서비스 구현 플랫폼	76
[그림 3-5] 학습용 시나리오	78
[그림 3-6] 표현에 따른 추가 시나리오	78

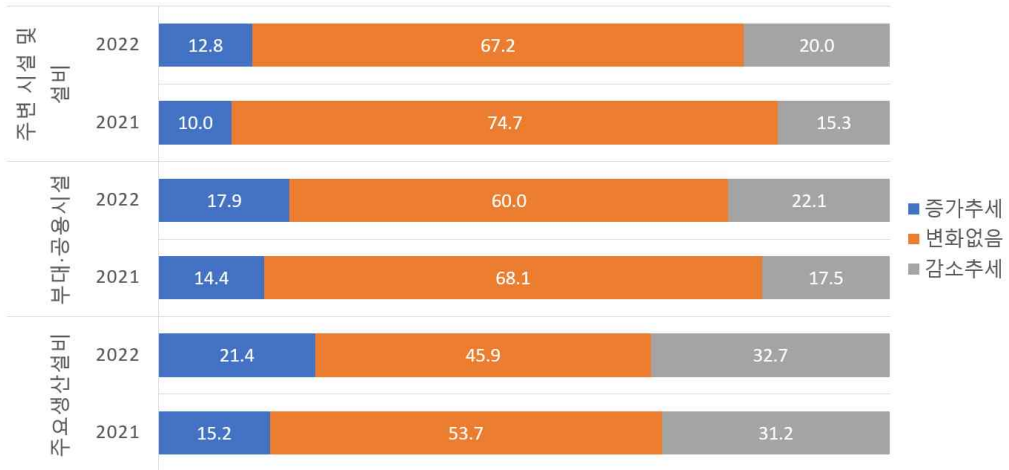
I. 서론

1.1 연구의 배경 및 목적

현대 제조업 분야는 원자재 및 인건비 상승과 유통비용의 증가로 인한 생산성 저하로 인해 글로벌 불황이 지속되면서 경영 위기를 겪고 있다. 우리나라의 경우 지정학적 리스크와 환율 상승으로 인해 이러한 어려움이 더욱 가중되어 경쟁력이 저하되고 있다. 이러한 문제를 해결하기 위해 스마트 팩토리(Smart Factory)의 도입은 업계의 혁신적인 변화를 주도하고 있다. 그러나 시스템 결함, 불안정성, 사업개선 실효성에 대한 불신, 유지관리의 어려움 등 주요 장애 요인이 이러한 시스템의 활성화를 계속 가로막고 있다(통계청, 2023).

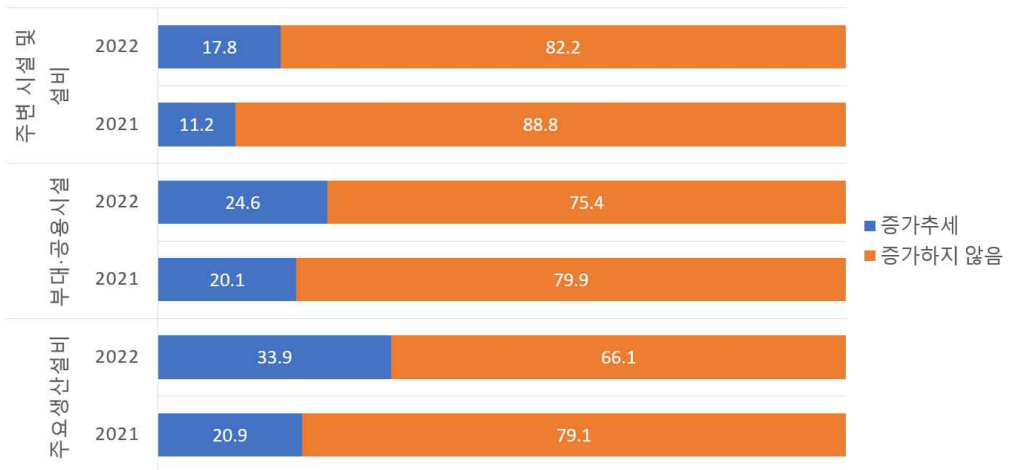
제조업은 4M(Man, Machine, Material, Method)과 1E(Environment)의 자원 활용에 따라 생산이 결정된다. 이 중 Machine으로 지칭되는 생산 설비는 조립 중심 제조에 비해 장치 중심 제조에서 더 큰 영향력을 갖는다(함효준, 2005). 주조 또는 사출과 같은 설비의 장시간 가동 중단은 생산성에 치명적인 영향을 미칠 수 있다(이명수, 2017). 결과적으로 설비 고장으로 인한 생산 중단은 제조기업의 경영 상태에 직접적으로 영향을 미치게 된다.

일본 플랜트 메인テナンス 협회의 2022년도 유지보수 실태조사 보고서에 따르면 생산 중단이 경영 상태에 미치는 영향의 심각성에 대한 중요성을 강조하고 있다. [그림 1-1]과 같이 주요 생산설비의 고장 증가 추세에 대해 ‘변함 없음’ 또는 ‘증가 추세’라고 응답한 비율이 68.5%에 달했다. 이는 많은 기업들이 여전히 설비 고장 문제에 효과적으로 대응하지 못하고 있음을 보여주고 있다((사) 일본 플랜트메인テナンス협회, 2023).



[그림 1-1] 고장 증가 추세 비교 ((사)일본플랜트메인テナンス협회, 2023)

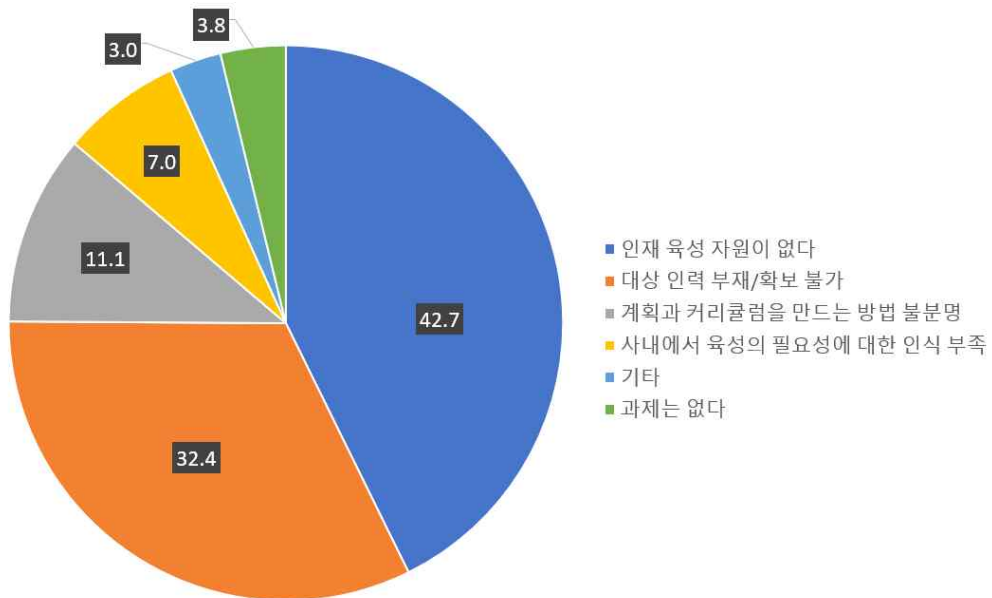
또한 [그림 1-2]와 같이 돌발 고장 증가 추세에서 ‘증가한다’고 응답한 비율이 20.9%에서 2022년에는 33.9%로 증가하였다.



[그림 1-2] 돌발고장 증가 추세 비교 ((사)일본플랜트메인テナンス협회, 2023)

그리고 [그림 1-3]과 같이 설비 고장 문제의 원인을 분석하는 데에도 여러 가지 어려움이 있는 것으로 나타났다. 응답 기업의 80.9%가 '원인 분석을 할 수 있는 사람의 능력이 따라가지 못함'을 주요 문제점으로 지적했으며,

53.5%는 '분석 기술 부족'을, 37.5%는 '범용 기술 정보 부족'을 언급했다. 이는 설비 고장의 원인을 정확히 파악하고 적절한 대응을 하는 데 있어 인력과 기술적 측면에서 문제가 있음을 보여준다.



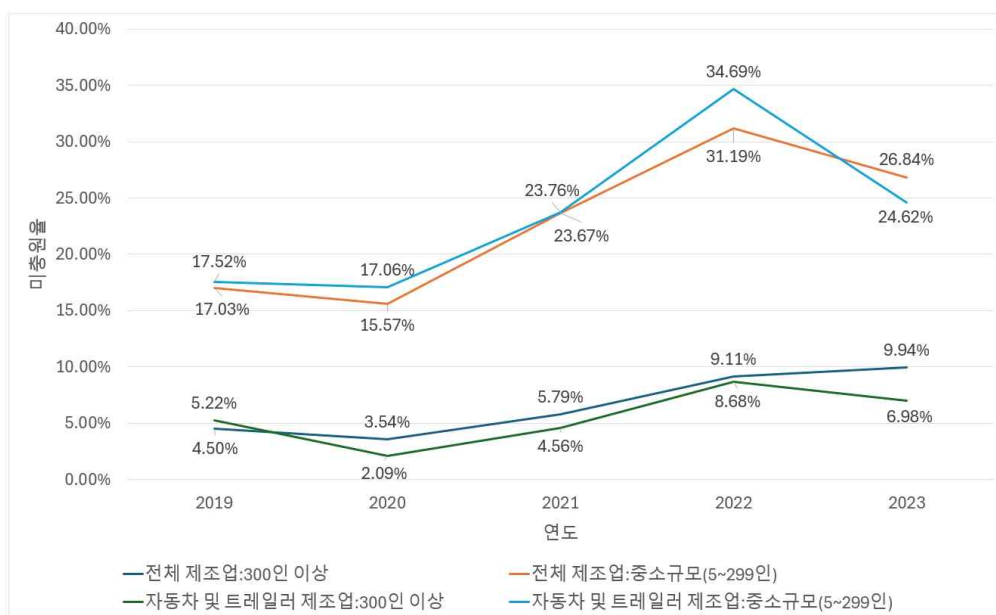
[그림 1-3] 설비관리 담당 인재 육성 과제 분석

((사)일본플랜트메인터넌스협회, 2023)

한편 설비보전에 관한 대부분의 연구는 스마트 팩토리의 핵심 기능인 장비 데이터를 분석하여 이상 징후를 사전에 식별하고 해결하는 예지보전을 강조하고 있다(강성우, 2019). 그리고 생산설비의 고장으로 인한 생산 중단 문제를 해결하기 위한 사후보전 조치 방안이나 설비관리 성과지표 개선 방안에 관한 연구는 상대적으로 부족한 실정이다. 그러나 예지보전 시스템을 구현하는 데 비용이 많이 들 수 있으며, 노후 된 설비를 계속 사용하는 중소기업의 경우 더욱 그렇다.

또한 중소기업은 숙련된 기술자의 경험과 지식의 비효과적인 공유로 인해 낙후된 설비관리 관행, 잦은 고장, 반복되는 문제로 어려움을 겪는 경우가 많다(김봉윤, 2008). 이러한 문제는 시간 부족, 과도한 업무량, 제한된 리소스,

인력의 노령화 등의 요인으로 인해 더욱 악화된다(Samuel Sam Santhose, 2023). 아래 [그림 1-4]와 같이 중소기업에서는 앞에서 언급한 문제를 해결하기 위한 숙련된 인력을 고용하는 데 상당한 어려움을 겪고 있다. 따라서 생산설비 고장 시 정확한 진단을 통한 개선된 사후 유지관리 조치가 시급히 필요하다.



[그림 1-4] 산업 구분 및 규모에 따른 연도별 미충원율¹⁾ (통계청, 2023)

본 연구는 제조업 분야 중소기업이 직면한 현실적 문제를 바탕으로 효율적인 설비보전 방안을 제안하는 것을 목적으로 한다. 최근 사회는 AI를 활용하여 경제적 가치를 창출하고, 복잡한 문제를 해결하며, 생산성과 효율성을 높이고, 비용을 절감하고 있다. 중국 A주 상장기업의 데이터를 활용해 15년간(2006~2020) AI 기술혁신이 기업 생산성에 미치는 영향을 실증 연구한 결과, AI 기술혁신이 중국 제조업체의 총요소생산성(TFP, Total Factory productivity)²⁾에 긍정적인 영향을 미치는 것으로 나타났다. AI 기술혁신은

1) 미충원율이란 구인 대비 미고용 비율을 말한다.

2) 총요소생산성이란 경제성장이라는 현상을 설명하기 위하여 그것을 두 생산요소인 자본과 노동이 행한 기여도를 척도로 분석함에 있어서 이 둘의 기여로는 설명되지 않는 나머지를 즉 잔차(residual) 부분을 의미한다.

운영 비용 절감, 고숙련 노동력 비율 증가, 디지털 전환 촉진을 통해 생산성을 향상시켰다(Shaoxuan Zhai, 2023). 이 연구에 따르면 AI 도입으로 시간당 처리 속도가 평균 14% 향상됐다.

이러한 배경에서 본 연구는 AI 기술, 특히 초거대 언어 모델(LLM)을 활용하여 중소기업의 설비보전 활동 중 사후보전의 효율성을 향상시키는 방안을 제시하고자 한다. 초거대 언어 모델(LLM)의 강력한 자연어 처리 능력과 광범위한 지식 기반은 설비 고장의 원인 분석과 해결책 제시에 큰 도움이 될 수 있다. 특히, 고장 원인, 고장 진단, 환경 요인, 고장 대응 시간 등 설비 유지관리 데이터를 활용함으로써 인력 부족과 기술 정보 부족 문제를 겪고 있는 중소기업에게 LLM은 풍부한 지식 자원으로 활용되어 지식 공유 문제를 해결할 수 있을 것이다. 이런 접근 방식을 통해 AI 사후보전 시스템 체계를 정립하여 설비종합효율(OEE, Overall Equipment Effectiveness)을 높임으로서 제조업의 생산성과 경쟁력을 향상시키고자 한다.

1.2 연구의 방법 및 구성

본 논문의 구성은 다음과 같다.

제1장은 서론으로서 제조업 분야의 생산설비 노후로 인한 보전의 문제점, 장애 발생 후 복구되는 시간 동안 발생하는 생산성 하락에 대한 해결 방안, 제조업 경영 상태 점검을 위한 설비종합효율(OEE) 판단, 생산설비의 경영상태 비중, 이러한 문제를 해결하기 위해 초거대 언어 모델(LLM) 기반으로 AI 설비보전 체계 모델을 개발하는 데 필요한 요구 사항을 기술하고 연구의 당위성을 설명하였다.

제2장은 본 연구의 이론적 배경으로서 인공지능(AI)의 정의와 역사, AI 기술의 적용 분야를 조사 분석하였다. 그리고 초거대 언어 모델(LLM)에 대한 정의와 적용 분야, 검색 증강 생성(RAG) 프로세스에 대한 정의를 분석하였다. 그리고 본 연구의 주제인 설비보전에 대한 정의를 통해 예방보전, 사후보전, 개량보전, 보전예방에 대한 개념을 재정립하고 설비종합효율(OEE) 관리 방안 관점에서 사후보전의 효율성 향상을 위한 요소들을 확인하였다.

제3장은 초거대 언어 모델(LLM) 설계를 위해 사후보전 매뉴얼을 AI 기술을 적용하여 설비 가동율을 향상시키고자 하였으며 설비종합효율을 높이는 측면을 고려하여 설계기준을 정의하였다.

제4장은 자동차 제조업체의 최근 1년 동안 사후보전 실측 데이터를 기반으로 LLM을 적용해 새로운 개념의 사후보전 체계와 매뉴얼을 구현하였다.

제5장은 본 연구의 결과에 대해 의미 있는 결과 도출 내용을 기술하였으며 해당 연구 결과에 대한 한계점, 향후 연구 방향을 제시하였다.

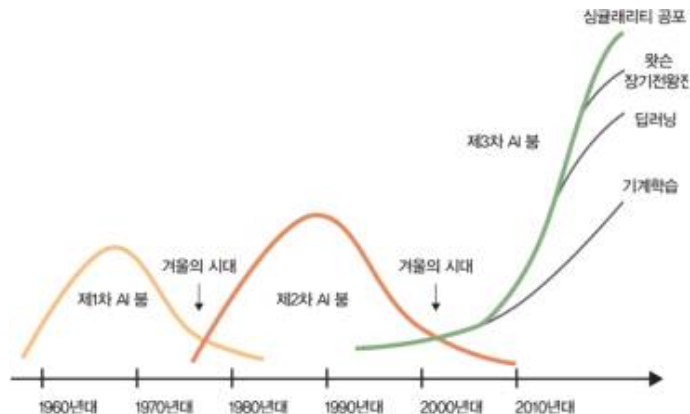
Ⅱ. 이론적 배경

2.1 인공지능(AI) 정의

2.1.1 AI 개요

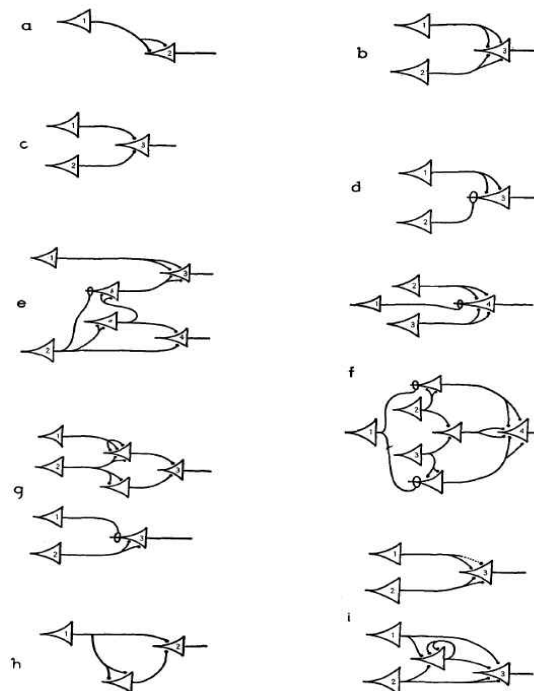
인공지능(AI, Artificial Intelligence)이란 고도의 문제 해결 능력을 가진 인공적으로 만들어진 지능을 말한다. 컴퓨터 공학에서는 이상적인 지능을 갖춘 존재, 혹은 시스템에 의해 만들어진 지능이라는 뜻으로 지칭되고 있다. 또 다른 정의는 인공적인 장치들이 가지는 지능이라는 표현으로도 사용되고 있다(기획재정부, 2024). AI는 다음과 같이 크게 네 가지 분류로 구분해 볼 수 있다. 첫 번째는 인간처럼 사고하는 시스템, 두 번째는 인간처럼 행동하는 시스템, 세 번째는 이성적으로 사고하는 시스템, 마지막으로 이성적으로 행동하는 시스템으로 분류할 수 있다.

인공지능이라는 용어는 1956년 다트머스 국제학회에서 존 매카시가 "기계를 인간 행동의 지식에서와 같이 행동하게 만드는 것"이라고 주창하면서 알려지게 되었다(McCorduck Pamela, 2004). 이는 스스로 지각력이 있으며 스스로에 대해 인식하고 스스로 판단하고 실행에 옮기는 단계까지를 AI라고 표현할 수 있다.



[그림 2-1] 인공지능 붐의 역사 (인공지능과 딥러닝, 2023)

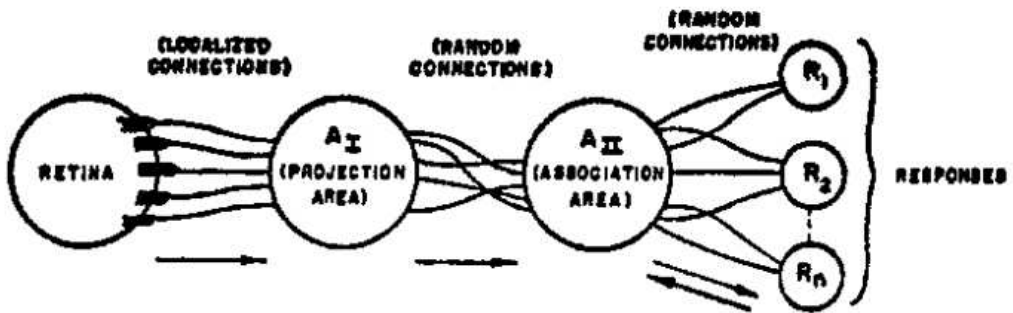
인공지능은 1940년대와 1950년대 초부터 수학, 공학, 컴퓨터 과학 분야의 과학자들이 연구해 왔다. 1943년, Warren S. McCulloch와 Walter Pitts는 신경 활동을 기반으로 한 논리적 계산을 제안했다. 그들은 신경망이 논리 회로로 작동할 수 있음을 입증하여 현대 딥러닝과 AI의 토대를 마련한 인공 신경망의 초기 개념을 확립했다(Warren S. McCulloch and Walter Pitts, 1943)(그림2). 1950년 Alan Turing은 "Computing Machinery and Intelligence"라는 논문에서 기계 지능에 대한 철학적 질문을 제기하고 이를 측정하기 위한 Turing 테스트를 제안했다(Alan Mathison Turing, 1950). [그림 2-2]와 같이 ‘심사관이 대화에서 인간과 기계를 구별할 수 없는 경우 기계는 지능을 갖춘 것으로 간주된다’는 “기계 지능” 또는 “기계학습”의 개념을 정의했으며 이는 AI 연구의 중요한 개념적 토대가 되었다.



[그림2-2] 인공 신경망 뉴런의 다양한 연결 패턴 (Warren S. McCulloch and Walter Pitts, 1943)

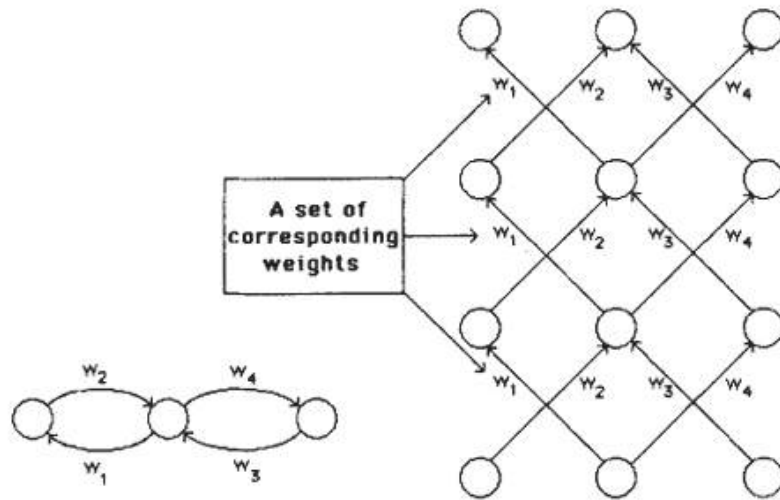
1954년 1월, 뉴욕에서는 IBM과 조지타운 대학은 러시아어 문장 60개를 영어로 번역하는 시스템을 공개적으로 시연하여 자연어 처리의 잠재력을 보여줬다(John Hutchins, 1997). 1956년 John McCarthy는 록펠러 재단이 자금을 지원한 다트머스 대학 회의에서 “인공지능”이란 용어를 “오토마타” 연구와 구별하기 위해 제안했다(McCorduck Pamela, 2004). 이 회의는 AI 연구를 독립적인 학문 분야로 확립하는 데 중요한 역할을 하게 되었다.

1958년 Frank Rosenblatt는 [그림 2-3]과 같이 비선형 분리성을 처리하는 데 한계가 있음에도 불구하고 입력 데이터부터 학습하여 패턴을 인식할 수 있는 단일 계층 신경망 모델인 퍼셉트론 개념을 정립하였다.



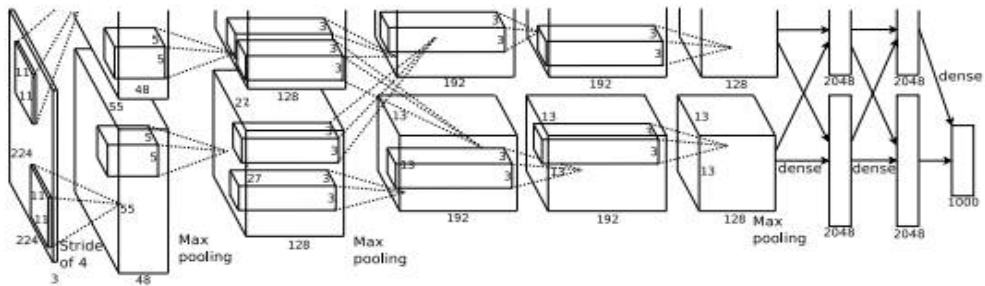
[그림2-3] 퍼셉트론의 구성 (Frank Rosenblat, 1958)

가중치 조정과 오류로부터의 학습 개념은 현대 신경망 연구의 기본 원칙으로 남아있다(Frank Rosenblat, 1958). 1986년 David Rumelhart, Geoffrey Hinton, Ronald Williams는 [그림 2-4]와 같이 다층 퍼셉트론 (MLP, Multi Layer Perceptron) 훈련을 위한 역전파 알고리즘을 도입했다. 이 알고리즘은 경사하강법을 통해 오류를 최소화하고 딥러닝 모델 개발을 크게 발전시켰으며 현대 신경망 훈련의 초석이 되었다는 평가를 받고 있다.



[그림2-4] 동기식 반복 네트워크와 이에 상응하는 계층화된 네트워크 (David E. Rumelhart et al., 1986)

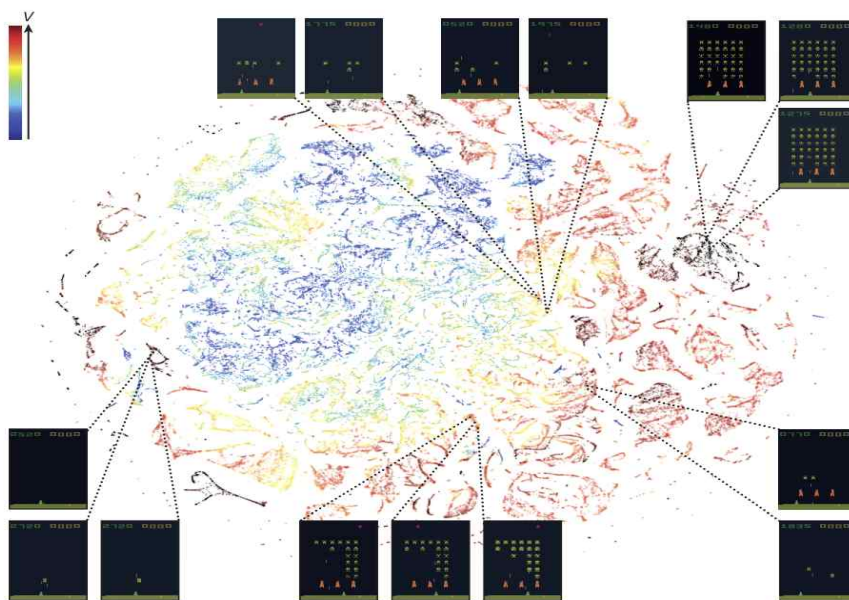
2006년 Geoffrey Hinton 등은 제한된 볼츠만 머신(RBM, Restricted Boltzmann Machines)과 심층 신뢰망(DBN, Deep Belief Networks)을 제안하여 딥러닝의 부활을 시작했다(Geoffrey Hinton et al., 2006). 딥 네트워크를 위한 이러한 빠른 학습 알고리즘은 복잡한 데이터 분포를 모델링하여 기계학습 성능을 향상시켰다.



[그림2-5] CNN의 아키텍처 (Alex Krizhevsky et al., 2012)

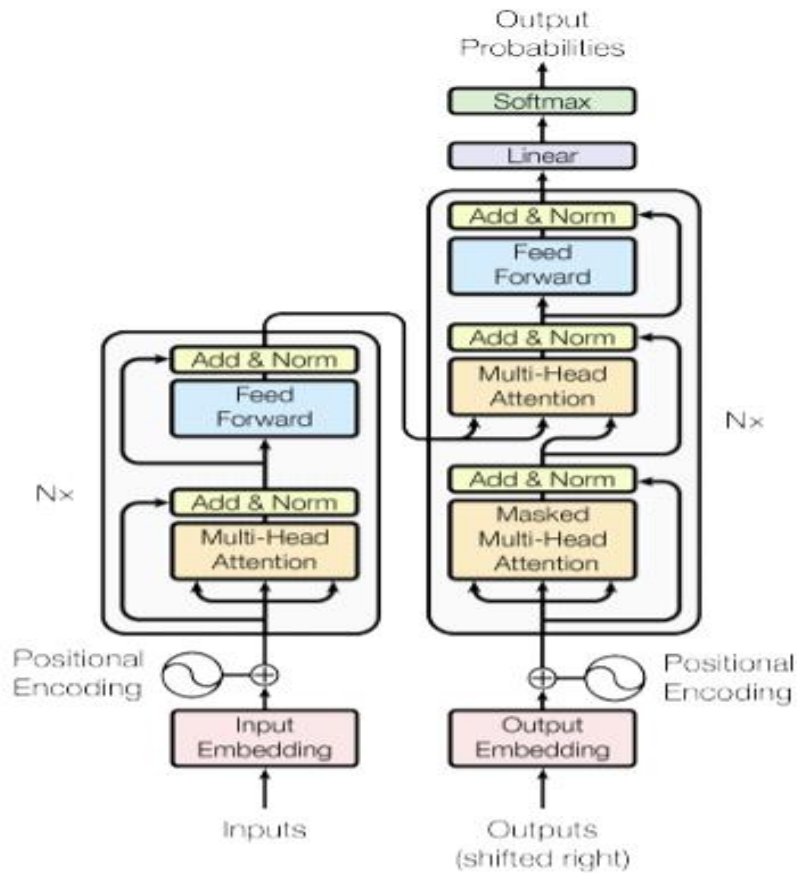
2012년 Alex Krizhevsky 등의 합성곱 신경망(CNN, Convolutional Neural Network)은 [그림 2-5]와 같이 ImageNet 대회에서 놀라운 성과를

거두며 이미지 인식 분야에서 딥러닝의 우수성을 입증했다(Alex Krizhevsky et al., 2012). 2015년 구글의 DeepMind의 연구팀은 [그림 2-6]과 같이 딥러닝과 강화학습을 결합하여 DQN(Deep Q-Network) 알고리즘을 개발하여 Atari 게임에서 인간 수준의 성능을 달성했다. 이는 게임 플레이와 같은 실시간 의사 결정 문제에서 AI의 가능성을 입증했다(Volodymyr Mnih et al., 2015).



[그림2-6] Atari의 게임 스페이스 인베이터의 상태에 대해 DQN이 예측한 상태 값 (Volodymyr Mnih et al, 2015)

2017년, Vaswani 등은 "Attention is All You Need" 논문에서 [그림 2-7]과 같이 Transformer 모델을 소개하여 병렬 처리 기능으로 자연어 처리(NLP)에 혁명을 일으켰으며(Ashish Vaswani et al., 2017) 이후 대형 언어 모델의 기초가 되었다. 트랜스포머는 순환 신경망(RNN)과는 달리 병렬 처리가 가능하여 훈련 속도를 크게 향상시켰다.



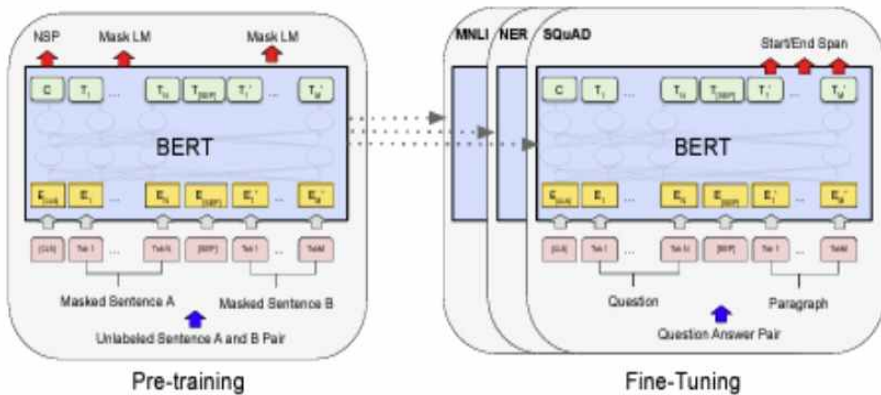
[그림2-7] 트랜스포머 모델의 아키텍처 (Ashish Vaswani et al., 2017)

2018년부터 시작된 OpenAI의 GPT(Generative Pre-trained Transformer) 시리즈는 [그림 2-8]과 같이 2020년에 출시된 1,750억 개의 매개변수를 사용한 GPT-3를 통해 다양한 자연어 처리 작업에서 초거대 언어 모델의 잠재력을 더욱 입증하였으며 문장의 생성, 번역, 요약, 질문 응답 등 여러 애플리케이션에서 최첨단 성능을 달성했다(Tom B. Brown et al, 2020).



[그림2-8] 언어 모델의 메타학습 (Tom B. Brown et al, 2020)

또한 Google의 BERT(Bidirection Encoder Representations from Transformers) 모델은 [그림 2-9]와 같이 NLP를 위한 새로운 사전 학습 기술을 도입하여 양방향의 컨텍스트를 고려하여 다양한 언어 이해 작업의 성능을 크게 향상시켰다(Jacob Devlin et al., 2018).



[그림2-9] BERT의 사전 학습 및 미세 조정 절차 (Jacob Devlin et al., 2018)

2.1.2 국내 AI 역사

국내에서 일반인에게 생소한 개념이었던 인공지능(AI, Artificial Intelligence)은 2016년 3월, 구글 DeepMind에서 개발한 AI 바둑 프로그램인 알파고와 프로바둑기사 이세돌과의 세기적인 대국을 통해 널리 알려지게 되었다.

국내에서는 1980년대부터 인공지능 기술 개발에 주력해 왔으며 1980년대는 주로 학계와 연구소에서 전문가 시스템과 자연어 처리(NLP)에 대한 연구가 이루어졌다. 대표적인 연구기관으로는 한국전자통신연구원이 있으며 초기 AI 연구와 개발의 중심 역할을 담당했다.

1990년대에는 정부의 적극적인 지원 아래 AI 연구가 확대되었다. 1996년 설립된 한국과학기술연구원(KIST) AI 연구소는 다양한 AI 프로젝트를 진행하며 국내 AI 기술 발전에 기여했다. 신경망, 패턴 인식, 음성 인식 등의 연구가 활발히 이루어졌다.

2000년대에는 인터넷과 모바일 기술의 발전과 함께 AI 기술은 음성 인식, 이미지 처리, 기계 번역 등에서 성과를 나타냈다. 네이버는 2000년대 중반부터 기계 번역 서비스인 ‘파파고’를 개발하여 AI 기반 번역 기술을 상용화했다. 카카오는 AI 기반의 음성 인식 서비스 ‘카카오미니’를 출시하며 음성 비서 시장에 진출했다.

2010년대에는 딥러닝과 빅데이터 기술의 도입으로 AI 연구와 응용이 본격적으로 증가했다. 주로 딥러닝을 활용한 이미지 인식, 자연어 처리, 자율주행 등의 기술이 발전하였다. 2013년부터 ETRI는 엑소브레인 프로젝트를 통해 딥러닝 기반의 자연어 처리 기술을 개발했다. 이 프로젝트에서 다양한 언어 데이터 분석과 처리에 높은 성과를 거두었다. 삼성전자와 LG전자는 딥러닝 기술을 적용한 스마트 가전제품을 개발하여 AI를 가전 분야에 적극적으로 도입했다. 삼성의 ‘빅스비’와 LG의 ‘씽큐’는 AI 기반의 음성 인식 기능을 제공하여 사용자 경험을 향상시켰다.

2019년 정부는 ‘AI 국가전략’을 발표하여 AI 기술 개발, 데이터 인프라 구축, AI 인재 양성 등을 목표로 하는 종합 계획을 제시했다. 이를 통해

AI 기술혁신을 선도하고 있다.

2.1.3 AI 적용 분야

AI 기술은 우리나라의 산업과 사회 전반에 걸쳐 폭넓게 적용되어 혁신적인 변화를 가져오고 있다. 전자상거래 분야에서는 고객 행동 데이터, 검색 이력, 구매 이력 등을 활용해 상품을 추천하고, 챗봇을 통해 실시간으로 사용자를 지원하며, 매출 예측과 공급망 관리를 통해 재고 관리를 최적화한다. 이러한 기술은 교육 분야로도 확장되어 AI 프로그램이 학생의 학습 스타일과 선호도에 따라 개인화된 콘텐츠를 제공하고, 학습 결과를 예측하며 교사의 일상적인 작업을 자동화함으로써 교육의 효율성을 크게 향상시키고 있다.

로봇 공학 분야에서는 다양한 센서의 데이터를 분석하여 산업용 로봇, 협동 로봇, 의료 로봇, 군용 로봇의 효율적이고 유연한 작동을 가능하게 한다. 이는 헬스케어 분야와도 밀접하게 연관되어 AI가 의료 데이터를 분석해 정확한 진단과 맞춤형 치료 계획을 지원하고 웨어러블 장치와 통합되어 환자의 활력 징후를 지속적으로 모니터링을 한다.

금융과 보안 분야에서 AI는 특히 중요한 역할을 한다. 금융 분야에서는 이상 징후 탐지와 데이터 분석을 통해 사기를 예방하고 효과적인 재무 계획을 수립하는 데 도움을 준다. 정보보안 분야에서는 네트워크 로그, 사용자 행동, 시스템 활동 등을 분석해 데이터 안전성을 유지하고 다양한 위협과 공격을 예방한다.

자율주행 기술에서는 센서 데이터를 실시간으로 분석하여 안전하고 효율적인 주행을 가능하게 한다. 이러한 기술은 스마트 시티 구축에도 적용되어, 도시 데이터를 수집하고 분석함으로써 교통 관리, 에너지 절약, 범죄 예방 등을 개선한다.

농업 분야에서는 작물 성장 모니터링, 해충 침입 예측, 최적 수확 시기 결정 등에 활용되며, 이는 환경 보호 분야와도 연결된다. AI 기반 모델은 대규모 기후 데이터를 분석하여 기후변화를 예측하고 자연재해에 대응하는데 도움을 준다.

엔터테인먼트 분야에서는 개인화된 콘텐츠 추천, 자동 콘텐츠 생성,

VR/AR을 통한 몰입형 경험의 제공 등에 활용되고 있다.

법률 분야에서는 대용량 법률 데이터 처리와 관련사건 및 판례 검색을 지원하며, 고객 서비스 분야에서는 AI 기반 챗봇과 음성 인식 기술을 통해 24시간 고객 지원을 제공한다.

이처럼 AI 기술은 다양한 산업 분야에서 혁신을 가져오고 있으며 과거 데이터 분석과 현재 데이터를 바탕으로 사용자에게 편리하고 다양한 서비스를 제공하고 있다. 또한, 다양한 기술과의 융합을 통해 과거에 없던 새로운 서비스를 창출해 내고 있다. AI 기술의 지속적인 발전은 향후 더 많은 산업 분야에서 혁신적인 변화를 가져올 것으로 예상된다.

2.2 초거대 언어 모델(LLM, Large Language Model) 정의

2.2.1 언어 모델(LM, Language Model) 정의

2.2.1.1 언어 모델(LM) 개요

언어 모델(Language Model)의 사전적 정의로는 인간의 언어를 이해하기 위해 단어나 토큰의 순서를 학습하여 주어진 텍스트의 다음 단어나 토큰의 확률을 예측하는 인공지능 모델로 정의할 수 있다(두산백과, 2024).

언어 모델은 자연어 처리(NLP, Natural Language Processing)에서 단어 시퀀스의 확률을 예측하는 데 사용되는 통계 도구이다. 언어 모델의 주요 목표는 다양한 단어 시퀀스의 가능성을 결정하여 인간의 언어를 이해하고 생성하는 것이다. 이를 위해 문장이나 문단 내의 문장을 기반으로 특정 단어의 확률을 계산하는 작업이 포함된다.

동작방식은 대규모 텍스트 데이터 모음에서 언어의 패턴과 구조를 학습하여 작동한다. 언어 모델의 작동 방법은 훈련을 위해 먼저 데이터를 수집하고 텍스트를 토큰³⁾의 단위로 나누는 토큰화 과정을 거치게 된다. 그리고 통계적 방법을 사용하여 단어 시퀀스의 확률을 추정한다. 초기 모델은 N-gram('N' 단어의 시퀀스)을 사용했지만 현대 모델은 신경망을 사용한다. 언어 모델에서 사용하는 모델의 유형은 이전 'n-1' 단어를 기반으로 단어의 확률을 계산하는 N-gram 모델과 순환 신경망(Recurrent Neural Network, RNN) 및 장단기 기억(LSTM, Long Short-Term Memory) 네트워크와 같은 아키텍처를 사용하여 텍스트의 더 긴 종속성을 모델링하는 신경망 모델과 주의 매커니즘을 활용하여 효율성을 위해 장거리 종속성과 병렬 처리를 하는 트랜스포머 모델 등이 있다. 이런 모델을 바탕으로 하여 이전 단어가 제공하는 문맥을 기반으로 문장의 다음 단어를 예측하고 전체 문장이나 구문에 확률을 할당하여 기계 번역 및 텍스트 생성과 같은 작업을 진행한다.

3) 단어, 하위 단어 또는 문자

2.2.1.2 언어 모델(LM)의 역사

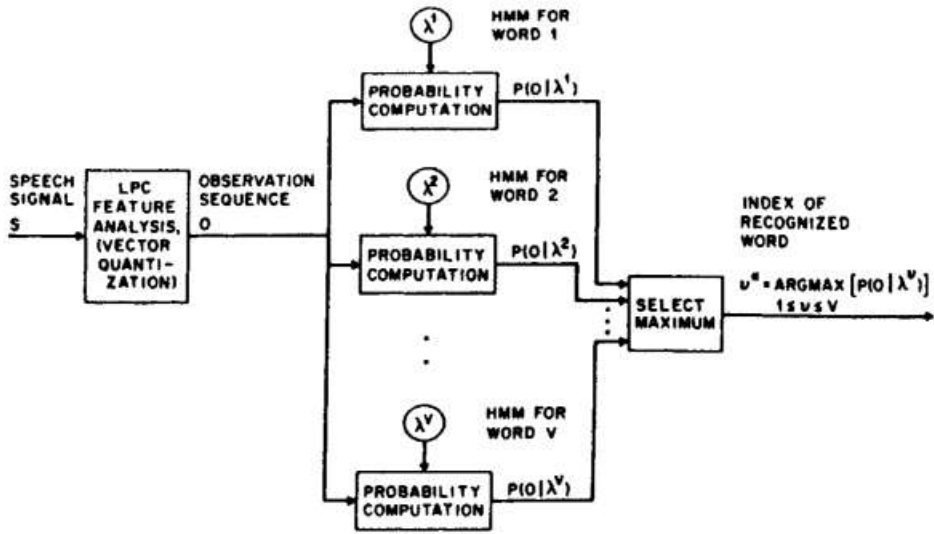
2.2.1.2.1 N-gram 모델

언어 모델의 역사는 Claude Shannon이 언어에 엔트로피 개념을 도입한 것에서부터 시작한다. Shannon은 상태 간의 전환 확률을 기반으로 하는 확률 모델인 마르코프 체인을 적용하여 문장의 다음 단어를 예측하는 데 사용했다(Claud E. Shannon, 1948). 이런 접근법은 언어의 통계적 특성을 모델링하는 데 효과적이었다. 또한 Claude E. Shannon은 통신 엔지니어링에서 정보를 다루는 방식에 대해 설명하며, 언어의 통계적 특성을 활용하는 새로운 방법으로 N-gram 모델을 제안했다(Claud Shnnon, 1951). 이 방법에서 $n-1$ 개의 연속된 단어에서 통계적으로 다음 글자를 예측하는 실험 결과에 의존하며, 초기의 언어 모델은 이런 단거리 종속성에 중점을 두고 언어를 모델링했다.

그러나 N-gram 기반 언어 모델은 언어 모델링 분야에 큰 변화를 가져왔지만 고정된 크기의 윈도우를 사용하기 때문에 윈도우 밖의 단어에 대한 정보를 포착하지 못하는 문제가 있었다.

2.2.1.2.2 통계적 언어 모델의 발전

1986년, Rabiner가 은닉 마르코프 모델(HMM, Hidden Markov Model)을 음성 인식에 어떻게 적용하는지에 대한 논문을 통해 음성 인식 시스템의 설계와 구현에 중요한 방향을 제시했다(Lawrence R. Rabiner, 1986). 아래 그림 [2-10]은 HMM 단어 인식기 다이어그램에 대한 회로도를 표현하고 있다.

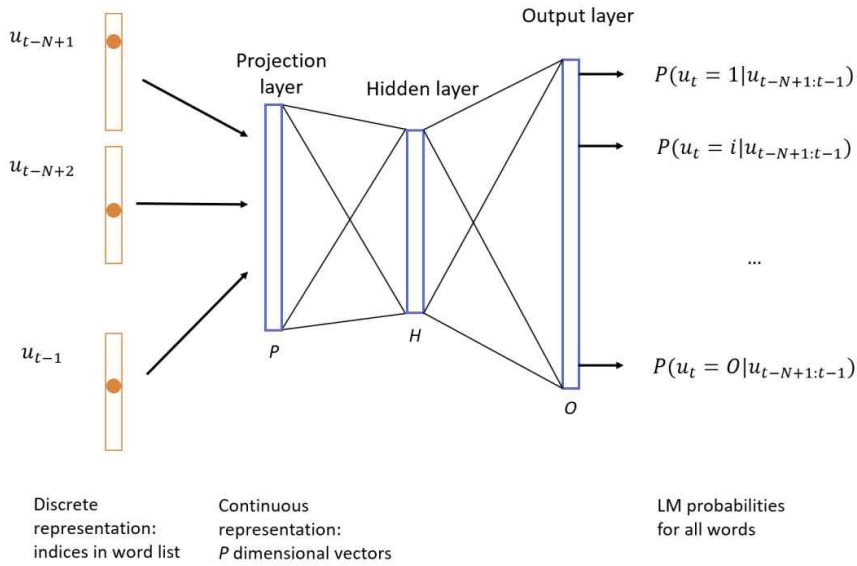


[그림 2-10] HMM 단어 인식기 다이어그램 (Leonard E. Baum, 1972)

은닉 마르코프 모델은 이후 문장의 품사를 예측하거나(Talal Almutiri, 2022) 음향 신호의 시퀀스를 단어의 시퀀스로 매핑하여 음성 인식에 널리 사용되었다(Przemyslaw Dymarkski, 2011). 그러나 상태 전이와 결과 생성이 독립적으로 이루어진다는 가정을 기반으로 하고 있어 이는 실제 언어 데이터의 복잡성을 완전히 반영하지 못하는 한계가 있었다.

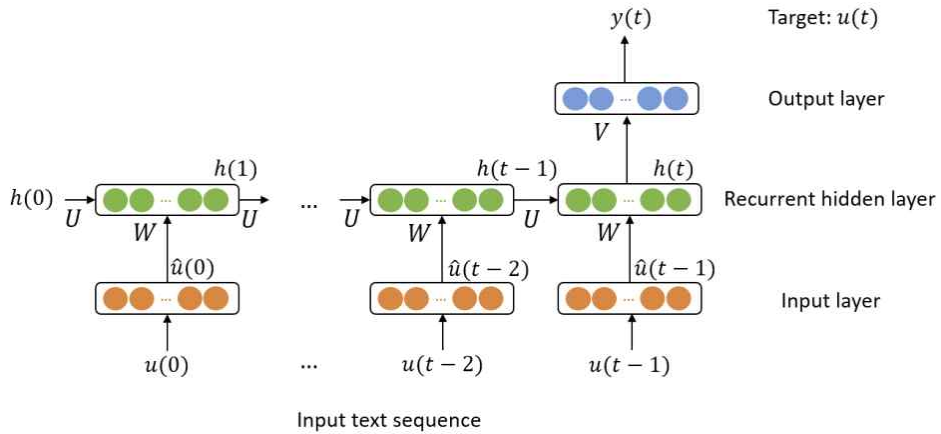
2.2.1.2.3 신경망 언어 모델 (Neural Language Models)

Justin T. Chiu와 Alexander M. Rush는 [그림 2-11]과 같이 이전 입력이 후속의 출력에 영향을 미친다는 가정을 기반으로 하는 순방향 신경망(FNN, Feed-forward Neural Network)의 도입을 제안했다(Yoshua Bengio et al., 2003).



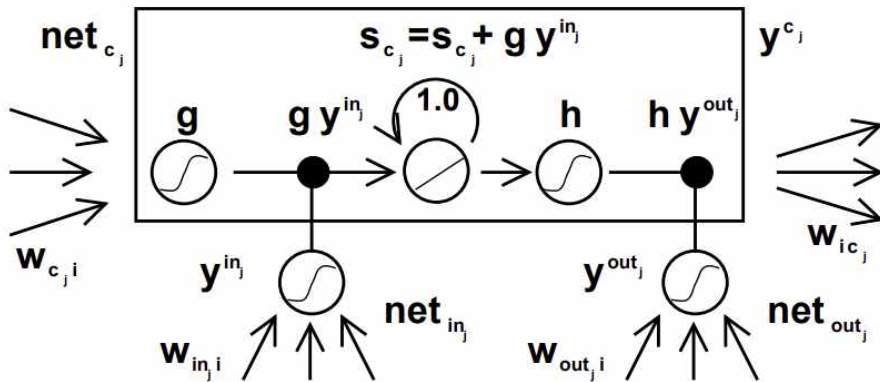
[그림 2-11] Feed-forward Neural Network의 구조 (Yoshua Bengio et al., 2003)

이 논문에서 다층 퍼셉트론(MLP, Multilayer Perceptron)을 사용해 다중 클래스 분류를 수행하고, 합성곱 신경망(CNN, Convolutional Neural Network)을 통해 창 필터의 속성을 통해 입력의 지역 패턴을 학습하도록 했다. 그러나 순방향 신경망은 입력이 출력으로 직접 흐르는 신경망 구조로 되어 있어 각 입력이 독립적으로 처리되기 때문에 순차적인 데이터를 처리하는 데 제한적이며 입력 간의 순서와 문맥을 고려하지 않게 된다. [그림 2-12]와 같이 이런 문제점을 해결하기 위해 Justin T. Chiu 등은 이전의 입력이 후속의 출력에 영향을 미친다는 가정을 기반으로 하는 순환신경망(RNN, Recurrent Neural Network)을 적용했다(Justin T. Chiu et al., 2020).



[그림 2-12] RNN 언어 모델의 구조 (Justin T. Chiu et al., 2020)

이를 통해 더 복잡한 패턴을 학습하고 긴 시퀀스에 대한 종속성을 모델링하게 되었다. [그림 2-13]과 같이 또한 순환 신경망에서도 긴 시퀀스를 처리할 때 장거리 종속성을 학습하기 어려운 문제가 있었는데 이를 보완하기 위해 메모리 셀이라는 특수 단위를 사용해 장기간 정보를 유지하는 장단기 메모리(LSTM, Long Short-Term Memory) 셀을 사용한 모델이 주목받게 되었다(Sepp Hochreiter, 1997).

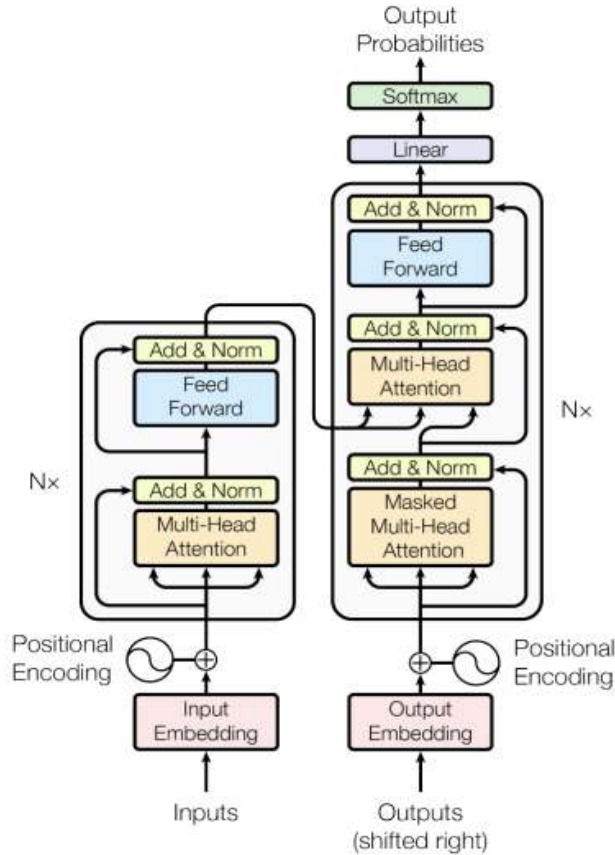


[그림 2-13] 메모리 셀의 아키텍처 (Sepp Hochreiter, 1997)

이러한 신경망 언어 모델은 딥러닝 기술을 활용해 문장의 전체적인 맥락을 고려할 수 있게 되었으며 연속적인 데이터를 처리하는 능력 덕분에 문장 내의 단어들 사이의 장기적인 의존성을 학습할 수 있게 되었다.

2.2.1.2.4 트랜스포머(Transformer)와 자연어 처리의 혁신

2017년 ‘Attention is All You Need’ 논문에서 처음 소개된 트랜스포머 아키텍처는 언어모델링과 자연어 처리 분야에 혁신을 가져왔다(Ashish Vaswani et al., 2017). [그림 2-14]와 같이 트랜스포머 모델은 순차 데이터 내의 관계를 추적하여 맥락과 의미를 학습하는 신경망으로, 주의 집중(Attention) 또는 자기 주의 집중(self-attention)이라는 수학적 기법을 활용하여 서로 떨어져 있는 데이터 요소들의 의미가 관계에 따라 미묘하게 달라지는 부분까지 감지한다. 순환신경망이나 장단기 메모리 셀과 같은 모델들은 입력이 길어질수록 이전 데이터들을 기억하기 힘들다는 문제가 있었으나 트랜스포머 모델은 이 문제를 해결했고, 이를 통해 문장의 길이에 대한 제약 없이 더욱 정교한 언어 이해 능력을 갖추게 되었다. 또한, 병렬 처리가 가능하여 대규모 데이터셋에 대한 학습 시간을 크게 줄일 수 있게 되어 초거대 언어 모델로의 발전으로 이끌었다.



[그림 2-14] 트랜스포머 모델의 구조 (Ashish Vaswani et al., 2017)

2.2.1.3 언어 모델(LM)의 적용 분야

언어 모델은 다양한 분야에서 폭넓게 적용되고 있다. 음성 인식 분야에서는 Google의 음성 검색 시스템이 언어 모델을 활용하여 음성 언어를 텍스트로 변환한다. 이 시스템은 음성 쿼리에서 가장 가능성이 높은 단어를 이해하고 예측함으로써 시끄러운 환경에서도 정확성을 향상시킨다(Alex Graves et al., 2013).

기계 번역 영역에서는 Google 번역이 언어 모델을 사용하여 번역을 예측하고 생성한다. 이 모델은 문장의 전체 맥락을 고려하여 보다 정확하고 유창한 번역을 생성할 수 있다(Yonghui Wu et al., 2016). 텍스트 생성 분야에서는 많은 고객 서비스 플랫폼이 언어 모델을 사용하여 고객 문의에

대한 자동 응답을 생성한다. 이런 시스템은 고객의 질문을 분석하고, 일반적인 문의 및 응답의 학습된 패턴을 기반으로 하여 관련 응답을 생성한다 (Oriol Vinyals et al., 2015).

자동 수정 및 예측 텍스트 기능에서는 Google의 Gboard와 같은 스마트폰 키보드 앱이 언어 모델을 사용하여 실시간 자동 수정 및 예측 텍스트 제안을 제공한다. 사용자가 입력을 하면 언어 모델은 문장의 맥락에 따라 다음 단어를 예측하고 오타를 수정한다.

감성 분석 영역에서는 Hootsuite Insights가 언어 모델을 사용하여 소셜 미디어 게시물을 분석하고, 게시물 이면의 정서를 파악한다. 이는 기업이 브랜드와 제품에 대한 대중의 인식을 이해하는 데 도움을 준다(Bing Liu, 2012).

정보 검색 분야에서 Google의 검색 엔진은 언어 모델을 사용하여 사용자의 쿼리를 해석하고 예측한다. 이 모델은 검색어 이면의 의도를 이해하기 쉽고 보다 정확하고 상황에 맞는 검색 결과를 제공하고 있다.

텍스트 요약 기능에서는 Inshort와 같은 뉴스 수집기 앱이 언어 모델을 사용하여 뉴스 기사를 간결한 주요 항목으로 요약한다. 이를 통해 사용자는 기사 전체를 읽지 않고도 기사의 주요 내용을 빠르게 파악할 수 있게 되었다.

질문 답변 시스템에서는 Amazon Alexa와 같은 가상 비서가 언어 모델을 사용하여 사용자 질문을 이해하고 응답한다. 언어 모델은 쿼리를 처리하고, 관련 정보를 검색하고, 적절한 응답을 생성한다.

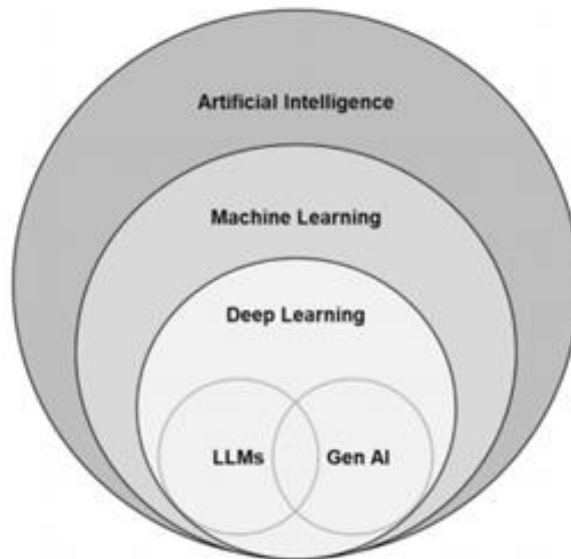
마지막으로, 자연어 이해 분야에서는 Zendesk Chat과 같이 언어 모델을 사용하여 고객 쿼리를 해석하고 관련 답변을 제공한다. 언어 모델은 챗봇이 대화의 맥락을 이해하고 적절하게 응답하도록 돕는다.

이와 같이 언어 모델은 인간 언어를 이해하고 생성하기 위한 기반을 제공함으로써 자연어 처리 분야를 크게 발전시켰다. 이는 다양한 영역에서 적용되고 있으며 음성 인식부터 감성 분석까지 애플리케이션 기능을 향상시키게 되었다. 최근에는 많은 AI 기반 시스템의 필수적인 요소로 부각되면서 점차 초거대 언어 모델로 전환하게 되는 계기가 되었다.

2.2.2 초거대 언어 모델(LLM)의 정의

초거대 언어 모델(LLM)은 인간의 언어를 처리하거나 이해하고 생성하도록 설계된 정교한 인공 지능 시스템이다. 이러한 모델은 딥 러닝 기술인 Transformer 아키텍처를 활용하여 다양한 자연어 처리(NLP) 작업을 높은 정확도로 수행한다. 초거대 언어 모델은 광범위한 데이터 세트에 대해 학습하고 수십억에서 수조 개의 매개변수를 보유하고 있어 복잡한 언어 패턴을 포착하고 일관성 있고 문맥에 적합한 텍스트를 생성할 수 있다(신중민 외, 2023).

[그림 2-15]는 초거대 언어 모델과 생성형 AI의 관계를 도식화하여 보여준다. 이 그림에서 초거대 언어 모델은 생성형 AI의 핵심 구성 요소로 작용하며, 다양한 응용 분야에서 활용되고 있음을 알 수 있다. 초거대 언어 모델은 대규모 언어 데이터를 학습하여 언어의 패턴과 구조를 이해하고, 이를 바탕으로 텍스트 생성, 번역, 요약 등 다양한 작업을 수행할 수 있다. 생성형 AI는 이러한 초거대 언어 모델의 능력을 기반으로 하여 더 복잡하고 창의적인 작업을 수행할 수 있게 된다.



[그림 2-15] LLM & Generative AI Relation Diagram (정천수, 2023)

초거대 언어 모델의 역사는 자연어 처리 기술의 발전과 밀접하게 연관되어 있다. 초기 NLP 모델은 규칙 기반 시스템과 n-그램 및 숨겨진 마르코프 모델(HMM)과 같은 통계 모델에 의존했다(Claud E. Shannon, 1948; Claud Shnnon, 1951). 2000년대 초반, 신경망과 단어 임베딩⁴⁾의 도입으로 모델이 문맥에 따라 단어의 의미적 의미를 파악할 수 있게 되면서 상당한 개선이 이루어졌다.

2017년 Vaswani 등이 "Attention is All You Need" 논문에서 Transformer 아키텍처를 소개하면서 NLP 분야에 혁명이 일어났다(Ashish Vaswani et al., 2017). 이 아키텍처는 모델이 전체 문장을 순차적으로 처리하는 것이 아니라 동시에 처리할 수 있게 했으며 이는 현대 초거대 언어 모델의 기반이 되었다.

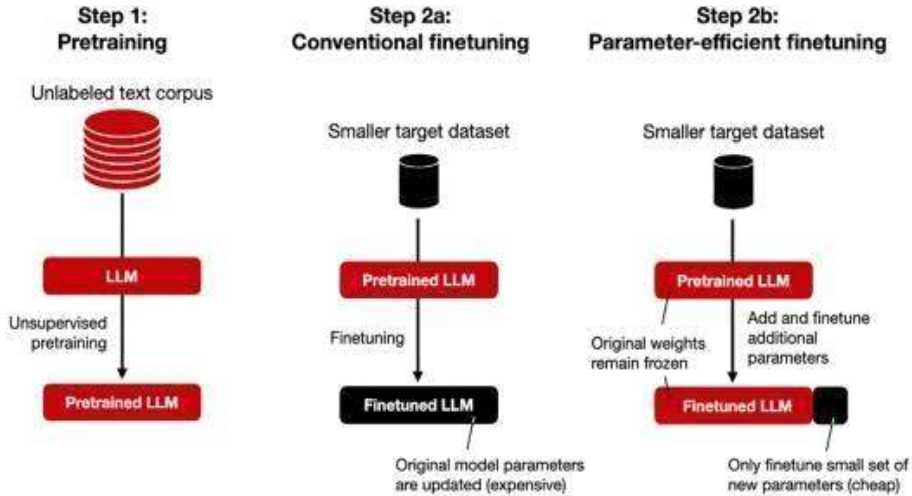
2018년 Google의 BERT(Bidirectional Encoder Representations from Transformers)는 단어의 문맥을 양방향으로 이해하는 모델을 도입하여 NLP를 크게 발전시켰다(Jacob Devlin et al., 2018). 이어 2018년부터 2020년까지 OpenAI의 GPT 시리즈, 특히 GPT-3는 최소한의 미세 조정으로 인간과 유사한 텍스트를 생성하고 다양한 NLP 작업을 수행하는 데 있어 사전 학습된 언어 모델의 힘을 보여주었다.

2020년 이후에는 GPT-4와 같은 후속 모델들이 아키텍처, 트레이닝 기법, 효율성의 개선을 통해 초거대 언어 모델 기능의 한계를 더욱 확장했다. 이러한 발전은 자연어 처리 분야에 혁명을 가져왔고 기계 번역, 콘텐츠 제작, 대화형 AI와 같은 분야에 혁신을 일으켰다(Humza Naveed et al., 2024).

초거대 언어 모델의 학습 과정은 데이터 수집, 전처리, 사전 훈련, 미세 조정, 평가 등의 단계로 이루어진다. [그림 2-16]은 이 과정 중 미세 조정 단계를 구체적으로 보여준다. 이 그림은 사전 훈련된 LLM이 특정 작업에 맞게 어떻게 적응되는지를 나타낸다. 미세 조정 과정에서는 작은 규모의 특

4) 단어를 컴퓨터가 이해하고 효율적으로 처리할 수 있도록 단어를 벡터화하는 기술로서 단어를 밀집 벡터의 형태로 표현하는 방법이다. 워드 임베딩을 거친 후 잘 표현된 단어 벡터들은 계산 처리도 가능하며 모델에 입력으로 사용할 수 있다는 특징이 있다.

정 작업 데이터셋을 사용하여 모델의 가중치를 조정한다. 이 과정을 통해 모델은 일반적인 언어 이해 능력을 유지하면서도 특정 작업에 대해 더 높은 성능을 발휘할 수 있게 된다(Humza Naveed et al., 2024).



[그림 2-16] LLM의 Fine-tuning 단계 (정천수, 2023)

초거대 언어 모델의 대표적인 예로는 OpenAI의 GPT 시리즈, Google의 BERT, Meta의 LLaMA 등이 있다. 이러한 모델들은 각각 고유한 특징과 강점을 가지고 있으며, 다양한 NLP 작업에서 뛰어난 성능을 보여주고 있다.

향후 초거대 언어 모델 연구는 모델의 효율성 향상, 더 나은 미세 조정 방법 개발, 이해력 향상, 윤리적 AI 구현, 학제 간 통합, 개인화 등의 방향으로 발전이 예상된다. 특히 모델의 크기를 줄이면서도 성능을 유지하는 모델 압축 기술, 적은 데이터로도 효과적으로 학습할 수 있는 few-shot 학습 기법, 그리고 모델의 결정 과정을 더 잘 이해하고 설명할 수 있는 설명 가능한 AI 기술 등이 중요한 연구 주제가 될 것으로 보인다. 멀티모달 학습을 통해 텍스트뿐만 아니라 이미지, 음성, 비디오 등 다양한 형태의 데이터를 처리할 수 있는 모델이 개발될 것으로 예상된다. 또한 모델의 추론 능력을 향상시키기 위해 외부 지식 베이스와의 연동을 강

화하는 연구도 진행될 것이다. 더불어 모델의 지속적인 학습과 적응을 위한 평생 학습(lifelong learning) 기법에 대한 연구도 활발히 이루어질 것으로 보인다. 이러한 연구들은 초거대 언어 모델의 능력을 더욱 확장하고, 실제 세계의 복잡한 문제를 해결하는 데 큰 도움을 줄 것으로 기대된다.

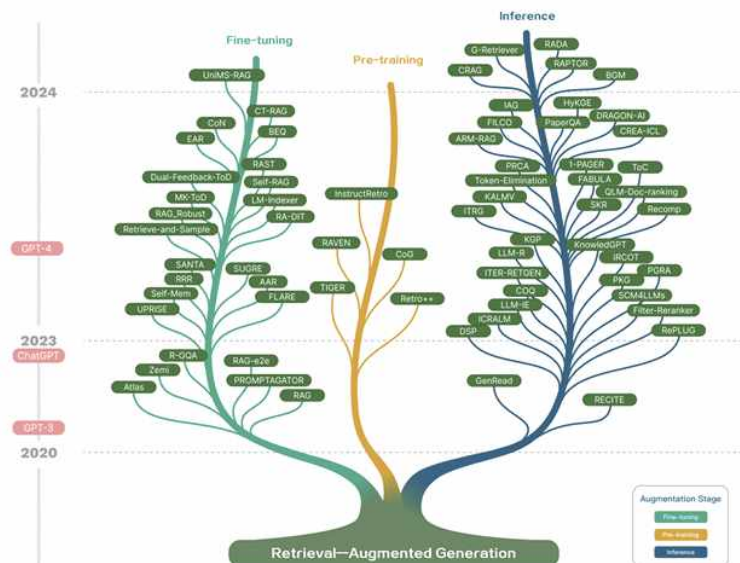
이러한 초거대 언어 모델은 고객 서비스, 콘텐츠 제작, 번역 서비스, 교육, 의료, 연구 및 개발 등 다양한 산업과 애플리케이션에서 광범위하게 사용되고 있다. 그러나 이러한 모델들은 높은 계산 비용, 데이터 편향성, 해석 가능성의 한계, 윤리적 고려사항 등의 과제에 직면해 있다.

초거대 언어 모델은 자연어 처리 분야에서 인공 지능의 중요한 발전을 의미한다. 이러한 모델은 다양한 애플리케이션에서 사람과 유사한 텍스트를 처리하고 생성하는 능력을 보여주고 있으며 앞으로도 계속해서 발전하고 더 넓은 적용 분야를 찾아갈 것으로 예상된다.

2.2.3 검색 증강 생성(RAG) 정의

검색 증강 생성(RAG, Retrieval-Augmented Generation)은 초거대 언어 모델(LLM, Large Language Model)의 출력을 최적화하여 응답을 생성하기 전에 학습 데이터 소스 외부의 신뢰할 수 있는 지식 베이스를 참조하도록 하는 프로세스이다(Yunfan Gao et al., 2024). 이는 초거대 언어 모델(LLM)의 학습 데이터에 따른 편향성이나 최신 데이터, 또는 오류를 보완할 수 있는 중요한 기능이다(정천수, 2023). 초거대 언어 모델(LLM)이 특정 주제에 대해 학습하지 않았거나 오래된 정보를 기반으로 학습한 경우 검색 증강 생성(RAG)은 신뢰할 수 있는 최신 정보를 검색하여 이를 보완할 수 있다(정천수, 2023).

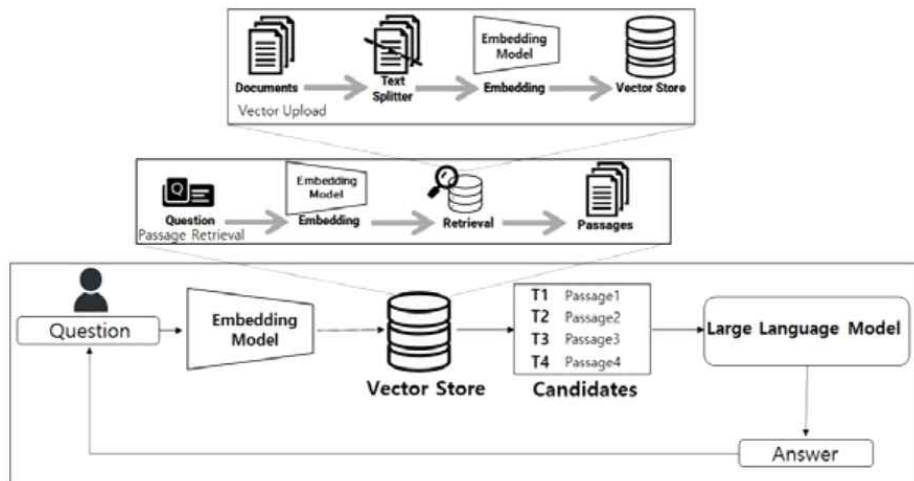
[그림 2-17]은 RAG 연구의 기술 트리를 보여주고 있으며 이 그림을 통해 RAG 기술의 발전 과정과 주요 구성 요소를 확인할 수 있다. RAG는 검색 기술과 생성 기술의 결합으로 이루어져 있으며, 각 기술의 발전이 RAG의 성능 향상으로 이어지고 있음을 알 수 있다.



[그림 2-17] RAG 연구의 기술 트리 (Yunfan Gao et al., 2024)

검색 증강 생성(RAG)은 검색기와 생성기라는 두 가지 주요 구성 요소로 이루어져 있다. 검색기는 대규모 텍스트 데이터베이스나 인터넷에서 관련성 높은 문서나 정보를 검색하는 역할을 한다. 주로 BM25⁵⁾나 TF-IDF(Term Frequency-Inverse Document Frequency)⁶⁾ 같은 전통적인 정보 검색 기법을 사용하거나 최근에는 머신 러닝 기반의 랭킹 모델을 적용하는 사례도 증가하고 있다. 생성기는 검색된 정보를 바탕으로 자연어 생성을 수행하며 GPT-4와 같은 초거대 언어 모델(LLM)이 사용된다. 이 생성기는 검색된 문서나 정보를 바탕으로 요약, 질의응답, 문장 생성 등의 작업을 수행한다.

[그림 2-18]과 같이 초거대 언어 모델을 활용해 검색 증강 생성(RAG)이 동작하는 구조를 확인할 수 있다. 이 그림에서 RAG 시스템은 사용자의 쿼리를 받아 관련 문서를 검색하고, 이를 초거대 언어 모델에 제공하여 최종 응답을 생성하는 과정을 거친다. 이 과정을 통해 초거대 언어 모델은 더 정확하고 최신의 정보를 바탕으로 응답을 생성할 수 있다.

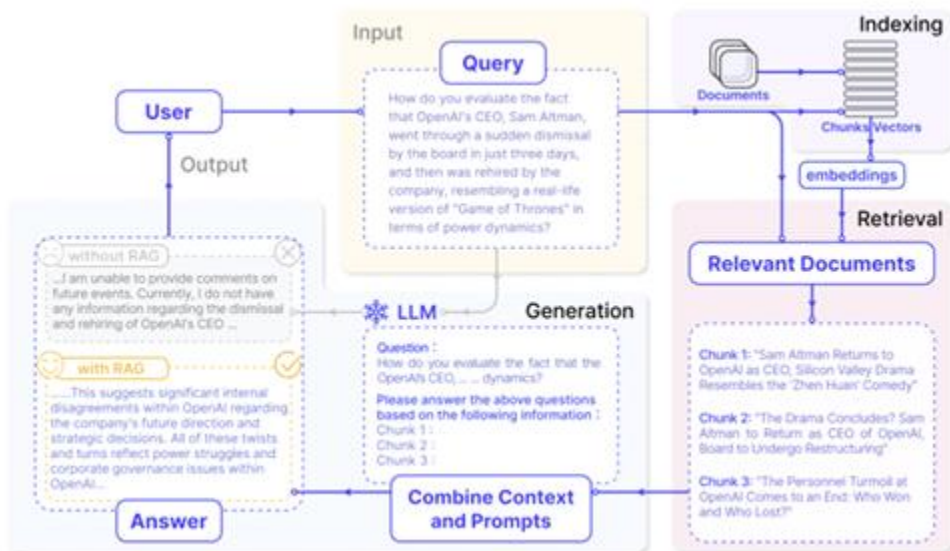


[그림 2-18] LLM을 활용한 RAG 구조도 (신중민 외, 2023)

5) 키워드 기반의 랭킹 알고리즘

6) BoW 표현 방법에서 파생된 빈도수 기반 문서 표현 방법. 단어가 문서에 몇 개나 포함되었는지(TF), 그 단어가 얼마나 많은 문서에 등장하였는지(IDF)를 모두 고려한다.

[그림 2-19]는 질의 응답에 적용된 검색 증강 생성(RAG)의 대표적인 사례를 보여준다. 이 그림에서 볼 수 있듯이, 사용자의 질문에 대해 관련 문서를 검색하고, 이를 바탕으로 초거대 언어 모델이 응답을 생성하는 과정이 잘 나타나 있다.



[그림 2-19] 질의 응답에 적용된 검색 증강 생성(RAG)의 대표적인 사례
(Yunfan Gao et al., 2024)

검색 증강 생성(RAG)의 주요 장점은 초거대 언어 모델의 생성 능력과 외부 지식 베이스의 최신성 및 정확성을 결합할 수 있다는 점이다. 이를 통해 모델의 응답이 더욱 정확하고 최신 정보를 반영할 수 있게 된다. 또한, 모델이 생성한 내용의 출처를 명확히 할 수 있어 설명 가능성도 향상된다.

RAG는 질의응답 시스템, 자동 요약, 콘텐츠 생성 및 개선, 맞춤 추천, 정보 검색 및 지식 관리, 대화형 AI 등 다양한 분야에서 활용되고 있다. 예를 들어, 질의응답 시스템에서는 사용자의 질문에 대해 최신 정보를 바탕으로 정확한 답변을 제공하며, 자동 요약 시스템에서는 대량의 문서를 검색하고 관련 정보를 추출하여 간결한 요약을 생성한다.

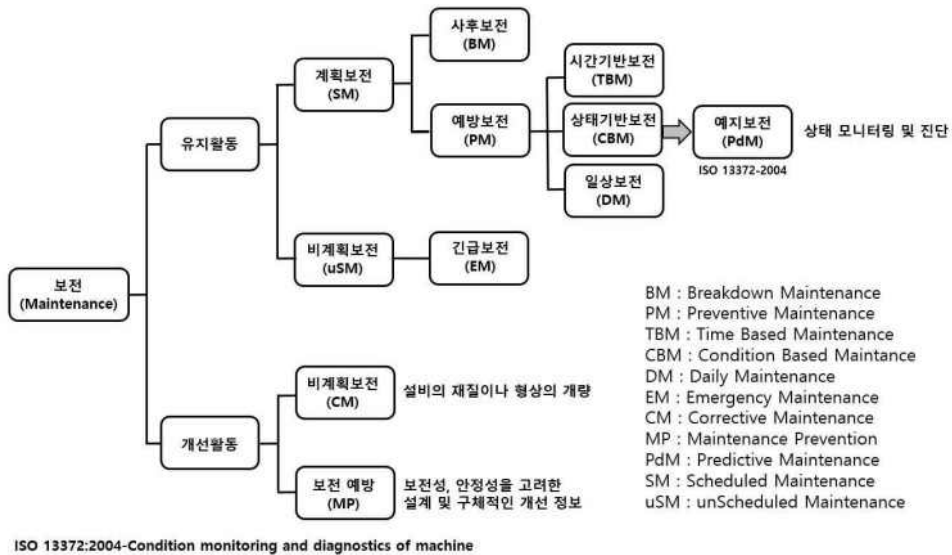
향후 검색 증강 생성(RAG) 연구는 더욱 효율적인 검색 알고리즘 개발, 다양한 형태의 데이터(텍스트, 이미지, 음성 등)를 통합적으로 처리할 수 있는 멀티모달 RAG 시스템 개발, 실시간으로 업데이트되는 동적 지식 베이스와의 연동 등의 방향으로 진행될 것으로 예상된다. 검색 증강 생성(RAG)은 초거대 언어 모델의 한계를 극복하고 더욱 정확하고 신뢰할 수 있는 정보를 제공하는 데 중요한 역할을 할 것으로 기대된다. 이는 다양한 분야에서 AI 시스템의 성능을 크게 향상시킬 수 있는 잠재력을 가지고 있다.

2.3 설비보전의 정의

설비보전은 제조 산업에서 생산 시설과 장비의 기능을 유지하고 향상시키기 위한 모든 활동을 포괄하는 중요한 개념으로서 설비의 운전과 유지보수에 소요되는 비용과 열화에 의한 손실 비용을 최소화하고, 설비 상태를 최상으로 유지하여 생산성과 품질을 높이고자 하는 모든 제반 활동들을 말한다(한국 생산성 본부, 1983). 이는 단순히 고장 난 장비를 수리하는 것을 넘어서 예방, 예측, 개선, 그리고 대응 전략을 포함하는 종합적인 관리 체계를 의미한다.

효과적인 설비보전은 여러 가지 중요한 목표를 가지고 있다. 주요 목표로는 장비의 수명 연장, 가동 중지 시간 최소화, 생산성 극대화, 생산 품질 향상, 안전성 확보 등이 있다. 이러한 활동들은 궁극적으로 기업의 수익성 향상으로 이어진다. 구체적으로, 설비보전은 설비의 가용성을 극대화하고, 생산 중단 시간을 최소화하며, 제품 품질을 향상시키고, 설비 수명을 연장하는 동시에 유지보수 비용을 최적화하는 것을 목표로 한다. 또한 작업 환경의 안전성을 확보하고 에너지 효율성을 향상시키는 것도 중요한 목표 중 하나이다.

설비보전은 방식에 따라 여러 가지로 분류될 수 있다. [그림 2-20]과 같이 유지 활동과 개선 활동으로 나누고 유지 활동에는 계획 보전(SM, Scheduled Maintenance)과 비계획보전(unScheduled Maintenance), 개선활동에는 비계획보전(Corrective Maintenance), 보전예방(Maintenance Prevention)으로 구분 된다. 계획보전 단계중 사후보전은 고장 발생 후 수리하는 방식이며 예방보전은 정기적인 점검과 유지보수를 통해 고장을 미리 방지하는 방식이다. 예지보전 단계는 최신 기술을 활용하여 설비의 상태를 실시간으로 모니터링하고 고장을 예측하여 대응하는 방식이고, 개량보전은 설비의 성능을 개선하거나 새로운 기능을 추가하는 방식이며 비계획보전(CM, Corrective Maintenance)과 보전 예방(MP, Maintenance Prevention)으로 나눌 수 있다. 각 보전방식은 고유한 특징과 장단점을 가지고 있으며, 기업의 상황과 필요에 따라 적절히 선택되어야 한다.



[그림 2-20] 설비보전의 분류 (ISO 13372, 2004)

설비보전의 역사는 산업화의 진전과 함께 발전해왔다. 초기에는 단순한 고장 수리에 중점을 두었지만, 시간이 지남에 따라 예방, 예측, 그리고 지능형 유지보수로 발전해왔다. 이러한 발전 과정은 각 시대의 기술적 진보와 산업의 요구 사항을 반영하고 있으며, 현대의 설비보전은 IoT, 빅 데이터, 인공지능 등의 첨단 기술을 활용하는 단계에 이르렀다.

효과적인 설비보전 전략은 이러한 다양한 방식을 기업의 상황에 맞게 적절히 조합하여 수립해야 한다. 이를 위해서는 기업의 생산 목표, 설비의 특성, 운영 환경, 비용구조 등을 종합적으로 고려해야 한다. 또한 기술의 발전과 산업 환경의 변화에 따라 설비보전 전략도 지속적으로 진화하고 있어, 이에 대한 꾸준한 관심과 투자가 필요하다.

설비보전은 현대 제조업에서 생산성, 품질, 안전성, 그리고 수익성을 결정짓는 핵심적인 요소이다. 따라서 기업은 설비보전에 대한 체계적인 접근과 지속적인 개선을 통해 경쟁력을 강화하고 지속 가능한 성장을 달성할 수 있다. 본 연구에서는 이러한 설비보전의 중요성을 인식하고, 특히 중소기업의 상황에 적합한 효과적인 설비보전 방안을 모색하고자 한다.

2.3.1 설비보전의 역사

설비보전의 역사는 산업화의 진전과 함께 발전해왔다. 19세기 후반부터 20세기 초반까지는 대부분의 기업들이 사후보전 방식을 채택했다. 이 시기에는 설비가 고장 날 때까지 사용하다가 문제가 발생하면 그때 수리하는 방식이 일반적이었다(Peter Poór et al. 2019). 설비의 복잡성이 낮고 생산 중단이 영향이 상대적으로 적었기 때문에 이러한 접근 방식이 통용되었다.

1차 세계대전을 거치면서 대량 생산 체제가 확립되고 설비의 복잡성이 증가함에 따라, 1920년대부터 예방보전의 개념이 도입되기 시작했다(Peter Poór et al. 2019). 기업들은 정기적인 점검과 유지보수를 통해 설비의 갑작스러운 고장을 예방하고 생산성을 향상시킬 수 있다는 점을 인식하기 시작했다. 이는 설비보전에 대한 인식의 중요한 전환점이 되었다(Peter Poór et al. 2019).

2차 세계대전 동안에는 군수 물자의 안정적인 공급을 위해 설비 신뢰성에 대한 관심이 크게 증가했다. 이 시기에 통계적 품질 관리 기법이 발전하면서 이를 설비관리에도 적용하려는 시도가 있었다. 1950년대에 General Electric(GE)에서는 예방보전의 개념을 공식화하고 정기적인 검사와 부품 교체를 통해 설비의 신뢰성을 향상시키는 방법을 체계화했다(최무호 외. 2010). 이는 현대적 의미의 예방 보전의 시작이라고 볼 수 있다.

1960년대에는 신뢰성 중심 유지보수(RCM, Reliability Centered Maintenance) 개념이 항공 산업을 중심으로 발전했다(Peter Poór et al. 2019). RCM은 설비의 기능과 고장모드를 분석하여 최적의 유지보수 전략을 수립하는 방법론으로, 이후 다양한 산업 분야로 확산되었다. 이는 설비보전에 대한 보다 체계적이고 과학적인 접근을 가능하게 했다.

1970년대에는 설비 상태 모니터링 기술의 발전으로 예측 유지보수의 기반이 마련되었다. 진동 분석, 열화상 분석 등의 기술을 통해 설비의 상태를 실시간으로 파악하고 잠재적인 문제를 조기에 발견할 수 있게 되었다. 이는 보다 선제적이고 효율적인 설비보전을 가능하게 했다.

1980년대에는 일본의 제조업 혁신 과정에서 총체적 생산보전(TPM,

Total Productive Maintenance) 개념이 등장하면서 모든 조직 구성원이 설비관리에 참여하는 종합적인 접근 방식으로 전 세계적으로 큰 영향을 미쳤다(최무호 외, 2010). 이는 설비보전을 단순한 기술적 문제가 아닌 조직 전체의 문화와 연결시키는 중요한 변화였다.

1990년대부터는 컴퓨터 기술의 발전으로 전산화된 유지보수 관리 시스템(CMMS, Computerized Maintenance Management System)이 도입되었다(Peter Poór et al. 2019). CMMS를 통해 유지보수 작업의 계획, 실행, 기록, 분석이 더욱 효율적으로 이루어질 수 있게 되었다(김연수 외, 2004). 이는 설비보전 데이터의 체계적인 관리와 분석을 가능하게 했다.

2000년대 이후에는 IoT(사물인터넷), 빅 데이터, 인공지능 기술의 발전으로 예지보전(Predictive Maintenance) 시스템이 급속도로 발전하고 있다(Peter Poór et al. 2019). 이러한 시스템은 설비의 상태를 실시간으로 모니터링하고 고장을 사전에 예측하여 최적의 시점에 유지보수를 수행할 수 있게 해준다. 이는 설비보전의 정확성과 효율성을 크게 향상시켰다.

현재 설비보전 분야는 Industry 4.0과 스마트 팩토리의 개념과 함께 더욱 진화하고 있다. 디지털 트윈(DX), 증강현실(AR), 가상현실(VR) 등의 기술을 활용한 새로운 유지보수 방식이 등장하고 있으며, 사람의 개입을 최소화하면서 자가 진단, 유지보수, 수리를 수행할 수 있는 시스템이 개발되면서 AI 및 자율 유지관리에 관심이 커졌다. 유지관리 일정을 최적화하고, 신뢰성을 향상하며 비용을 절감하기 위해 AI 기반 솔루션이 점점 더 늘어나고 있다. 이러한 설비보전의 역사적 발전 과정은 단순한 고장 수리에서 시작하여 예방, 예측, 그리고 지능형 유지보수로 발전해 온 과정을 보여준다. 각 시대의 기술적 진보와 산업의 요구 사항이 설비보전 방식의 진화를 이끌어왔으며, 앞으로도 새로운 기술의 등장과 함께 설비보전 분야는 계속해서 발전할 것으로 예상된다.

2.3.2 예방보전의 정의

예방보전(PM, Preventive Maintenance)은 설비보전 방식 중 하나로써 장비의 고장을 미연에 방지하고 최적의 운전상태를 유지하는 것을 목표로 하는 선제적인 유지보수 전략이다(권오운 외, 2003). 이는 정기적이고 계획적인 유지보수 활동을 통해 설비의 신뢰성을 향상시키고 예상치 못한 가동 중지 시간을 줄이며 장비의 수명을 연장하는 것을 목표로 한 매우 유용한 전략이다.

예방보전의 개념은 1920년대부터 도입되기 시작했으며 1950년대 General Electric(GE)에 의해 공식화되고 체계화되었다(Peter Poór et al., 2019). 당시 대량 생산 체제의 확립과 설비의 복잡성 증가로 인해 기업들은 갑작스러운 설비 고장이 생산성에 미치는 심각한 영향을 인식하게 되었다. 이에 따라 정기적인 점검과 유지보수를 통해 고장을 사전에 방지하고자 하는 예방보전의 필요성이 강조되었다.

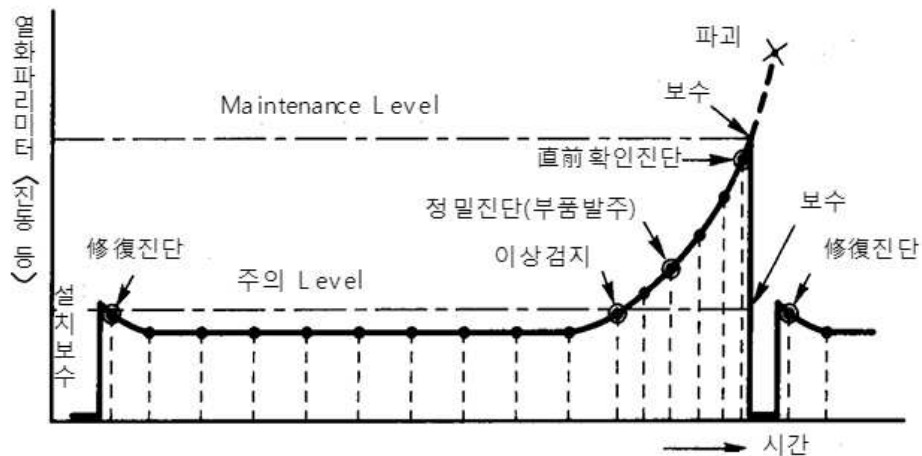
예방보전의 주요 구성 요소로는 정기 점검, 계획된 유지보수, 상태 모니터링, 데이터 분석, 그리고 설비 이력 관리가 있다. 정기 점검을 통해 설비의 상태를 주기적으로 검사하여 잠재적인 문제를 조기에 발견하고 계획된 유지보수를 통해 미리 정해진 일정에 따라 부품 교체, 윤활, 청소 등의 작업을 수행한다(Liliane Pintelon et al., 2008). 상태 모니터링은 설비의 주요 파라미터를 지속적으로 관찰하여 성능 저하나 이상 징후를 감지하며, 데이터 분석을 통해 수집된 유지보수 데이터를 분석하여 최적의 유지보수 주기와 방법을 결정한다. 설비 이력 관리의 각 설비의 유지보수 이력을 체계적으로 기록하고 관리하는 것을 포함한다.

예방보전의 주요 장점으로서는 설비의 신뢰성 향상, 가동 중지 시간 감소, 설비 수명 연장, 유지보수 비용의 장기적 절감, 생산 품질 향상, 작업 안전성 증대 등이 있다(Peter Poór et al., 2019). 특히, 계획된 유지보수를 통해 생산 일정을 보다 예측 가능하게 만들어 전체적인 생산 효율성을 높일 수 있다.

그러나 예방보전에는 몇 가지 주요 단점도 존재한다. 우선 초기 구현

비용이 높다는 단점이 있다. 이는 전문적인 설비 진단 기술의 습득, 초기 투자 비용, 그리고 시스템 유지 및 관리 비용 등이 소요되기 때문이다. 또한, 때로는 불필요한 유지보수 작업으로 인한 비용 낭비 가능성이 있으며 설비의 실제 상태와 관계없이 정해진 일정에 따라 유지보수를 수행함으로써 발생할 수 있는 비효율성도 존재한다(Peter Poór et al., 2019). 그리고 예상치 못한 고장을 완전히 방지할 수는 없다는 한계가 있다.

예방보전은 시간 기반 유지보수(TBM)와 상태 기반 유지보수(CBM)로 세분화될 수 있다. TBM은 정해진 시간 간격으로 유지보수를 수행하는 방식이며, CBM은 설비의 실제 상태를 모니터링하여 필요할 때 유지보수를 수행하는 방식이다(권오운 외, 2004). [그림 2-21]은 CBM에 있어서 진단 데이터, 진단 주기, 보전 타이밍의 관계를 중심으로 상태기준보전(CBM)의 실시과정을 표현하고 있다. 현대의 예방보전은 컴퓨터화된 유지보수 관리 시스템(CMMS)의 도입으로 더욱 효율적으로 수행되고 있으며 최근에는 IoT 센서, 빅데이터 분석, 인공지능 기술 등을 활용한 고급 예방보전 시스템이 등장하고 있다.



[그림 2-21] 상태기준보전(CBM)의 실시 과정 (권오운 외, 2004)

예방보전은 많은 기업들이 채택하고 있으며 특히 대기업에서 주로 활용되고 있는 보전방식이다. 그러나 중소기업의 경우에는 초기 투자 비용의 부

담과 전문인력 확보의 어려움 등으로 인해 완전한 예방보전 시스템을 구축하는 데 큰 부담이 되고 있다. 따라서 중소기업의 현실에 맞는 효과적이고 경제적인 예방보전 전략의 개발이 필요한 상황이다.

2.3.3 사후보전

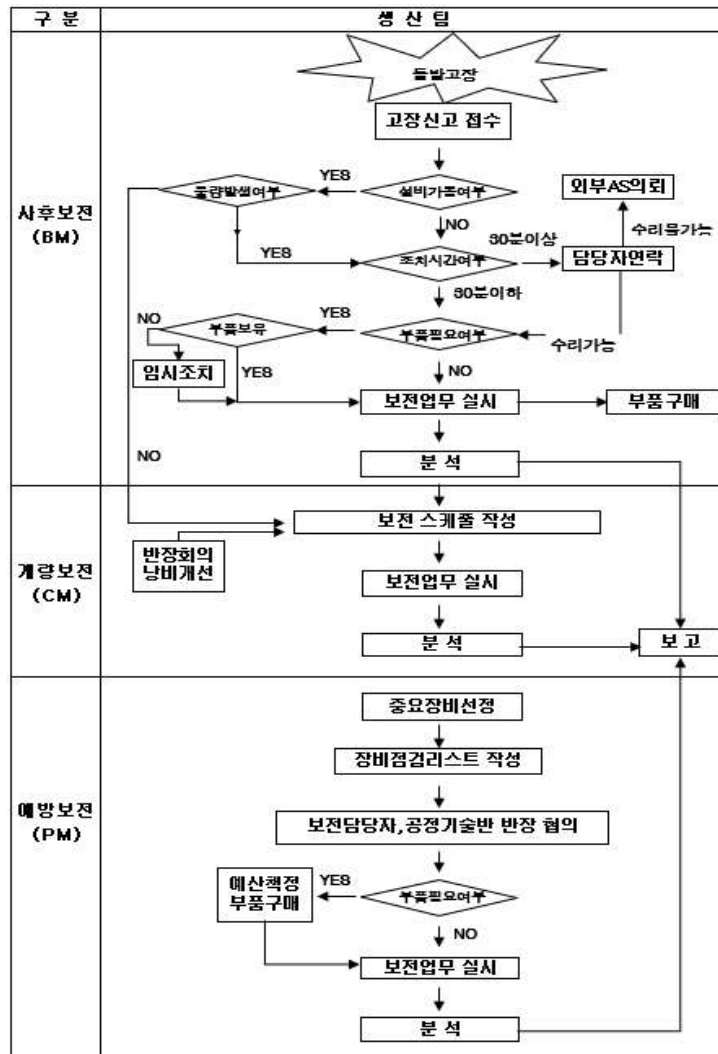
사후보전은 기계나 설비에 문제가 발생한 후에 적용되는 유지 관리 전략으로서 고장 유지 관리(Breakdown Maintenance, BM)라고 표현되기도 한다. 이는 기계 고장이 발생한 후 기계나 설비를 운영이 가능한 상태로 회복하기 위해서 필요한 부품을 수리하거나 교환해서 기능을 복구하는 활동을 의미한다. 사후보전의 주요 목표는 필요한 부품을 수리하거나 교체하여 장비를 작동 상태로 복원하는 것이며 이는 예상치 못한 장비 오류를 해결하고 복구하여 중단을 최소화하면서 운영을 계속할 수 있도록 하는 데 중점을 둔다(Peter Poór et al., 2019).

사후보전의 단계는 크게 다음과 같이 네 가지로 분류할 수 있다. 첫 번째는 문제 진단 단계로 장비에 고장이 발생하면 원인을 분석하고 장애 문제를 진단한 후 적절한 수리 방법을 결정한다. 두 번째는 복원 단계로 손상된 부품을 교체하거나 조정하여 생산설비의 성능을 복원시킨다. 세 번째는 문제 해결 단계로 설비가 고장 부위가 크거나 전반적인 문제가 발생할 때 이를 수리하여 원래 상태로 복구한다. 마지막 네 번째 단계는 긴급 대응 단계에서는 예상하지 못한 급박한 고장에 신속하게 대응하여 생산 중단시간을 최소화하고 생산성을 향상시키는 데 주력하는 단계이다.

사후보전의 특징으로는 낮은 초기 비용과 자원 활용의 효율성이 높다는 장점이 있다. 고장이 발생한 경우에만 비용이 발생하므로 초기 유지 관리 비용이 낮고 장비가 실제로 고장 날 때까지 운영할 수 있어 자원을 절약할 수 있게 된다. 또한, 예방보전 비용이 고장 후 복구비용을 초과할 때 특히 유용하게 활용될 수 있다. 그러나 대형 설비 사고의 위험 가능성이 있으며 돌발 고장의 경우 수리 시간 예측이 어렵다는 단점이 있다. 설비의 고장이 발생하면 그에 따른 사고 위험이 증가하며 이는 설비의 손상뿐만 아니라 작업자의 안전에도 위협을 가할 수 있다. 그리고 고장이 예기치 않게 발생하면 수리 시간을 정확히 예측하는 것이 어렵고 이로 인해 생산 일정이 크게 영향을 받을 수 있다.

본 연구의 샘플링이 된 브랜드 자동차 부품 공급기업인 제조업체 A사

는 [그림 2-22]와 같이 사후보전에 대응하고 있으며 사후보전 내용에 따라 개량보전 및 예방보전을 시행하고 있다. 이와 같이 제조업체에서는 품질과 가동률을 안정시키기 위해 종합적인 유지보수 전략을 수립하여 사후보전과 예방보전을 균형 있게 적용하는 것이 필요하다. 제한된 자원으로 인해 완전한 예방보전 시스템을 구축하기 어려운 상황에서는 차선택으로서 사후보전의 효율성을 높이는 방안의 적용이 필요하다.



[그림 2-22] 제조설비 보전업무 흐름도 (A사, 2022)

2.3.4 개량보전의 정의

개량보전(CM, Corrective Maintenance)은 장비의 고장빈도를 줄이고 수명을 연장하며 복구비용을 절감하는 방법으로서 고장 원인을 규명하여 유지보수성을 향상시키기 위해 장비를 개선하는 활동이다(권오운 외., 2003). 이는 단순히 설비를 유지하는 것을 넘어서 보존하면서 개선하는 데 중점을 두고 접근하는 개념이다(Annie Francie Kouedeu et al., 2011).

개량보전을 통해 성능 저하를 해결하기 위해 부품을 변경하거나 성능을 개선하기 위해 재료 처리를 변경할 수 있다(Annie Francie Kouedeu et al., 2011). 내마모성을 강화하기 위해 표면 처리를 적용하거나 성능을 개선하기 위해 형태를 수정하기도 하며 응력을 감소시키기 위해 메커니즘을 변경하는 등의 활동을 하게 된다. 기존 장비의 재질 및 형상 등 설계 취약점을 체계적이고 적극적으로 개선하는데 중요한 역할을 한다. 또한, 설비의 가용성과 유지 관리 가능성을 향상시키기 위해 일상적인 계획 보전을 개발한다. 상적인 개선을 통해 얻은 유지보수 정보를 활용하는 것은 신규 설비의 향후 유지보수 방지나 기존 설비의 개선에 매우 중요하다. 이를 위해서는 유지보수 부서와 생산 기술부서 간의 협력이 필요하며, 매일 기계를 다루는 직원의 피드백을 통합해야 한다.

개량보전의 주요 목적은 노후화를 개선하고 고장을 줄이며 궁극적으로는 유지보수가 거의 필요 없는 설비나 장비로 유지시키는 역할을 하고 있다. 이를 통해 설비의 수명과 가동시간을 효과적으로 늘릴 수 있게 되었다. 이러한 이유로 개량보전은 설비관리에 있어 필수적인 개선 활동으로 자리 잡게 되었다. 그리고 신뢰성 향상, 유지 관리성 개선, 잦은 유지 관리 문제 개선, 작동성 향상, 안전성 증대, 유연성 증대, 경제적 효율성 향상 등을 주요 목표로 한다.

개량보전은 초기 관리 활동부터 폐기까지 설비의 수명주기 전반에 걸쳐 촉진되어야 한다. 이는 장기적으로 설비의 성능을 향상시키고 유지보수 비용을 줄일 수 있지만 초기 투자비용이 높고 개선 효과를 정확히 예측하기 어렵다는 단점이 있다. 그럼에도 불구하고 개량보전은 현대 제조업에서 설

비의 지속적인 개선과 효율성 향상을 위해 매우 선호하는 방식으로 산업 현장에서 활용되고 있다.

2.3.5 긴급보전 정의

긴급보전(EM, Emergency Maintenance)은 운영을 크게 방해하는 갑작스럽고 예상치 못한 장비 고장이나 고장을 해결하기 위해 수행되는 예정되지 않은 대응적 유지보수 활동을 지칭한다(권오운 외, 2004). 기본적으로 가동 중지 시간과 생산 손실을 최소화하기 위해 가능한 한 빨리 장비를 작동 상태로 복원하는 것이 주요한 목적이다.

긴급 보전은 긴급성 때문에 생산과 안전에 미치는 영향을 완화하기 위한 신속한 조치가 필요하며 보다 영구적인 해결 방법이 마련될 때까지 기능을 신속하게 복원하기 위해 임시 수리를 하는 경우도 있다. 즉각적인 위험을 해결하고 정상적인 운영을 복원하여 안전과 운영 연속성을 보장하는 것이 주요 목표이다. 갑작스러운 모터 고장, 전기 단락, 유압 시스템 고장 등의 장비 고장, 잘못된 설정, 부적절한 취급 또는 부적절한 유지 관리 절차로 인한 인적 오류, 전력 서지, 자연재해 또는 공급망 중단과 같은 외부 요인이나 검사 누락 및 상태 모니터링 시스템의 경고 신호 무시와 같은 유지 관리 소홀도 긴급 보전의 원인이 될 수 있다.

긴급 보전 절차는 초기 대응 및 상황 평가, 안전 확보, 신속한 진단, 자원 동원, 수리 및 복원, 테스트 및 검증, 문서화 및 분석, 후속 조치로 진행된다. 이 과정에서 오류의 정도와 운영에 미치는 영향을 신속히 평가하고, 피해지역을 확보하여 안전을 확보하고 추가 피해를 방지한다. 그 다음 진단 도구와 전문 지식을 활용하여 문제를 신속하게 찾아내고 실패의 근본 원인을 식별한다.

긴급 보전은 즉각적인 대응, 생산 중단 시간 최소화, 안전 확보, 고객 신뢰 유지 등의 측면에서 중요하다. 그러나 높은 비용, 리소스 할당 문제, 임시 수정의 위험, 운영 중단 등의 단점도 있다. 긴급 수리 중에 적용된 임시 수정 사항은 장기적인 신뢰성을 보장하기 위해 후속 조치가 필요할 수 있으며, 복원에 즉각적인 초점을 맞추다 보면 때로는 철저함이 타협될 수 있다.

사후보전 방식과 유사한 측면이 있지만 몇 가지 중요한 차이점이 있다.

긴급 보전은 즉각적인 대응이 필요한 계획되지 않은 활동인데 비해 사후보전은 고장 후 계획된 대응이 가능한 방식이다. 또한 긴급 보전은 자원에 더 큰 부담을 주고 긴급 조달이 필요할 수 있는 반면 사후보전은 더욱 통제되고 계획된 방식으로 자원을 활용할 수 있다.

2.4 설비종합효율(OEE) 정의

설비종합효율(Overall Equipment Effectiveness, OEE)은 생산설비의 효율성을 종합적으로 평가하는 핵심 지표로, 설비의 가동 상태를 시간적, 속도적 측면에서 평가하고 생산된 제품의 품질을 고려하여 설비가 부가가치 창출에 얼마나 효과적으로 기여하고 있는지를 나타낸다(백종엽 외, 2010). 이 지표는 시간가동율, 성능가동율, 양품율의 세 가지 요소를 곱하여 산출되며, 각 요소는 설비의 가용성, 성능, 그리고 생산품의 품질을 각각 반영한다.

설비종합효율(OEE)은 중소기업의 생산성 향상과 비용 절감에 중요한 역할을 한다. 중소기업은 대기업에 비해 상대적으로 제한된 자원과 인력을 가지고 있기 때문에, 효율적인 설비 관리가 더욱 중요하다. OEE를 통해 중소기업은 자원을 최적화하고, 비용을 절감하며, 품질을 향상시킬 수 있다. 또한, OEE 데이터는 의사결정을 지원하여 제한된 자원의 효율적 배분을 가능하게 하며, 지속적인 모니터링과 개선을 통해 장기적인 성장과 경쟁력 강화를 도모할 수 있다.

설비종합효율의 개념은 현대 경제에서 지속적인 투자 효과를 기대하는 환경에서 중요성을 갖게 되었다. 효율적으로 가동되는 공장에서는 새로운 투자를 통한 추가적인 효과를 얻기 어려워지면서, 기업들은 생산비용을 감소시키면서도 현재의 설비로 더 많은 생산을 해야 하는 상황에 직면하게 되었다(Kathleen E. Mckone, 2001). 이러한 배경에서 설비종합효율은 현재 설비의 시간적, 속도적 상태와 양품률을 종합하여 부가가치 창출에 대한 공헌도를 평가하는 중요한 도구로 자리잡게 되었다. 이를 통해 목표로 하는 가동 범위에 적절한 프로세스의 안정 상태와 신뢰성을 평가할 수 있으며, 고객이 요구하는 제품의 품질과 공급자의 요구사항을 만족시키기 위해 설비와 인력을 포함한 자원을 어떻게 잘 활용하는가를 보여준다(백종엽 외, 2010).

설비종합효율은 제조업체가 생산설비의 성과를 종합적으로 평가하고 개선하기 위한 핵심적인 지표로, 기업의 전반적인 생산 효율성과 경쟁력 향상에 직접적인 영향을 미치는 중요한 개념이다. 지속적인 OEE 모니터링과 개선을 통해 기업은 생산성 향상, 비용 절감, 품질 개선, 납기 준수를 향상 등 다양

한 경영 목표를 달성할 수 있다.

2.4.1 설비 효율화

설비 효율화는 설비의 가동상태를 양적, 질적인 면에서 파악하여 부가가치 창출을 최적화하는 활동을 의미한다(한국표준협회, 2000). 이는 설비의 고유 능력을 최대한 발휘하고 유지하면서 인적 자원의 능력을 최대한 활용하여 전체 시스템의 최적 상태를 유지하는 것을 목표로 한다. 생산이 증가하는 경우 생산량 증대 대책을 취하는 것은 물론 원가절감을 꾀하고, 생산이 감소될 경우 원가가 높아지지 않도록 인원 감축, 부자재의 원 단위 원가의 삭감 등 전체 가공비를 낮추는 활동을 추진한다(백종엽 외, 2010).

설비 효율화는 관리 기준에 따라 양적 측면과 질적 측면으로 나뉜다. 양적 측면은 설비의 가동시간 증대와 단위시간 내의 생산량을 말하며, 질적 측면은 불량품의 감소와 품질의 안정화 및 향상을 의미한다. 이를 통해 설비종합효율은 설비의 효율성을 제고하여 생산성을 향상하기 위한 지표로서 관리되고 있다(백종엽 외, 2010).

2.4.2 설비종합효율 측정

설비종합효율의 측정은 시간가동율, 성능가동율, 양품율을 각각 계산하고 이를 곱하여 이루어진다. 여기에 사용된 값은 전체 프로세스 공장, 프로세스 라인 또는 각각의 설비를 반영하고 있다(백종엽 외, 2010).

크리스터 이드하머(1997)는 설비종합효율에 의해 계산된 요인을 추적하면 설비가 예상한 것보다 더 많은 고장시간이 있었는지, 설비가 느린 속도로 가동되었거나 또는 경미한 중지 상태인지, 또는 더 많은 결함을 만들어냈는지 알 수 있다고 했다. 원인 분석은 설비종합효율의 퍼센트 수치 자체를 검토하는 것이 아니고 설비종합효율을 계산한 요인을 추적하는 것이다.

설비종합효율의 측정은 시간 가동율, 성능 가동율, 양품율을 각각 계산하고 이를 곱하여 이루어진다. 아래 [표 2-1]에서는 글로벌 수준의 이상적인 설비종합효율 수준을 제시하고 있다.

[표 2-1] 글로벌 수준의 설비종합효율 측정 기준 (Seichi Nakajima, 1998)

시간가동률(Availability)	>90%
성능가동률(Performance rate)	>95%
양품률(Quality Achieved)	>99%
설비종합효율(OEE)	>85%

설비종합효율과 관련된 주요 관리지표는 다음과 같다(이영상 외, 2003).

- ① 시간가동율 = (부하시간 - 정지시간)/부하시간
- ② 성능 가동율 = 속도가동율 × 정미가동율
- ③ 속도가동율 = 이론 Cycle Time / 실제 Cycle Time
- ④ 정미가동율 = 완성량 × 실제 Cycle Time/가동시간
- ⑤ 양품율 = 양품생산량 / 완성량
- ⑥ 설비종합효율 = 시간가동율 × 성능가동율 × 양품률

2.4.2.1 시간가동율

시간가동율은 설비의 계획 가동 시간 대비 실제 가동 시간의 비율을 나

타내며, 고장이나 셋업 등으로 인한 비가동 시간을 최소화하는 것이 중요하다. 시간가동율이 높다는 것은 더 많이 생산할 수 있다는 것을 의미하고, 총자산 이익률을 높게 할 수 있다는 의미이다.

설비는 사용할수록 생산성과 성능이 감소된다. 계획정지 동안에 일반적인 설비와 고장 수리를 수행하여 시간의 낭비를 최소화하는 한편 설비의 성능을 회복시킬 수 있다. 예방 보전과 예지보전 활동을 통해 계획되지 않은 정지시간을 줄이고 설비의 가용성을 향상시킬 수 있다(백종엽 외, 2010).

2.4.2.2 성능 가동율

성능가동율은 설비의 이론적 생산 능력 대비 실제 생산량의 비율을 의미한다. 많은 생산 공장은 현재 달성하고 있는 것보다 더 많은 생산성을 올릴 수 있는 잠재력을 가지고 있다.

프로세스 최적화를 통해 생산성을 높일 수 있으며, 이는 부하의 변경과 목표값의 변경을 최상의 효과성을 갖도록 제어하고 예기치 않은 이상 상태를 최소화하는 것을 포함한다. 데이터 분석과 통계적 도구를 활용하여 효과적인 진단과 성능 개선을 할 수 있으며, 전체적인 경제적 최적화 관점에서 접근해야 한다(백종엽 외, 2010).

최근 IoT와 빅데이터 분석 기술의 발전은 OEE 측정과 개선에 새로운 가능성을 열어주고 있다. 이러한 기술 동향은 AI 기반 사후보전 시스템과 밀접하게 연관되어 있으며, OEE를 다양한 방식으로 향상시킬 수 있다. 실시간 데이터 수집, 예측적 분석, 품질 관리 고도화, 통합 분석 플랫폼, 디지털 트윈, 모바일 모니터링 등의 기술이 OEE 개선에 활용되고 있다. 이러한 최신 기술 동향은 특히 중소기업에게 중요한 의미를 갖는다. 과거에는 고비용으로 인해 대기업 위주로 도입되던 첨단 설비 관리 시스템들이, 클라우드 서비스와 IoT 기술의 발전으로 인해 중소기업도 적은 초기 투자로 도입할 수 있게 되었기 때문이다. 이는 AI 기반 사후보전 시스템과 결합하여 중소기업의 OEE를 획기적으로 개선할 수 있는 기회를 제공한다.

2.4.2.3 양품율

양품율은 전체 생산량 중 품질 기준을 충족하는 제품의 비율로, 생산된 제품의 품질 수준을 반영한다. 품질을 향상시키는 가장 큰 기회는 프로세스의 가변성(Variability)를 줄이는 것이다.

프로세스의 가변성을 줄이기 위한 방법으로는 지능적인 필드 장치 활용과 향상된 조절 제어(Regulatory Control) 시스템 도입 등이 있다. 센서, 발신기, 제어 밸브와 같은 필드 장치의 정밀도 향상은 프로세스의 가변성을 감소시키는 데 도움이 되며, 조절 제어 시스템은 공급 원료의 품질, 주변 조건, 장비 성능 등의 변화에도 불구하고 프로세스에서의 변화를 최소화하여 일관된 품질의 제품을 생산할 수 있게 한다(백종엽 외, 2010).

2.4.3 설비종합관리(TPM) 측정 방안

설비종합관리(TPM, Total Productive Maintenance)는 전사적 차원에서 조직 구성원 전체가 참여하는 종합적인 설비관리 활동이다. TPM은 주로 5S(정리, 정돈, 청소, 청결, 습관화), 자주보전, 계획보전, 개별개선, 품질보전, MP활동, 교육 및 훈련, 소집단 활동 등 8대 활동을 중심으로 이루어진다. 이러한 TPM 활동은 설비종합효율(OEE) 향상에 직접적으로 기여하며, 생산성 향상과 비용 절감을 통해 기업의 경쟁력 강화에 중요한 역할을 한다.

TPM의 성과 측정은 기업의 지속적인 개선과 경쟁력 유지를 위해 필수적이다. 조희중(2007)의 연구에 따르면, TPM 성과 측정에는 국내외에서 약 500여 종의 지표가 활용될 정도로 다양한 접근 방식이 존재한다. 그러나 이러한 다양한 지표들 중에서도 설비종합효율(OEE)은 TPM의 핵심적인 성과 지표로 널리 인식되고 있다. OEE는 TPM 활동의 효과를 시간가동율, 성능가동율, 양품율이라는 세 가지 핵심 요소를 통해 종합적으로 평가할 수 있게 해주므로, TPM 성과 측정의 중요한 도구로 활용되고 있다.

2.5 선행연구 분석

2.5.1 선행연구 분석 결과

국내의 설비보전 관련 선행연구들을 아래 [표 2-2]와 같이 정리하였다. 설비보전 관련 선행연구들에서 설비관리(TPM)와 관련된 다양한 연구가 진행되어 왔음을 확인할 수 있다. 이러한 연구들은 주로 예방보전, 예지보전, 그리고 전반적인 설비종합효율(OEE) 향상에 초점을 맞추고 있다.

[표 2-2] 설비보전 선행연구 분석 (연구자 재구성, 2024)

연구자	연구주제	연구방법	연구결과	주제어
능률협회 컨설팅 (1991)	TPM (설비보 전) 정의	TPM 탄생 배 경 및 보급 이 유, TPM 5대 기능과 특징 분석	- 설비의 광학적 해설 - 설비 효율화를 도모 - 설비공사 향상에 기여하는 계기 마련 - 자신의 설비는 자신이 유지·관리 대명제 제시	TPM 배경, TPM 이득, TPM 정의, TPM 5대 기능, 한국기업 TPM 적용사례
김정식 외 (1997)	보 전 계 획 을 위한 체계적 인 설비 선정 및 분석	Stacker Crane 특징 분석,	- 평가지도 및 점수 배점 (고장빈도, 고장 강도, 고장이 미치는 범위, 품질, 수리 비용, 치명도)	고장빈도, 고장강도, 수리비, 치명도,
함효준 외 (2001)	보전정책 수립을 위한 설비구조분석	설비 구조분석 개념 정립, 실증사례연구,	- 효율적인 보전을 위한 설비의 설계 측면 이해 필요 - 설비 부품정보 체계적 분석 - 설비구조분석의 통제 가능 활동의 구분 필요	설비구조분석, Black Box, 생산보전, Drive Unit,
김연수 외 (2004)	물류설비 보 전관리를 한 고장 진단정보 원 시스템	Stacker Crane 특징 분석,	- 물류설비 Stacker Crane 유지관리 및 고장진단	설비관리, 고장진단, 고장예측, 설비정보

연구자	연구주제	연구방법	연구결과	주제어
최병관 (2006)	활동기준원가를 이용한 설비보전비 계산 시스템에 관한 연구	TPM과 ABC의 장점을 이용하여 설비보전비를 정량화	- 설비보전 활동의 필요성을 강조 (생산성 품질 극대화) - 전사적생산보전(TPM), 활동기준원가계산(ABC) 적용	설비보전, TPM, ABC, 불비보전 비용구조 분석, 보전비 계산, 경제성 분석
박선호 외 (2007)	제조현장의 설비보전 능력향상을 위한 자주학습체계에 관한 연구	기존 교육훈련 문제점 분석, 현장 실상 검토, 자주학습 OJT 체계 적용	- 설비 보전원의 역할이 매우 중요함을 강조 - 기존 교육훈련의 문제점을 지적하고 자주학습 중심의 교육훈련을 통해 보존능력 향상을 기대	Maintenance, OJT, Autonomous, Training System
장원태 (2007)	최적보전정책을 위한 경제적 설비구조분석	설비구조분석, Black Box 결정, 알고리즘 모의실험	- 보전비 고려 경제적 설비구조분석 모형 개발 - 총보전비용의 정확한 산출에 관한 연구 - 사후보전의 의사결정 방법론 개발에 관한 연구 필요성 강조	장비구조 분석, 유지비용, 유지관리 정책, 총보존비
김봉운 (2008)	보전 수준을 고려한 효율적 설비보전 시스템 구축	설비 보전 수준 진단, 최적화 시스템 구현, 모형 개발 및 적용, 성과측정	- 설비 보전원의 역할이 80% 차지함 - 설비 보전원의 학습능력을 향상, 정신적 멘탈 훈련 필요	설비보전, 설비관리, KPI, 보전수준, CMMS, RCM, 보전체계
황호연 (2010)	발전설비 신뢰도 향상을 위한 보전관리시스템 구축 방안 연구	ERP 및 Portal 환경 구현, 업무지침 개선, Process 개선, RCM 운영, 표준화	- 발전설비 보전관리시스템 신뢰도 향상 방안 연구 - 표준화, 선진 정비관리 기법 도입 - 관리 주체인 전문 인력 관리의 중요성 강조	설비관리, 보전관리시스템, PIS, PCM, RBI, PM
최무호 외 (2010)	중소기업 TPM 성과에 영향을 주는 요인에 대한 실증적 연구	TPM 성과요인 분석, 통계분석, 구조방정식	- 대기업에서 증명된 TPM을 중소기업에 적용, 실증연구 - 300명 이사 중소기업 설문 - TPM의 활동 요인과 성과 요인 간의 인과관계 증명	TPM, 중소기업, 활동요인, 성과요인, 인과관계, 경영성과 향상 인자
정민형 (2011)	사례를 통한 TPM 성과측정 KPI 및 경제적 효과 산출 방법	TPM 활동 측정 요소 분석, TPM 성과 측정 및 정량화	- 성과관점과 활동 관점의 15개 KPI 선정 - AHP 기법 활용 성과 측정 - 경제적 가치 환산 방식의 성과분석 활용 기대	TPM, BSC, KPI, 경제적 효과, 수량화, 성과 측정

연구자	연구주제	연구방법	연구결과	주제어
이명수 (2017)	다이캐스팅 설비를 위한 예지보전 기반 설비보전관리 시스템	모델별 기능 정의, 고장모드 FMEA 실시, AI 기반 알고리즘 활용	- 장시간 다운타임 발생으로 인한 제조업 생산성 악영향 우려 - 사후보전 이전에 예지보전 시스템 강화 필요 - 설비 효율 및 생산성 향상 기대	설비보전관리시스템, 신뢰성 중심 유지보수, 고장모드 영향분석, 기계학습
오창현 (2021)	예지보전 기술을 활용한 스마트 팩토리 효율화에 관한 연구	스마트 팩토리 효율화 실험, IoT 센서 활용, I-Forest 알고리즘 적용	- 설비 이상 진단 알고리즘 - 설비 가동시간 단축, 유휴시간 감소, 효율적 계획 정비 - 설비 가동율 향상을 목표	스마트팩토리, I-Forest알고리즘, 예지보전
김순철 (2023)	필름제조 연속생산시스템의 설비효율 향상을 위한 FOM-AI기반 설비 예지보전 연구	FOM 솔루션 분석, 예지보전 시스템 구축, 특정 회사 실증 사례 적용 및 평가	- 생산설비 연속 가동 기업 대상으로 실증 연구 - 설비관리 항목 중 설비 효율 향상을 위한 분야 선정 - 사후보전의 어려움으로 인해 예지보전으로 실증	생산시스템 개선, 예지보전, 설비효율 향상
박상현 (2023)	딥러닝을 이용한 음향기반 현장 적용형 공조설비 결함진단 시스템	센서를 활용해 공조설비에서 발생하는 음향신호 분석, 딥러닝 알고리즘으로 음향신호 분석, 공조설비 결함 진단	- 매우 강한 잡음 환경에서도 효과적인 음향 향상이 가능 - 멀티태스크 러닝 기반 음향신호처리 프레임워크를 구현 - 음향구간 검출 기법은 약 80%의 정확도를 나타냈으며, 음원 방향 추적기법은 프레임단위에서 39%, 발언 단위에서 83%의 정확도를 보임	음향신호, 딥러닝, 공조설비, 고장진단
이병욱 (2023)	IIoT 기반한 핵심유틸리티의 유지보수 최적화를 위한 공동활용 시스템	전용 컨트롤보드 개발, 클라우드 서버에 정보 수집, 통합 정보를 모든 운영자 및 관리자가 실시간으로 관리	- M/S의 단계를 9단계에서 3단계로 줄임 - 시스템 장애 상태 공유로 유지보수 지연시간을 획기적으로 줄임 - 유지보수 직원의 실시간 상태 감지 제공 - 비용과 공간적 문제를 획기적으로 개선	Repository Utilization, optimizing, IIoT, maintenance, Utilities

연구자	연구주제	연구방법	연구결과	주제어
이재호 외 (2023)	생성적 딥러닝 모형 기반 시계열 데이터의 이상치 탐지 비교 연구	생성적 딥러닝 모형 기반의 여러 모형 소개, 합성 데이터와 실제 데이터를 기반으로 성능 비교를 수행	- 시계열 데이터에 대한 이상치 탐지에 대해 LSTM 구조와 GAN 모형을 결합한 TadGAN 모형이 가장 우수한 성능을 나타냄	공정 데이터, 딥러닝 모형, 생성적 모형, 시계열 데이터, 이상치 탐지, Anomaly detection, deep learning models, generative models, sensor data within process, time series
조운형 외 (2023)	딥러닝 알고리즘을 활용한 진동 신호 기반 회전설비 예지보전	센서를 활용해 진동 신호 분석, 오토 인코더 알고리즘 활용, 회 전 설비에 가속도 센서를 부착해 데이터 수집	- 오토 인코더의 잠재적 벡터를 디코딩하여 오차율 계산 - 정상 작동과 다르게 작동하는 이상치 탐지를 통해 고장 예지 - 높은 예측 정확도를 보여주며, 회전 설비의 고장 예지에 효과적으로 활용될 수 있음을 시뮬레이션으로 입증	딥러닝, 이상치 탐지, 회전설비, 예지보전

황호연(2010)은 발전설비 분야에서 신뢰도 향상을 위해 개선된 보전관리시스템의 구축이 필요함을 강조하였다. 장원태(2007), 송태영(2002), 함효준(2001)은 보전정책의 수립과 집행의 어려움을 지적하며, 이를 극복하기 위한 방안으로 관리 대상 설비에 대한 구체적인 보전방식을 규명할 수 있는 설비구조분석 방법을 제시하였다.

이명수(2017)는 설비관리 시 사후보전에만 의존하는 경우 예기치 못한 오류나 다운타임 감소에 큰 영향이 없다고 지적하였다. 이는 사후보전의 한계를 인식하고 개선의 필요성을 제시한 중요한 연구이다. 김봉운(2008)은 설비 보전원의 역할이 80% 차지한다고 밝히며, 설비 보전원의 학습 능력 향상과 정신적 훈련의 필요성을 강조하였다.

최무호 외(2010)는 대기업에서 증명된 TPM을 중소기업에 적용하여 실증 연구를 수행하였으며, TPM의 활동 요인과 성과 요인 간의 인과관계를 증명하였다. 이는 중소기업에서의 TPM 적용 가능성을 보여주는 중요한 연

구이다.

오창현(2021)과 김순철(2023)은 각각 예지보전 기술을 활용한 스마트 토리 효율화와 필름제조 연속 생산시스템의 설비 효율 향상을 위한 예지보전 연구를 수행하였다. 또한 박상현(2023), 이병옥(2023), 이재호 외(2023), 조운형 외(2023)의 최근 연구들은 딥러닝, IoT, 빅데이터 분석 등 최신 기술을 활용한 설비보전 방법을 제시하고 있다.

능률협회컨설팅(1991)은 TPM의 배경과 정의, 5대 기능 등을 제시하며 한국기업의 TPM 적용사례를 분석하였다. 김정식 외(1997)와 김연수 외(2004)는 각각 보전계획을 위한 체계적인 설비 선정 및 분석과 물류설비 보전관리를 위한 고장 및 진단정보 지원 시스템에 대해 연구하였다.

최병관(2006)은 활동기준원가를 이용한 설비보전비 계산 시스템에 관한 연구를 수행하였다. 박선호 외(2007)는 제조 현장의 설비보전 능력향상을 위한 자주 학습 체계에 관한 연구를 수행하였다. 정민형(2011)은 TPM 성과측정 KPI 및 경제적 효과 산출방법에 대한 연구를 수행하였다.

이영상 외(2007)는 KPI를 활용한 TPM 성과 지표에 대한 연구를 수행하였으며, 조희중(2007)은 TPM 추진 및 성과측정 방법론에 관한 연구를 통해 정량적 성과의 중요성을 강조하였다. 이희양(2009)은 설비종합효율 향상을 위한 TPM 활동 사례연구를 수행하였다.

문태수 외(2003)와 문혁동 외(1997)는 기업들이 생산성을 높이기 위해 정보화 투자를 늘려가고 있음을 지적하였다. 이는 설비보전 분야에서도 정보화와 디지털화의 중요성이 증가하고 있음을 시사한다. 그러나 박성현(1997)은 정보화 투자를 통한 생산성 향상의 결과가 과도한 투자 비용 및 관리비용의 증가로 인해 실패 사례도 증가하고 있음을 지적하였다. 이는 효율적이고 비용 효과적인 설비보전 시스템의 필요성을 강조한다.

이러한 선행연구들은 설비보전의 중요성과 다양한 접근 방식을 제시하고 있으나, 대부분 예방 보전이나 예지보전에 초점을 맞추거나 전통적인 TPM 방식에 의존하고 있다. 특히 중소기업의 현실을 고려한 사후보전 효율성 향상에 대한 연구는 미흡하다. 또한, 최근 급속도로 발전하고 있는 AI 기술을 설비보전, 특히 사후보전에 적용한 연구는 거의 없다.

2.5.2 설비보전 개선방안 연구

2.5.2.1 현재 설비 사후보전의 문제점

일반적으로 장비 고장으로 인한 사후보전 수행할 때, 단순한 문제, 또는 이전 경험과 동일하거나 유사한 문제가 발생한 경우에는 원인 파악에 많은 시간이 걸리지 않았다. 그러나 사전 경험이 없거나 숙련된 관리자가 없는 경우 문제를 진단하는 데 시간이 많이 걸릴 수 있다. 또한, 누구도 경험하지 못한 문제가 발생하거나 경험이 풍부한 선임 관리자가 없는 경우, 장비의 매뉴얼을 검색하거나 과거 수리 이력 조회에 의존해야 하므로 원인 파악에 많은 시간이 걸릴 수 있다. 이로 인해 장비 비가동 시간이 길어지고 잠재적인 손상이 발생하며, 결과적으로 비용 손실로 이어진다.

2.5.2.2 해결방법 검토

이러한 문제를 해결하기 위해 다양한 기술 솔루션을 검토한 결과, 초거대 언어 모델과 검색 증강 생성 시스템의 결합이 효과적인 해결책이 될 수 있다고 판단하였다. 초거대 언어 모델은 복잡한 자연어 처리 능력을 통해 설비 유지보수 관련 질문에 대한 정확하고 상세한 답변을 제공할 수 있으며, RAG 시스템은 기존의 매뉴얼과 과거 수리 이력 등 신뢰할 수 있는 정보를 실시간으로 검색하여 초거대 언어 모델의 응답을 보강함으로써 정확성과 신뢰성을 크게 향상시킬 수 있다.

① 초거대 언어 모델

Gemini와 같은 초거대 언어 모델은 시설 유지 관리 직원이 제기하는 복잡한 질문을 이해하고 답변하는 탁월한 능력을 갖추고 있다. 이러한 모델은 자연어 쿼리를 해석하고 상세한 상황 인식 응답을 제공할 수 있으므로 문제 해결 및 유지 관리 지침에 이상적이다.

② RAG(검색 증강 생성)

RAG는 초거대 언어 모델이 그럴듯하지만 부정확하거나 관련 없는 정보를 생성할 수 있는 일반적인 문제인 환각을 방지한다. 검색 메커니즘을

통합함으로써 RAG는 응답이 실제의 정확한 문서 및 데이터에 기반을 두도록 보장하여 신뢰성을 향상시킨다.

2.5.2.3 예비실험

예비 실험을 위해 Google의 Gemini 모델을 선택하여 Samsung S24 스마트폰의 매뉴얼을 데이터 세트로 사용해 RAG 모델을 학습시켰다.

```
from google.colab import drive
drive.mount('/content/drive')

from IPython.display import Markdown

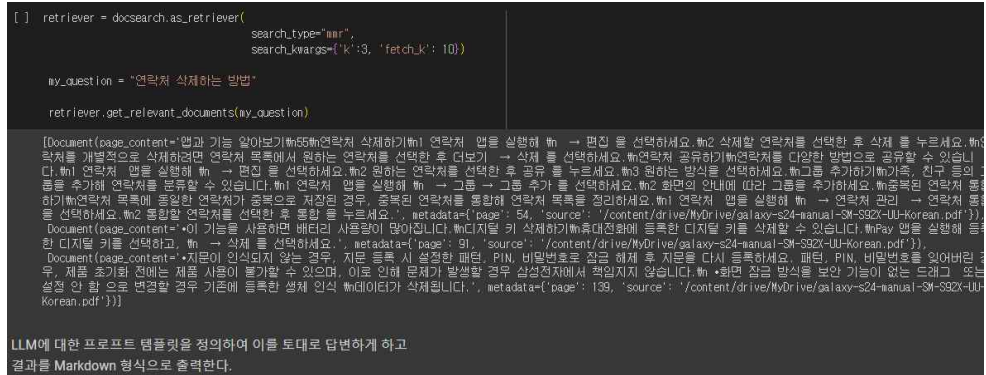
import os
os.environ["GOOGLE_API_KEY"] = "AIzaSyA5zRl4juLkpvzre*****"

from langchain_google_genai import ChatGoogleGenerativeAI
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.vectorstores import Chroma
from langchain.document_loaders import PyPDFLoader
from langchain.text_splitter import CharacterTextSplitter

loader =
PyPDFLoader("/content/drive/MyDrive/galaxy-s24-manual-SM-S92X-UU-K
orean.pdf")
pages = loader.load_and_split()

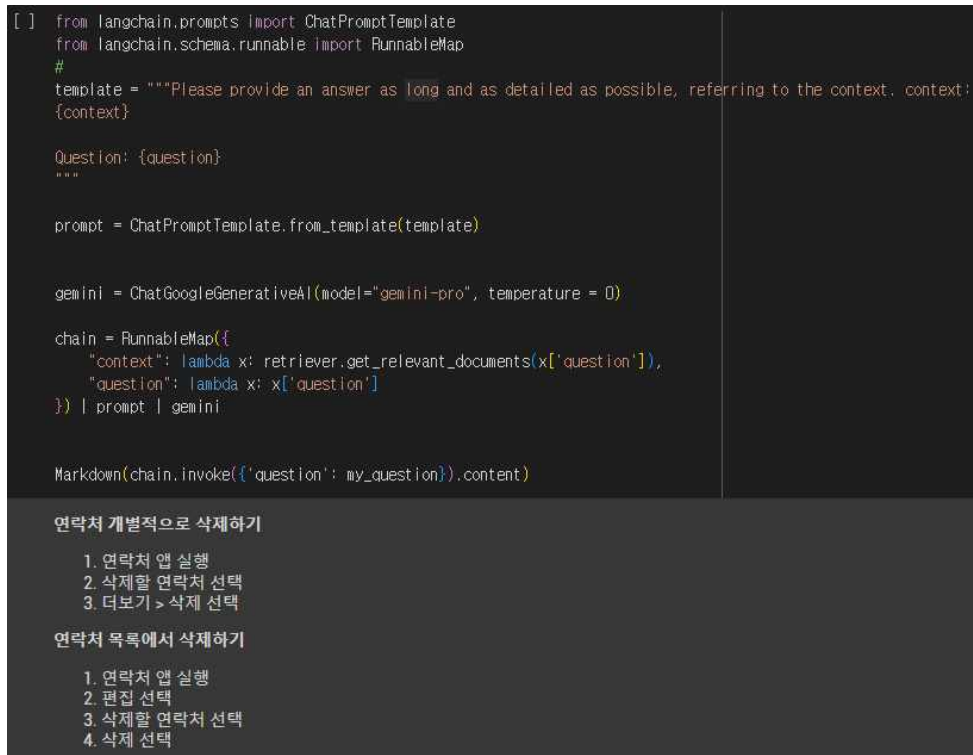
text_splitter = RecursiveCharacterTextSplitter(chunk_size=256,
                                                chunk_overlap=20)
texts = text_splitter.split_documents(pages)
```

[그림 2-23]과 같이 시스템의 효율성을 테스트하기 위해 ‘연락처 삭제 방법’을 질의했고, 시스템이 등록된 매뉴얼에서 답변을 검색하는 과정을 표시하도록 했다.



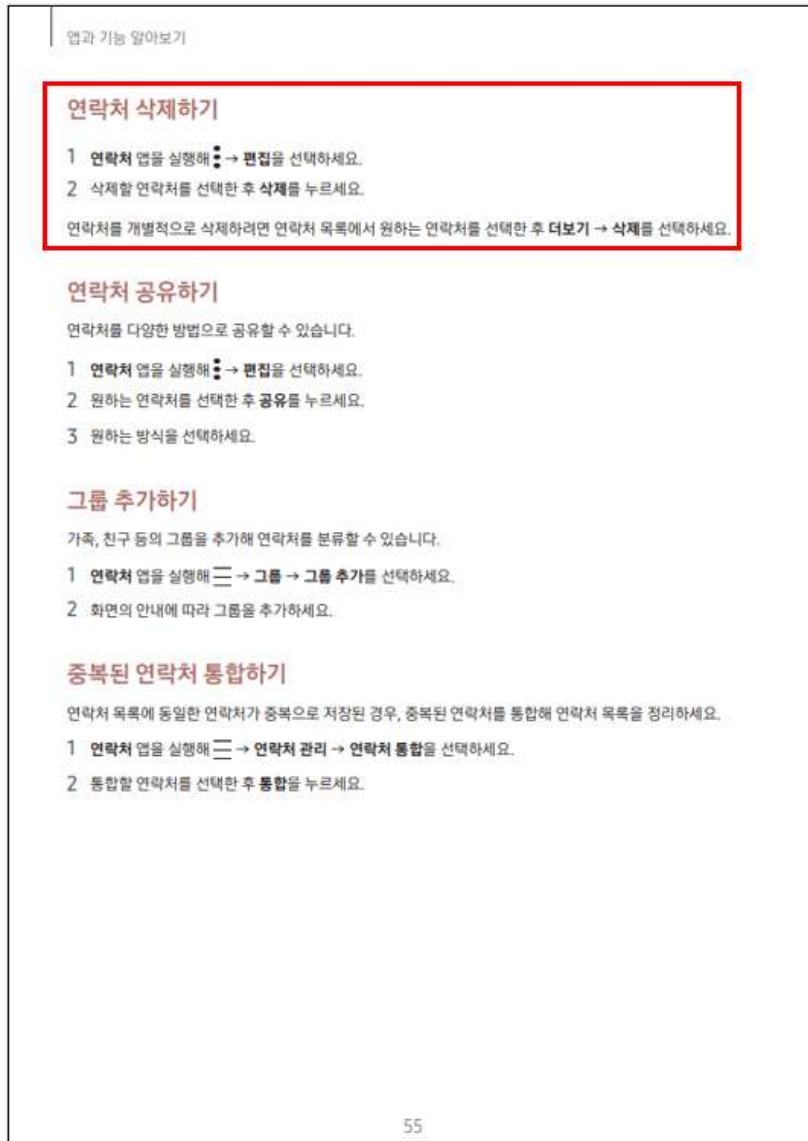
[그림 2-23] 질의 및 RAG 내부 검색 결과

[그림 2-24]와 같이 시스템은 매뉴얼의 정확한 내용을 바탕으로 상세하고 정확한 답변을 제공하였으며, 이는 실제 설비 매뉴얼과 과거 수리 이력을 활용할 경우 유사한 수준의 정확성과 상세함을 기대할 수 있음을 시사한다.



[그림 2-24] 초거대 언어 모델을 통한 답변 처리

[그림 2-25]의 내용과 같이 초거대 언어 모델을 통한 답변의 처리가 정확함을 확인할 수 있다.



[그림 2-25] 매뉴얼의 실제 내용

2.5.2.4 타당성 및 가치

이 예비 실험의 결과는 확인된 유지 관리 문제를 해결하기 위해 초거대 언어 모델 및 RAG를 사용하는 타당성과 잠재적 가치를 확인했다. 이

시스템의 실제 적용 시 다음과 같은 구체적인 효과를 예상할 수 있다.

① 향상된 효율성

시스템은 유지 관리 문의에 빠르고 정확하게 응답하여 가동 중지 시간을 크게 줄이고 운영 효율성을 향상시킬 수 있다. 이는 사후보전뿐만 아니라 계획 보전 및 예방 보전에도 적용된다.

② 정확도 향상

기술 매뉴얼에서 직접 제공되는 상세하고 정확한 지침으로 유지 관리 절차에서 인적 오류를 최소화할 수 있다.

③ 정보를 바탕으로 한 의사 결정

신뢰할 수 있는 상황별 정보를 제공하여 경험이 부족한 유지 관리 담당자도 충분한 정보를 바탕으로 결정을 내릴 수 있다.

2.5.2.5 연구방향

향후 연구는 다음과 같은 방향으로 진행할 계획이다

① 데이터 다양성 확대

RAG 모델을 교육하기 위해 광범위한 유지 관리 매뉴얼과 기술 문서를 통합하여 다양한 유형의 시설에 걸쳐 다양한 쿼리를 처리할 수 있도록 한다.

② 모델 개선

초거대 언어 모델 및 RAG 모델을 지속적으로 개선하여 특히 복잡하고 미묘한 시나리오에서 이해 및 검색 기능을 향상시킨다.

③ 통합 및 배포

사용자 친화적인 인터페이스를 개발하고 시스템을 기존 유지 관리 위

크플로에 통합하여 유지 관리 담당자가 원활하게 채택하고 사용할 수 있도록 한다.

④ 성능 평가

실제 환경에서 포괄적인 테스트와 평가를 수행하여 시스템이 유지 관리 작업에 미치는 영향을 측정하고 추가 개선이 필요한 영역을 식별한다.

초거대 언어 모델과 RAG를 사용한 예비 실험에서는 이러한 기술을 활용하면 사후보전 시 직면한 문제를 효과적으로 해결할 수 있음이 입증되었다. 숙련된 인력 없이도 문제를 진단하고 해결하는 데 필요한 시간을 줄이는 시스템 기능은 상당한 이점을 제공한다. 앞으로 본 연구는 다양한 설비 전반에 걸쳐 유지보수 작업을 향상시킬 수 있는 강력하고 효율적이며 사용자 친화적인 시스템을 개발하는 것을 목표로 할 것이다.

Ⅲ. 연구설계 및 구현

3.1 실험 환경 및 구성

3.1.1 연구설계

사후보전의 문제점을 파악하기 위해 [그림 3-1]과 같이 A사의 협조를 통해 2023년 1월에서 2023년 12월까지 1년간 761개의 사후보전의 내용을 제공받아 비가동 유형별로 횟수와 비율, 누적 비가동 시간, 평균 비가동 시간, 전체 비가동에서 차지하는 비율을 분석하였다.

1	점검구분	클린트 작업장 작업라인	고장설비	공정	고장위치	고장 시작일자	고장 시작시간	접수일자	접수시간	도착일자	도착시간	부탁 시작일자	부탁 시작시간	부탁 완료일자	완료시 간	고장코드	고장분류
2	JM:사후보				Cover 가조립 & 놓이림/	2023-01-03	09:15	2023-01-03	09:15	2023-01-03	09:15	2023-01-03	09:15	2023-01-03	09:30	04: 작동불	M: 기계고장
3	JM:사후보				Cover 가조립 & 놓이림/	2023-01-03	15:50	2023-01-03	15:50	2023-01-03	15:50	2023-01-03	15:50	2023-01-03	16:05	14: 수정	M: 기계고장
4	JM:사후보				Grease 도포 Unit(스핀들	2023-01-03	08:30	2023-01-03	08:30	2023-01-03	08:30	2023-01-03	08:30	2023-01-03	10:00	17: 교체	M: 기계고장
5	JM:사후보				Pin 압입	2023-01-04	14:10	2023-01-04	14:10	2023-01-04	14:10	2023-01-04	14:10	2023-01-04	15:00	14: 수정	M: 기계고장
6	JM:사후보				재갈 Unit	2023-01-05	09:00	2023-01-05	09:00	2023-01-05	09:00	2023-01-05	09:00	2023-01-05	11:00	14: 수정	M: 기계고장
7	JM:사후보				순차제어 Unit	2023-01-06	16:50	2023-01-06	16:50	2023-01-06	16:50	2023-01-06	16:50	2023-01-06	17:05	18: 청소	M: 기계고장
8	JM:사후보				JIG	2023-01-10	08:40	2023-01-10	08:40	2023-01-10	08:40	2023-01-10	08:40	2023-01-10	10:40	17: 교체	M: 기계고장
9	JM:사후보				JIG	2023-01-11	08:30	2023-01-11	08:30	2023-01-11	08:30	2023-01-11	08:30	2023-01-11	12:30	01: 마모	M: 기계고장
10	JM:사후보				재갈 Unit	2023-01-11	14:00	2023-01-11	14:00	2023-01-11	14:00	2023-01-11	14:00	2023-01-11	14:20	17: 교체	E: 전기고장
11	JM:사후보				Riveting Unit	2023-01-11	12:10	2023-01-11	12:10	2023-01-11	12:10	2023-01-11	12:10	2023-01-11	12:20	14: 수정	M: 기계고장
12	JM:사후보				Robot #2	2023-01-16	16:00	2023-01-16	16:00	2023-01-16	16:00	2023-01-16	16:00	2023-01-16	16:30	14: 수정	M: 기계고장
13	JM:사후보				재갈 Unit	2023-01-16	12:20	2023-01-16	12:20	2023-01-16	12:20	2023-01-16	12:20	2023-01-16	14:40	13: 끼임	M: 기계고장
14	JM:사후보				Robot #2	2023-01-17	10:20	2023-01-17	10:20	2023-01-17	10:20	2023-01-17	10:20	2023-01-17	11:40	05: 단락	M: 기계고장
15	JM:사후보				Torque 측정 Unit	2023-01-18	13:42	2023-01-18	13:42	2023-01-18	13:42	2023-01-18	13:42	2023-01-18	15:00	04: 작동불	E: 전기고장
16	JM:사후보				이송 Unit	2023-01-20	09:20	2023-01-20	09:20	2023-01-20	09:20	2023-01-20	09:20	2023-01-20	10:40	17: 교체	M: 기계고장
17	JM:사후보				Grease 도포 Unit(스핀들	2023-01-20	08:50	2023-01-20	08:50	2023-01-20	08:50	2023-01-20	08:50	2023-01-20	09:35	17: 교체	M: 기계고장
18	JM:사후보				재갈 Unit	2023-01-25	21:40	2023-01-25	21:40	2023-01-25	21:40	2023-01-25	21:40	2023-01-25	22:20	17: 교체	M: 기계고장
153	JM:사후보				SC3 이송 unit	2023-12-28	04:52	2023-12-28	04:52	2023-12-28	04:52	2023-12-28	04:52	2023-12-28	05:03	14: 수정	M: 기계고장
153	JM:사후보				SC2 이송 Unit	2023-12-28	03:33	2023-12-28	03:33	2023-12-28	03:33	2023-12-28	03:33	2023-12-28	03:41	18: 청소	M: 기계고장
154	JM:사후보				ST4 Adapter 공급 Unit	2023-12-28	23:22	2023-12-28	23:22	2023-12-28	23:22	2023-12-28	23:22	2023-12-28	23:38	13: 끼임	M: 기계고장
155	JM:사후보				Grease 도포 Unit	2023-12-29	02:48	2023-12-29	02:48	2023-12-29	02:48	2023-12-29	02:48	2023-12-29	03:09	13: 끼임	M: 기계고장
156	JM:사후보				ST4 Adapter 공급 Unit	2023-12-29	11:05	2023-12-29	11:05	2023-12-29	11:05	2023-12-29	11:05	2023-12-29	11:15	16: 점검	M: 기계고장
157	JM:사후보				롤타이트 도포 Unit	2023-12-29	13:10	2023-12-29	13:10	2023-12-29	13:10	2023-12-29	13:10	2023-12-29	13:30	14: 수정	M: 기계고장
158	JM:사후보				소재결합 Unit	2023-12-29	22:03	2023-12-29	22:03	2023-12-29	22:03	2023-12-29	22:03	2023-12-29	22:16	13: 끼임	M: 기계고장
159	JM:사후보				Greasing 도포 unit	2023-12-29	15:10	2023-12-29	15:10	2023-12-29	15:10	2023-12-29	15:10	2023-12-29	22:14	17: 교체	M: 기계고장
160	JM:사후보				ST7 Robot	2023-12-30	02:15	2023-12-30	02:15	2023-12-30	02:15	2023-12-30	02:15	2023-12-30	02:25	14: 수정	M: 기계고장
161	JM:사후보				검사가 1호	2023-12-30	05:47	2023-12-30	05:47	2023-12-30	05:47	2023-12-30	05:47	2023-12-30	05:52	13: 끼임	M: 기계고장
162	JM:사후보				Laser marking unit	2023-12-30	02:06	2023-12-30	02:06	2023-12-30	02:06	2023-12-30	02:06	2023-12-30	02:12	13: 끼임	M: 기계고장

[그림 3-1] A사의 사후보전 처리 현황 (A사, 2023; 연구자 재구성, 2024)

아래의 [표 3-1]과 같이 A사의 2023년 1월에서 12월까지의 전체 사후보전 이력을 집계해 고장 유형별로 분류하였다. 우선 고장 횟수와 고장유형이 전체 사후보전에서 차지하는 비율, 유형별 누적 비가동 시간, 유형별 평균 비가동 시간, 전체 비가동 시간에서 유형별 누적 비가동 시간이 차지하는 비율을 분류한 후 고장 비율과 비가동 시간 비율과의 대비성을 분석하였다.

[표3-1] 2023년도 A사 사후보전 이력 현황 (A사, 2023; 연구자 재구성, 2024)

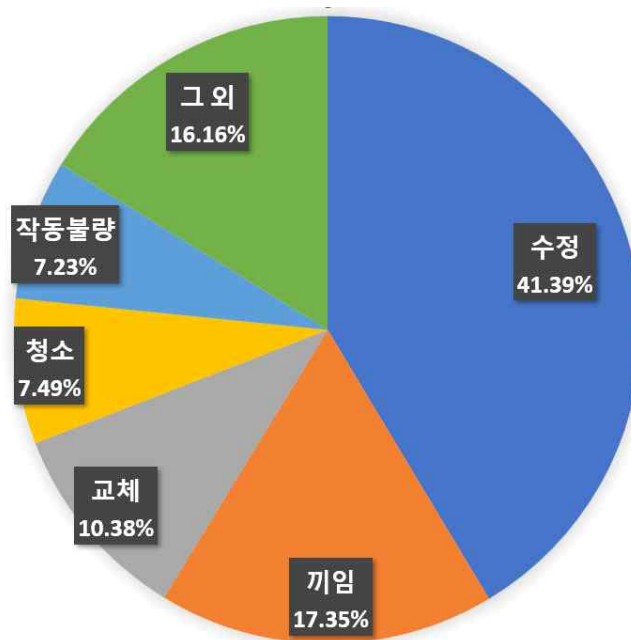
고장유형	횟수	고장비율 (%)	누적 비가동 시간(분)	평균 비가동 시간(분)	비가동 시간 비율
마모	5	0.66%	365	73.00	0.50%
파손	19	2.50%	1,650	86.84	2.26%
작동불량	55	7.23%	3,216	58.47	4.41%
단락	11	1.45%	1,071	97.36	1.47%
유동	32	4.20%	1,289	40.28	1.77%
막힘	2	0.26%	310	155.00	0.43%
누(설)유	8	1.05%	455	56.88	0.62%
끼임	132	17.35%	3,671	27.81	5.04%
수정	315	41.39%	53,196	168.88	73.01%
가공	2	0.26%	41	20.50	0.06%
점검	27	3.55%	494	18.30	0.68%
교체	79	10.38%	5,546	70.20	7.61%
청소	57	7.49%	869	15.25	1.19%
PLC, PCB이상	2	0.26%	195	97.50	0.27%
기타	15	1.97%	491	32.73	0.67%
합계	761		72,859		

[표 3-2]와 [그림 3-2]는 비가동 중에서 고장 횟수의 비율이 5%를 초과하는 비가동은 수정, 끼임, 교체, 청소, 작동 불량이며, 이들이 차지하는 비율은 83.84%에 달하는 것으로 나타났으며 그 외 마모, 단락, 누설, 누유, 막힘 등과 같은 고장의 비율은 16.16%를 차지하였다.

[표 3-2] 2023년도 유형별 고장 비율 (A사, 2023; 연구자 재구성, 2024)

고장유형	수정	끼임	교체	청소	작동 불량	그 외
횟수	315	132	79	57	55	104
고장 비율(%)	41.39%	17.35%	10.38%	7.49%	7.23%	16.16%

[그림 3-2]와 같이 그래프로 표현하면 수정, 끼임, 교체, 청소, 작동 불량
의 비율에서 84.84%를 차지하고 있다.



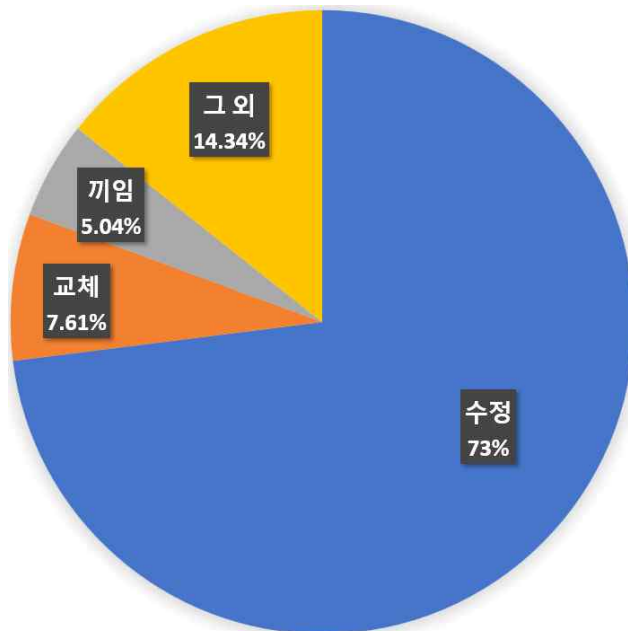
[그림 3-2] 2023년도 유형별 고장 비율 (A사, 2023; 연구자 재구성, 2024)

[표 3-3]과 같이 전체 누적 비가동 시간에서 차지하는 유형별 누적 비가
동 시간 비율이 5%를 초과하는 유형을 정리했다. 고장유형 중 수정의 비가
동 시간 비율은 73.01%, 교체는 7.61%, 끼임과 그 외 비가동은 각각
5.04%와 14.34%를 차지하고 있다.

[표 3-3] 2023년도 유형별 누적 비가동 시간 비율 (A사, 2023; 연구자 재구성, 2024)

고장유형	수정	교체	끼임	그 외
누적 비가동 시간(분)	53,196	5,546	3,671	10,446
비가동 시간 비율	73.01%	7.61%	5.04%	14.34%

[그림 3-3]과 같이 그래프로 표현하면 수정, 교체, 끼임이 전체 누적 비가동 시간의 비율에서 차지하는 영역은 85.66%이다. 누적 비가동 시간을 기준으로 했을 때는 특정 비가동 유형에 더욱 집중되는 것을 확인할 수 있다.



[그림 3-3] 2023년도 유형별 누적 비가동 시간 비율 (A사, 2023; 연구자 재구성, 2024)

A사에서는 단순 교체 또는 수정으로 인한 비가동 시간을 10분을 기준으

로 하고 있어, 비가동 이력에서 20분을 초과하는 비가동 내역을 중심으로 동일 항목에 대해 비가동 시간의 차이가 많은 것을 확인해 보았다.

[표 3-4]와 같이 같은 설비에서 발생한 같은 고장 현상, 원인, 그리고 수리 내역에 대한 사후보전 이력을 고장 현상별로 비교분석 하였다. 이 데이터를 분석하면 같은 문제에 대한 대응 시간이 크게 차이가 나는 것을 확인할 수 있다. 예를 들어 'Spring Ass'y M/C' 설비의 'SC1, 2 간섭 발생 문제'에 대해 소요 시간이 10분에서 60분까지 다양하게 나타나고 있으며 'St9 제품 이송 실린더 전진 후진 알람 발생' 문제에 대해서는 소요 시간이 10분에서 230분까지 나타나 최소 2배에서 많게는 20배 이상의 차이를 보이는 것을 확인할 수 있다.

[표 3-4] 동일한 문제에 대한 보전 이력 (A사, 2023; 연구자 재구성, 2024)

설비 : Spring Ass'y M/C, 공정 : Spring Assembly, 위치: 체결 Unit					
발생 시간	완료 시간	소요 시간	현상	원인	수리내역
08-07 14:00	08-07 14:10	10	제품 조립 시 SC1, 2 간섭 발생	중간 클램프 위치 맞지 않음 확인	중간 클램프 위치 수정
08-07 22:50	08-07 23:50	60			
08-08 08:40	08-08 09:40	60			
08-08 21:40	08-08 22:30	50			
08-11 09:00	08-11 09:10	10			
08-15 08:45	08-15 09:10	25			
설비 : Spring greasing, 공정 : Greasing, 위치: Washer 체결 unit					
10-20 07:07	10-20 07:09	2	SC2 와셔 감지 NG 발생	근접센서 위치 이상 확인	근접센서 위치 조정
11-08 08:30	11-08 08:54	24			
설비 : Spring Ass'y M/C, 공정 : Spring Assembly, 위치: 체결 Unit					

08-07 14:00	08-07 14:10	10	제품 조립 시 SC1, 2 간섭 발생	중간 클램프 위치 맞지 않음 확인	중간 클램프 위치 수정
08-07 22:50	08-07 23:50	60			
설비 : Spring ass'y & EOL, 공정 : Spring Assembly & EOL, 위치: Lotation table unit					
08-29 19:14	08-29 19:21	7	인텍스 St4 가이드서포 트 기구부 유동	가이드서 포트 기구부 볼트 풀림 확인	기구부 재고정
11-15 15:30	11-15 16:00	30			
설비 : Brake module CG M/C, 공정 : Brake module CG, 위치: 이송 & INDEX Unit					
02-25 12:20	02-25 12:30	10	St9 제품 이송 실린더 전진 후진 알람 발생	전후진 오토 스위치 고장 확인	오토 스위치 교체
06-07 11:40	06-07 15:30	230			
설비 : Spring ass'y & EOL, 공정 : Spring Assembly & EOL, 위치: Lotation table unit					
02-25 15:40	02-25 16:20	40	너트런너 전후진 알람 발생	전후진 오토스위 치 고장 확인	오토스위 치 교체
03-16 21:00	03-16 23:10	130			
10-30 08:45	10-30 09:55	70			

이러한 차이는 크게 두 가지 원인에서 비롯된다는 것을 알 수 있다. 문제의 원인을 파악하는 데 걸리는 시간이 다르거나 복구를 위한 조치 시간이 다르다는 것이다. 즉, 설비보전 담당자의 개인적 역량과 경험의 차이가 큰 영향을 미치게 된다. 동일한 문제에 대해 처리 시간이 크게 차이가 난다는 것은 생산성이 저하되고 효율성 측면에서 심각한 불균형을 초래할 수 있으므로 제조기업의 입장에서는 큰 문제가 될 수 있다.

그러나 현재까지 이러한 문제에 대한 체계적인 연구는 매우 제한적으로 이루어졌다. A사의 보전 이력의 시간 기록으로는 문제 원인 파악에 걸린 시간과 실제 수리에 소요된 시간의 정확성이 보장되지 않는다는 것을 알 수 있다. 그럼에도 불구하고 담당자의 숙련도에 따라 소요 시간이 크게 좌우된다는 것을 인정하고 있다는 게 현실이다.

이러한 상황에서 AI 기반의 사후보전 시스템은 숙련자와 비숙련자 간의 소요 시간 차이를 줄일 수 있는 획기적인 방법이 될 수 있다. 이 시스템은 과거의 모든 보전 이력과 관련 기술 문서를 학습하여, 문제 발생 시 신속하게 가능한 원인과 해결 방안을 제시할 수 있다. 이를 통해 비숙련자도 숙련자 수준의 문제 진단과 해결이 가능해지며, 전반적인 보전 작업의 효율성과 일관성을 크게 향상시킬 수 있다.

더불어 이 시스템은 지속적으로 새로운 사례를 학습하고 업데이트되므로, 시간이 지날수록 더욱 정확하고 효과적인 해결책을 제시할 수 있게 된다. 결과적으로, 설비 가동 중단 시간을 최소화하고 생산성을 향상시키는 데 크게 기여할 수 있을 것이다. 이는 특히 숙련된 인력 확보에 어려움을 겪는 중소 기업에게 큰 도움이 될 수 있는 혁신적인 접근 방식이 될 것이다.

3.1.2 실험 환경 및 구성

실험 환경을 로컬에 구성하기 위해 [표3-5]와 같이 서버를 구성하고 LLM 및 RAG 시스템을 구성했다. 운영 모델은 Llama 3을 사용하였다.

[표 3-5] 실험 시스템의 구성

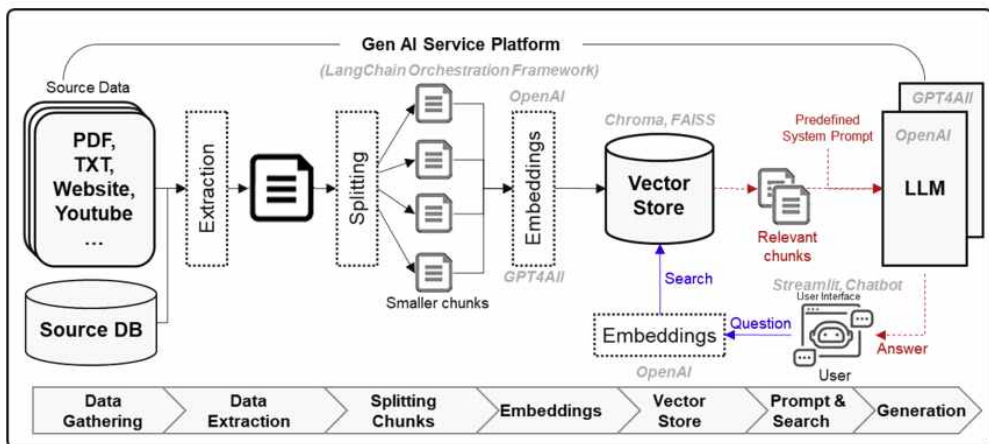
구분	항목
CPU	AMD 라이젠 스레드리퍼 3960X 24Core 48Thread
Meory	삼성전자 DDR4 128GB
VGA	Geforce RTX 3090 24GB
SSD	삼성전자 m.2 NVME 2TB * 4 Raid
HDD	Western Digital Ultrastar 7200rpm 18TB * 2
OS	Ubuntu Linux Server 22.04LTS
운영 모델	Llama 3
사용 언어	Python 3.11

3.1.3 Llama 3

Llama 3은 페이스북으로 유명한 Meta에서 개발한 오픈 소스 초거대 언어 모델이다. Llama 3은 700억 개의 파라미터를 보유하고 있으며, 요구 성능이 낮아 컴퓨팅의 효율성이 높게 설계되어 있다. Llama 3은 유연한 확장성을 지원하기 때문에 소규모 애플리케이션에도 적합한 장점이 있다. 그리고 사전 학습된 언어 모델과 함께 전체 소프트웨어를 고객사 내부에 직접 구축할 수 있어 안정적인 성능의 유지가 가능하고, 민감한 데이터가 고객사 인프라 내에 유지되기 때문에 데이터 침해 및 무단 액세스의 위험이 줄어든다. 또한 고객사의 요구나 특징에 따라 미세 조정 및 맞춤화를 할 수 있고, 기타 내부 시스템과 통합을 진행할 수 있는 장점이 있다. Llama 3은 현재 지마켓 등의 여러 IT 기업에서 내부에 구축하여 실제 업무에 활용하고 있다.

3.1.4 시스템의 구성

먼저 정보를 보관하고 있는 위키피디아와 같은 구조화된 데이터베이스나 매뉴얼 등의 문서를 참조할 수 있도록 지식 베이스를 구축한다. 사용자가 질의를 하면 초거대 언어 모델이 먼저 내용을 분석해서 RAG 체인을 실행한다. RAG 체인은 필요한 정보를 지식 베이스에서 문장이나 어휘 단위로 쪼개어 검색하고, 검색한 내용을 초거대 언어 모델의 입력에 추가해 답변을 생성한다. 이렇게 생성한 답변을 사용자에게 전달한다. 전체 시스템의 구성도는 [그림3-4]와 같다.



[그림 3-4] RAG기반 생성형 AI 서비스 구현 플랫폼 (정천수, 2023)

3.1.5 코드의 구현

LLM과 RAG를 구현한 시스템은 Python 3.11로 작성하였으며, 구현한 코드는 아래와 같다.

PDF 문서 등 문서를 읽어들이고 잘라서 벡터화로 변환하여 저장할 수 있도록 관련 라이브러리를 불러온다.

```
from langchain.text_splitter import RecursiveCharacterTextSplitter
from langchain.vectorstores import Chroma
from langchain.document_loaders import PyPDFLoader
from langchain.text_splitter import CharacterTextSplitter
```

load_documents 함수를 통해 경로에 있는 모든 PDF 문서를 읽어들이고, split_documents 함수로 문서를 잘라 RAG에서 불러들일 수 있도록 처리한다.

```
def load_documents():
    document_loader = PyPDFDirectoryLoader(DATA_PATH)
    return document_loader.load()

def split_documents(documents: list[Document]):
    text_splitter = RecursiveCharacterTextSplitter(
        chunk_size=800,
        chunk_overlap=80,
        length_function=len,
        is_separator_regex=False,
    )
    return text_splitter.split_documents(documents)
```

앞에서 처리한 문서를 add_to_chroma 함수를 통해 정의된 chunk의 사이즈를 기준으로 문장을 잘라 벡터로 변환한다.

```
def add_to_chroma(chunks: list[Document]):
    db = Chroma(
        persist_directory=CHROMA_PATH,
        embedding_function=get_embedding_function()
    )

    chunks_with_ids = calculate_chunk_ids(chunks)
```

```

existing_items = db.get(include=[])
existing_ids = set(existing_items["ids"])
print(f"현재 DB에 있는 문서의 수: {len(existing_ids)}")

new_chunks = []
for chunk in chunks_with_ids:
    if chunk.metadata["id"] not in existing_ids:
        new_chunks.append(chunk)

if len(new_chunks):
    print(f"새 문서 추가: {len(new_chunks)}")
    new_chunk_ids = [chunk.metadata["id"] for chunk in new_chunks]
    db.add_documents(new_chunks, ids=new_chunk_ids)
    db.persist()
else:
    print("추가된 문서가 없음")

def calculate_chunk_ids(chunks):
    last_page_id = None
    current_chunk_index = 0

    for chunk in chunks:
        source = chunk.metadata.get("source")
        page = chunk.metadata.get("page")
        current_page_id = f"{source}:{page}"

        if current_page_id == last_page_id:
            current_chunk_index += 1
        else:
            current_chunk_index = 0

        chunk_id = f"{current_page_id}:{current_chunk_index}"
        last_page_id = current_page_id

        chunk.metadata["id"] = chunk_id

    return chunks

```

RAG에 질의를 하여 응답을 받아서 처리한다.

```

def query_rag(query_text: str):
    embedding_function = get_embedding_function()

```



```

db = Chroma(persist_directory=CHROMA_PATH,
            embedding_function=embedding_function
            )

results = db.similarity_search_with_score(query_text, k=5)

context_text = "WnWn--WnWn".join([doc.page_content for doc,
                                   _score in results])

prompt_template = ChatPromptTemplate.from_template(PROMPT_TEMPLATE)

prompt = prompt_template.format(context=context_text,
                                question=query_text
                                )

response_text = model.invoke(prompt)

sources = [doc.metadata.get("id", None) for doc, _score in results]
formatted_response = f"Response: {response_text}WnSources: {sources}"

return response_text

```

3.1.6 실험 데이터 생성

정확한 학습을 위해 엑셀로 된 사후보전 이력 데이터를 설비명, 공정, 고장 현상, 고장 원인, 수리 내역을 발췌하여 시나리오를 작성하여 [그림 3-5]와 같은 내용으로 PDF 파일을 생성하였다.

질의 : GN7 파워 프레임 CO2 용접, 레그 브라켓 후단 와이어 직경 변경
 응답 : E.O 변경이 원인. CO2 안착 블록 수정

질의 : GN7 파워 프레임 CO2 용접, 백프레임 브라켓 서브 지그 이젝터 작동시 브라켓 구부러짐 발생
 응답 : 기준핀 끼임 발생이 원인. 기준핀 가공

질의 : GN7 PWR FRM CO2, RECLILING 레그 간 거리 불량
 응답 : LH,RH 간 거리 수정

질의 : MQ4 메카 조립라인, 파렛트 스톱퍼 브라켓 파손.
 응답 : 스톱퍼 브라켓의 적정 하중 초과로 인한 충격 파손이 원인. 파손부 용접 및 보강 BAR 추가 용접.

[그림 3-5] 학습용 시나리오

[그림 3-6]과 같이 일부 단어에 대해 다르게 표현할 수 있는 유사한 단어들로 변경해 시나리오를 추가하였다.

질의 : GN7 파워 프레임 CO2 용접, 레그 브라켓 후단 와이어 직경 **변경**
 응답 : E.O 변경이 원인. CO2 안착 블록 수정

질의 : GN7 파워 프레임 CO2 용접, 레그 브라켓 후단 와이어 직경 **수정**
 응답 : E.O 변경이 원인. CO2 안착 블록 수정

질의 : GN7 파워 프레임 CO2 용접, 백프레임 브라켓 서브 지그 이젝터 작동시 브라켓 **구부러짐** 발생
 응답 : 기준핀 끼임 발생이 원인. 기준핀 가공

질의 : GN7 파워 프레임 CO2 용접, 백프레임 브라켓 서브 지그 이젝터 작동시 브라켓 **휨** 발생
 응답 : 기준핀 끼임 발생이 원인. 기준핀 가공

질의 : GN7 PWR FRM CO2, RECLILING 레그 간 거리 **불량**
 응답 : LH,RH 간 거리 수정

질의 : GN7 PWR FRM CO2, RECLILING 레그 간 거리 **NG**
 응답 : LH,RH 간 거리 수정

질의 : MQ4 메카 조립라인, 파렛트 스톱퍼 브라켓 **파손**
 응답 : 스톱퍼 브라켓의 적정 하중 초과로 인한 충격 파손이 원인. 파손부 용접 및 보강 BAR 추가 용접.

질의 : MQ4 메카 조립라인, 파렛트 스톱퍼 브라켓 **마모**
 응답 : 스톱퍼 브라켓의 적정 하중 초과로 인한 충격 파손이 원인. 파손부 용접 및 보강 BAR 추가 용접.

[그림 3-6] 표현에 따른 추가 시나리오

IV. 연구 결과

4.1 실험 결과 및 분석

4.1.1 실험 결과

생성한 시나리오를 서버에 등록하고, 시스템을 운영하여 아래와 같이 질의하였다.

① 질의 : EPL-M-001에서 리테이너 너트가 삽입이 안됨

응답 :

****고장 원인: 체결 톨 이물질 오염****

체결 톨이 이물질로 오염되면 정상적인 작동이 방해될 수 있습니다. 이물질은 체결 톨 내부의 움직이는 부품에 끼어들어 마찰과 결함을 유발할 수 있습니다.

****수리 내역: 체결 톨 청소****

체결 톨의 이물질 오염 문제를 해결하려면 톨을 철저히 청소해야 합니다. 이는 다음 단계를 포함합니다.

- * 톨을 전원에서 분리합니다.
- * 모든 부착물과 액세서리를 제거합니다.
- * 압축 공기 또는 부드러운 솔을 사용하여 톨에서 먼지와 이물질을 제거합니다.
- * 필요한 경우 청소용 용액 또는 윤활유를 사용하여 톨의 움직이는 부품을 청소합니다.
- * 톨을 조립하고 다시 전원에 연결합니다.

이러한 단계를 수행하면 체결 톨의 이물질 오염이 제거되고 정상적인 작동이 복구됩니다.

② 질의 : EPL-M-003에서 바코드 리딩 불가능 현상

응답 :

****고장 원인****

이 기기는 프로그램에 이상이 발생하여 정상적으로 작동하지 못했습니다.

****수리 내역****

기술자가 기기의 프로그램을 재시작하여 이전에 발생했던 이상을 해결했습니다. 이를 통해 기기가 다시 정상적으로 작동할 수 있게 되었습니다.

③ 질의 : EPL-M-006에서 설비 리노핀 부서짐

응답 :

****고장 원인:****

바코드 리더의 커넥터가 체결하는 동안 파손되었습니다. 이로 인해 리더가 스캐너에 신호를 전송할 수 없게 되었습니다.

****수리 내역:****

1. 파손된 하우징을 제거했습니다.
2. 새로운 하우징을 리더에 납땜했습니다.
3. 커넥터를 새로운 하우징에 다시 체결했습니다.

수리 후 바코드 리더는 정상적으로 작동하고 있으며 제대로 스캔할 수 있습니다.

④ 질의 : EPL-M-008에서 피더 디스크 공급 불가 발생
 응답 :

****고장 원인:****
 장치의 공급부에 이물질이 발견되었습니다. 이 이물질은 장치의 정상적인 작동을 방해하여 고장을 일으켰습니다.
****수리 내역:****
 고장을 해결하려면 다음 단계를 수행했습니다.
 1. 장치를 분해하고 공급부에 접근했습니다.
 2. 이물질을 신중하게 제거했습니다.
 3. 공급부를 조립하고 장치를 재조립했습니다.
 이물질을 제거한 후 장치는 정상적으로 작동하는 것으로 확인되었습니다.

[표 4-1] 시스템 적용 전과 적용 후 증감 비교

고장유형	횟수 (2024년3~5월)	고장비율 (%)	평균 비가동 시간(분)		
			적용전 (2023년기준)	적용후 (2024년3~5월)	증감(%)
마모	2	1.04%	73	63.50	-13.0%
파손	4	2.08%	86.84	88.25	1.6%
작동불량	14	7.29%	58.47	36.50	-37.6%
단락	3	1.56%	97.36	101.33	4.1%
유동	7	3.65%	40.28	35.57	-11.7%
막힘	1	0.52%	155	159.00	2.6%
누(설)유	3	1.56%	56.88	46.33	-18.5%
끼임	31	16.15%	27.81	20.42	-26.6%
수정	79	41.15%	168.88	101.43	-39.9%
가공	0	0.00%	20.5	0.00	0.0%
점검	8	4.17%	18.3	16.88	-7.8%
교체	16	8.33%	70.2	44.44	-36.7%
청소	19	9.90%	15.25	14.74	-3.4%
PLC, PCB이상	0	0.00%	97.5	0.00	0.0%
기타	5	2.60%	32.73	29.40	-10.2%
합계	192	100%	18,138.3	11,761.0	-37.7%

위 [표 4-1]에서 2024년 3월~5월의 3개월간 발생한 비가동 내역과 각

각의 비가동 평균 시간(분)을 정리하여 표시하였다. 또한 시스템 적용 전과 비교하기 위해 2023년 1월~2023년 12월까지의 평균 비가동 시간(분)을 같이 표시하였으며, 2023년과 2024년의 비가동 시간의 차이를 비율로 표시하였다.

전체 내역 중 2023년에 비해 평균 비가동 시간이 감소한 항목은 마모, 작동불량, 유동 등 10가지이며, 비가동 시간이 증가한 항목은 파손, 단락, 막힘의 3가지이다.

시스템 적용 후, 전체적으로 비가동 시간이 37.7% 감소하여 18,138.3분에서 11,761분으로 줄어들었다. 이는 대부분의 유형에서 개선이 있었음을 나타낸다. 가장 큰 개선을 보인 유형은 '수정'으로, 전체 고장의 41.15%를 차지하며 평균 비가동 시간이 39.9% 감소했다. '작동불량'과 '교체' 유형도 각각 37.6%와 36.7%의 큰 개선을 보였다. '끼임' 유형은 두 번째로 빈번하게 발생하며(16.15%), 26.6%의 비가동 시간 감소를 보였다. '누(설)유'(18.5% 감소), '마모'(13% 감소), '유동'(11.7% 감소) 유형에서도 상당한 개선이 있었다. '기타' 유형은 10.2% 감소했으며, '점검'과 '청소' 유형도 각각 7.8%와 3.4%의 개선을 보였다.

반면, 일부 유형에서는 비가동 시간이 증가했다. '파손' 유형은 1.6%, '단락' 유형은 4.1%, '막힘' 유형은 2.6% 증가했다. 이러한 증가는 전체적인 개선에 비해 상대적으로 작은 폭이지만, 주의가 필요한 부분이다. '가공'과 'PLC, PCB이상' 유형은 2024년 3월~5월 기간 동안 발생하지 않았다. 이는 이전 기간에 비해 개선되었거나, 해당 기간 동안 관련 문제가 없었음을 의미할 수 있다.

종합적으로, 대부분의 유형에서 비가동 시간이 감소했으며, 특히 자주 발생하는 유형들에서 큰 개선이 있었다. 일부 유형에서의 소폭 증가에도 불구하고, 전체적인 비가동 시간이 크게 줄어든 것은 시스템 적용이 효과적이었다는 것을 보여준다.

그러나 [표 4-2]을 기준으로 대응표본 t-검정을 수행하게 되면

$$\text{평균 차이} : \frac{67.45 + 25.76 + 7.39 + 21.97 + 4.71}{5} = 25.46$$

$$\text{차이의 표준편차} : \sqrt{\frac{\sum (x - \mu)^2}{(n-1)}} = 24.94$$

$$\text{표준 오차} : SE = \frac{s}{\sqrt{n}} = \frac{24.94}{\sqrt{5}} = 11.13$$

$$t\text{-통계량} : t = \frac{(\text{평균차이})}{(\text{표준오차})} = \frac{25.46}{11.13} = 2.29$$

$$\text{자유도} : df = n - 1 = 4$$

$$p\text{-value} : p\text{-value} = 0.10 - \frac{(2.29 - 2.132)}{(2.776 - 2.132)} * (0.10 - 0.05) \approx 0.0877$$

와 같이 계산된다. t-통계량의 값은 2.29, p-value의 값은 0.088로 일반적인 유의 수준 0.05에서 통계적으로 유의하지 않다. 이는 AI 기반 사후보전 시스템의 도입이 비가동 시간을 감소시키는 경향이 있지만, 정확한 결론을 내기 위해서는 더 많은 데이터와 더 긴 관찰 기간이 필요하다는 것을 시사한다.

[표 4-2] A사의 고장 유형별 평균 비가동 시간 증감률 (2023년 A사 누적 비가동 시간 상위 5가지 기준), 연구자 재구성(2004)

고장유형	2023년 평균(분)	2024년 3~5월 평균(분)	증감(%)
수정	168.88	101.43	-39.9
교체	70.20	44.44	-36.7
끼임	27.81	20.42	-26.6
작동 불량	58.47	36.50	-37.6
유동	40.28	35.57	-11.7

V. 결 론

5.1 결론 및 시사점

본 연구는 초거대 언어 모델 기반 AI 언어 모델과 RAG 시스템을 활용한 사후보전 방식을 적용하여, 중소 제조업체의 생산성 향상과 비용 절감을 위한 실질적인 해결책을 제시하고자 하였다. 본 연구에 적용한 시스템은 자연어 처리 능력을 갖춘 초거대 언어 모델과 기존 데이터베이스를 효과적으로 활용하는 RAG 시스템을 결합하여 설계되었다. 이 시스템은 설비 고장 시 방대한 데이터를 신속하게 분석하고, 과거 사례와 현재 상황을 종합적으로 고려하여 정확한 원인 분석과 해결 방안을 제시할 수 있다.

실증 연구 대상인 A사의 경우, 이 시스템 도입을 통해 비가동 시간을 약 37.7% 줄이는 효과를 보였다. 이는 설비종합효율(OEE)의 핵심 구성 요소인 시간 가동율의 뚜렷한 개선으로 이어졌다. 시간 가동율 = (부하 시간 - 정지 시간) / 부하 시간 공식에 따라, 정지시간의 큰 폭 감소는 시간 가동율의 상당한 향상을 의미한다. 또한, '작동 불량'과 '교체' 유형의 비가동 시간이 각각 37.6%와 36.7% 감소한 것은 성능 가동율 향상에 직접적인 영향을 미쳤다. 성능 가동율은 속도 가동율(이론 Cycle Time / 실제 Cycle Time)과 정미 가동율(완성량 × 실제 Cycle Time / 가동 시간)의 곱으로 계산되는데, 이러한 개선은 실제 Cycle Time을 단축시켜 성능 가동율을 향상시켰을 것으로 추정된다. 이러한 개선 효과를 종합적으로 고려할 때, AI 기반 사후보전 시스템의 도입은 전체 설비종합효율($OEE = \text{시간가동율} \times \text{성능가동율} \times \text{양품율}$)을 상당히 향상시켰을 것으로 추정된다. 이는 연간 최소 8억 8천만 원에서 최대 13억 2천만 원의 손실액 절감 효과로 이어질 것으로 예측되어, 시스템 도입의 경제적 가치를 입증하고 있다.

또한, 설비 고장 진단의 정확성과 신속성이 전반적으로 향상되었으며, 이는 평균 수리 시간 단축으로 이어졌다. 중소기업의 경영 현실을 고려할 때, 이러한 효과는 기업의 생존과 경쟁력 강화에 직접적인 영향을 미칠 수 있는 수준이다.

최근 미국의 인플레이션 감축법(IRA, U.S. Inflation Reduction Act)이나 유럽의 그린딜 산업계획(Green Deal Industry Plan for the Net-Zero Age) 등으로 인해 현지 생산 강화 추세가 강화되고 있다. 이는 한국 중소기업들에게 새로운 도전이자 기회로 작용하고 있다. 2023년 기준으로 한국 중소기업의 해외 직접 투자는 전년 대비 22.2% 하락하였으나, 글로벌 공급망 재편 전략에 따라 제조업의 해외 투자는 36.5% 증가했으며(한국수출입은행 해외경제연구소, 2023), 현지 인력 관리와 설비 유지보수 문제가 주요 과제로 대두되고 있다. 본 연구에서 적용한 AI 기반 사후보전 시스템은 이러한 문제를 효과적으로 해결할 수 있는 방안을 제시한다. 특히, 언어 장벽이나 문화적 차이로 인한 어려움을 AI 시스템이 일부 해소할 수 있어, 해외 공장 운영의 효율성을 크게 높일 수 있다.

AI 시스템이 모든 현장 인력을 대체할 수 있다고 보기는 어렵다. 여전히 현장에서의 실제 정비와 긴급 상황 대응을 위한 인력은 필요할 것이다. 본 연구에서는 AI와 인간의 장점을 결합한 협업 모델을 지향한다. 예를 들어, AI 시스템이 초기 진단과 해결 방안을 제시하면, 인간 작업자가 이를 바탕으로 최종 판단을 내리고 실제 정비 작업을 수행하는 방식이다. 이를 통해 정확성과 효율성을 높이는 동시에, 인간의 전문성과 경험을 최대한 활용할 수 있을 것이다.

본 연구는 AI 기술을 활용한 제조업 혁신의 가능성을 실증적으로 보여주고 있다. 이는 한국 중소기업들이 글로벌 경쟁력을 강화하고, 해외 진출 시 직면할 수 있는 다양한 문제들을 효과적으로 해결할 수 있는 방향을 제시하고 있다. 향후 이러한 AI 기반 시스템의 도입이 확대된다면, 중소기업의 디지털 전환을 가속화하고, 중소기업의 제한된 자원과 전문 인력 부족 문제를 극복하는 데 큰 도움이 될 수 있다. 그리고 궁극적으로는 한국 제조업 전반의 경쟁력 강화에 기여할 수 있을 것으로 기대된다.

5.2 연구의 한계점 및 향후 연구 방향

본 연구는 A사의 2023년 한 해 동안의 사후보전 데이터를 기반으로 분석을 수행하였다. 이러한 데이터의 한정성으로 인해 연구 결과를 일반화하는 데는 제한이 있다. 또한 초거대 언어 모델과 RAG 시스템의 실효성을 검증하는 과정에서도 실제 현장에서의 파일럿 테스트와 성능 평가가 충분히 이루어지지 않아 이론적인 추정에 그친 부분이 있다.

향후 연구에서는 이러한 한계점을 보완하기 위해 다양한 기업들의 데이터를 포함하여 분석 범위를 확대하는 것이 필요하다. 이를 통해 연구 결과의 신뢰성을 높이고 일반화할 수 있을 것이며, 다양한 산업과 기업의 데이터를 활용함으로써 AI 기반 사후보전 시스템의 적용 가능성과 효용성을 검증할 수 있다. 또한, 장기적인 실증을 통해 추세를 분석한다면 통계적 분석의 신뢰성을 높이고, 시스템 도입의 지속 가능성을 확인할 수 있을 것이다.

‘통계플러스 2024년 봄호’의 관련 보고서에 따르면 인공지능(AI) 등 4차 산업혁명 관련 신기술 도입률은 대기업과 중소기업 간 두 배 이상 차이가 나고 있다(이유지, “중소기업 AI 도입률 대기업 절반... 벌어지는 신기술 격차”, 한국일보, 2024.03.25.). 또한 디지털 마케팅 전 업체인 WSI의 최고 AI 책임자 로버트 미첼은 “핵심 프로세스에 AI를 통합하면 기업은 효율성을 높이고 고객 만족도를 향상시키며 혁신을 이끌어 낼 수 있다.”라고 전했다(이채환, “중소기업의 새로운 기회 ‘AI’”, GTT Korea, 2024.06.08.). 재정적 지원, 기술 컨설팅, 인력 양성 프로그램 등을 포함하여 초기 도입 비용에 대한 부담을 줄이기 위한 세제 혜택이나 보조금 정책에 대한 연구가 요구된다. 연구 결과를 바탕으로 중소기업의 AI 도입을 지원하기 위한 정책적 제언이 필요하다. 이러한 정책적 지원을 통해 중소기업이 AI 기반 사후보전 시스템을 보다 쉽게 도입할 수 있도록 도와야 한다.

이러한 연구 방향을 통해 AI 기반 사후보전 시스템의 실질적 효과를 검증하고, 중소 제조업체의 경쟁력 강화와 지속 가능한 성장에 기여할 수 있을 것이다.

참 고 문 헌

1. 국내문헌

- 권오운 외. (2003). 설비에방보전을 위한 TBM의 최적보전주기 설정모델 연구. 한국경영과학회/대한산업공학회 춘계공동학술대회 논문집, pp106-118
- 권오운 외. (2004) 최적 설비보전정책 수립을 위한 보전방식 선택 모델. 대한설비관리학회, vol9.
- 김봉운. (2008). 보전수준을 고려한 효율적 설비보전시스템 구축. 동아대학교, 박사학위논문.
- 김연수 외. (2004). 물류설비 보전관리를 위한 고장 및 진단정보 지원 시스템. 대한설비관리학회, vol9. no.1
- 김정식 외. (1997). 보전계획을 위한 체계적인 설비선정 및 분석. 대한설비관리학회. Vol.2 no.1
- 나승훈 외. (2012). 설비효율화를 위한 설비관리시스템. 형설출판사.
- 남석주. (2023). 중소기업 스마트팩토리 생산시스템의 효율적 설비관리 방안 에 관한 연구. 경남대학교, 박사학위논문.
- 도성욱. (2020). 카메라 모듈 검사설비에 대한 setup순서와 설비 보전의 효율적 관리 방안에 관한 연구. 금오공과대학교, 석사학위논문.
- 문태수 외. (2003). 자동차 부품산업의 전자조달 EIP 시스템 설계 및 구현. 한국경영과학회 및 대한산업공학회 2023 춘계공동학술대회 논문, pp.28-35.
- 박기철. (2022). TPM 활동요인이 설비종합효율에 미치는 영향에 관한 연구. 전주대학교, 박사학위논문.
- 박선호. (2007). 제조현장의 설비보전 능력향상을 위한 자주학습체계에 관한

- 연구. 대한설비관리학회, vol 12, no.1.
- 신중민 외. (2023). LLM 답변 향상을 위한 검색 기반 생성 기법: GPT3.5, GPT4의 Zero-shot, RAG 비교 연구, 한국정보통신학회 종합학술대회 논문집, 2023.10, pp.350-352
- 안성원 외. (2023). 초거대언어모델의 부상과 주요이슈 - ChatGPT의 기술적 특징과 사회적·산업적 시사점, 소프트웨어정책연구소.
- 오창현. (2021). 예지보전 기술을 활용한 스마트팩토리 효율화에 관한 연구. 한국기술교육대학교, 석사학위논문.
- 이명수. (2017). 다이캐스팅 설비를 위한 예지보전 기반 설비보전관리 시스템. 대한기계학회, KSME17-Th16B008.
- 이영상 외. (2003). 이론과 실무를 접목시킨 하이브리드 TPM, KSA.
- 이영상. (2006). TPM활동을 통해 나타난 추진성과의 핵심 측정지표 개발. 아주대학교, 박사학위논문.
- 이영상 외. (2007). KPI를 활용한 TPM 성과지표에 대한 연구”, 대한설비관리학회지, 제10권, 제3호 pp.101~109.
- 이희양. (2009). 설비종합효율 향상을 위한 TPM 활동 사례연구. 한밭대학교, 석사학위논문.
- 장원태. (2007). 최적보전정책을 위한 경제적 설비구조분석. 대한설비관리학회, vol 12, no.1.
- 정민형. (2011). 사례를 통한 TPM 성과측정 KPI 및 경제적 효과 산출 방법. 한국산학기술학회, 추계 학술발표논문.
- 정천수. (2023). LLM 애플리케이션 아키텍처를 활용한 생성형 AI 서비스 구현: RAG모델과 LangChain 프레임워크 기반.
- 조희중. (2007). TPM 추진 및 성과측정 방법론에 관한 연구. 고려대학교, 석사학위논문.
- 최무호 외. (2010). 중소기업 TPM 성과에 영향을 주는 요인에 대한 실증적

- 연구. 품질경영학회지, 제38권 제3호.
- 최병관. (2006). 활동기준원가를 이용한 설비보전비 계산 시스템에 관한 연구. 아주대학교, 박사학위논문.
- 최영구. (2018). 데이터베이스 기반 설비보전을 위한 개방형 PLC 구현과 적용 연구. 창원대학교, 석사학위논문.
- 함효준. (1997). 한국제조업체의 설비관리 조직 변화와 TPM추진 현황분석에 관한 연구. 대한설비관리학회지, 제2권(2호).
- 함효준 외. (2001). 보전정책 수립을 위한 설비구조 분석. 대한설비관리학회 제6권 4호, pp.36~42.
- 함효준. (2005). 수익성 중심의 설비관리. 동현출판사, pp.49~79
- 황호연. (2010). 발전설비 신뢰도 향상을 위한 보전관리시스템 구축 방안 연구. 한양대학교, 석사학위논문.

2. 국외문헌

- Alan Turing. (1950). "Computing Machinery and Intelligence", Mind 49, pp.433~460.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. (). "Language Models are Unsupervised Multitask Learners".
- Alex Graves, Abdel-rahman Mohamed, Geoffrey Hinton. (2013). "Speech Recognition with Deep Recurrent Neural Networks." IEEE International Conference on Acoustics, Speech, and Signal Processing.
- Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton. (2012). "ImageNet Classification with Deep Convolutional Neural Networks", Advances in Neural Information Processing Systems 25 (NIPS 2012).
- Annie Francie Kouedeu, Jean-Pierre Kenné, Victor Songmene. (2011). "Production, Preventive and corrective maintenance planning in manufacturing systems under imperfect repairs".
- Ashish Vaswani, NoamShazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. (2017). "Attention Is All You Need".
- Bing Liu. (2012). "Sentiment Analysis and Opinion Mining".
- Christer Idhammer. (1997). "In search of conservation without waste", Chemical Engineering.
- Claude E. Shannon, (1948). "A Mathematical Theory of Communication", The Bell System Technical Journal, Vol. 27, pp. 379–423, 623–

656.

Claude E. Shannon, (1951). "THE REDUNDANCY OF ENGLISH".

David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams. (1986).
"Learning representation by back-propagating errors", Nature Vol.
323, pp.533~536.

Frank Rosenblatt. (1958). "The Perceptron: A Probabilistic Model for
Information Storage and Organization in the Brain", Psychological
Review Vol. 65, No. 6, 1958.

Geoffrey E. Hinton, Simon Osindero, Yee-Whye Teh. (2006). "A Fast
Learning Algorithm for Deep Belief Nets", Volume 18, Issue 7,
Neural Computation, pp.1527~1554.

Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam
Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G,
Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V.
Vasilakos, Thippa Reddy Gadekallu. (2023). "GPT (Generative
Pre-trained Transformer) – A Comprehensive Review on
Enabling Technologies, Potential Applications, Emerging
Challenges, and Future Directions".

Goled, Shraddha (2021). "Self-Supervised Learning Vs Semi-Supervised
Learning: How They Differ". Analytics India Magazine.

Humza Naveed, Asad Ullah Khana, Shi Qiub, Muhammad Saqib, Saeed
Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnesh, Ajmal
Miani. (2024). A Comprehensive Overview of Large Language
Models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. (2018).
BERT: Pre-training of Deep Bidirectional Transformers for

Language Understanding.

John Hutchins. (2004). "The first public demonstration of machine translation: the Georgetown-IBM system".

Justin T. Chiu, Alexander M. Rush. (2020). "Scaling hidden Markov language models".

Kathleen E. Mckone, (2001). "The impact of total productive maintenance practices on manufacturing performance", Journal of operations management 19.

Lawrence R. Rabiner. (1989). "A tutorial on hidden Markov models and selected applications in speech recognition", Proceedings of the IEEE Volume: 77, Issue: 2, pp.257~286.

Leonard E. Baum. (1972). "An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes", Volumn 3, Inequalities, pp.1~8

McCorduck Pamela. (2004). "Machines Who Think" 2nd Edition, pp.111~136.

Oriol Vinyals, Quoc V. Le. (2015). "A Neural Conversational Model".

Peter Poór, David Ženíšek, Josef Basl. (2019). "Historical Overview of Maintenance Management Strategies: Development from Breakdown Maintenance to Predictive Maintenance in Accordance with Four Industrial Revolutions"

Qiao Jin, Yifan Yang, Qingyu Chen, Zhiyong Lu. (2023). "GeneGPT: Augmenting Large Language Models with Domain Tools for Improved Access to Biomedical Information".

Samuel Sam Santhosé, Lawrence Noble Lawrence (2023). "Understanding the implementations and limitations in knowledge management

- and knowledge sharing using a systematic literature review”
Volume 42, pp. 32427~32442, Current Psychology.
- Seichi Nakajima. (1998). Introduction to TPM, Productivity Press,
Cambridge,MA.
- Sepp Hochreiter, Jürgen Schmidhuber. (1997). “Long Short-Term
Memory”.
- Shaoxuan Zhai, Zhenpeng Liu. (2023). “Artificial intelligence technology
innovation and firm productivity– Evidence from China”.
- Talal Almutiri, Farrukh Nadeem. (2022). “Markov Models Applications in
Natural Language Processing: A Survey”, Information Technology
and Computer Science, 2022, 2, 1–16.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared
Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam ,
Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel
Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child,
Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter,
Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott
Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam
McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. (2020).
“Language Models are Few-Shot Learners”.
- Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel
Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller,
Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles
Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan
Kumaran, Daan Wierstra, Shane Legg & Demis Hassabis. (2015).
“Human-level control through deep reinforcement learning”,

Nature volume 518, pp.529–533.

Warren S. McCulloch, Walter Pitts. (1943). “A logical calculus of the ideas immanent in nervous activity”, Volume 5, pp.115~133, The bulletin of mathematical biophysics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean. (2016). “Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation”.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin. (2003), “A Neural Probabilistic Language Model”.

Yunfan Gaoa, Yun Xiongb, Xinyu Gaob, Kangxiang Jiab, Jinliu Panb, Yuxi Bic, Yi Daia, Jiawei Suna, Meng Wangc, and Haofen Wang a. (2024). “Retrieval-Augmented Generation for Large Language Models: A Survey”.

3. 연구보고서 및 단행본

- 기술표준원. (2004). 설비보전 관리 관련 용어 표준화 및 용어집 발간을 위한 조사 연구. pp.164~165.
- 능률협회컨설팅(KMAC). (1991). TPM(설비보전).
- 산업연구원. (2023). 2024년 경제산업전망.
- 실무노동용어사전. (2014).
- 중소기업진흥공단. (2009). TPM(설비효율화) 종합.
- 한국생산성본부. (1981). 생산성 저해 요인에 관한 조사연구.
- 한국생산성본부. (1983). 생산성대전.
- 한국표준협회. (2000). 생산혁신을 위한 신 TPM 전개프로그램(장치공업편)
- Liliane Pintelon, Alejandro Parodi-Herz, Khairy A. H. Kobbacy, D. N. Prabhakar Murthy. (2008). “Maintenance: An Evolutionary Perspective”, Complex System Maintenance Handbook

4. 정부기관 및 연구기관

- 사단법인 일본 플랜트 메인テナンス 협회. (2023). 2022년도 유지보수 실태조사
보고서 개요
- 통계청. (2019). 2019년 고용데이터
- 통계청. (2020). 2020년 고용데이터
- 통계청. (2021). 2021년 고용데이터
- 통계청. (2022). 2022년 고용데이터
- 통계청. (2023). 2023년 고용데이터
- 한국수출입은행 해외경제연구소. (2024). 2023년 4분기 분기 보고서 해외직
접투자 동향분석

ABSTRACT

Empirical Study on How to Improve Production Machine Management System Using Large AI Language Model

– Focusing on automotive parts manufacturing –

Kim, Hyun-Jin

Major in Smart Convergence

Technology Consulting

Dept. of Smart Convergence Consulting

Graduate School of Knowledge Service
Consulting

Hansung University

In modern society, the efficiency of preservation, operation, and maintenance of production facilities to improve productivity of manufacturing companies is becoming increasingly important. Since production facilities are always in operation 24 hours a day, abnormal symptoms such as wear, damage, and malfunctions due to aging of major parts occur regularly. When these abnormalities occur, production facilities are stopped and productivity is reduced, causing serious problems for SME management. To solve these problems, various improvement measures are being promoted, such as the introduction of smart factories

and the application of predictive maintenance. However, the high costs associated with these systems pose a major challenge to their practical application.

As such, there has not been much research on preventive measures or improvement of facility management performance indicators, which are studies to solve the problem of production stoppage due to production equipment failure. The reason for this is not only the failure of production facilities, but also the difficulty of research due to unexpected special situations such as switching to or restoring alternative facilities due to facilities with specifications that do not meet the performance level. These include complex issues such as determining the cause of the failure, procuring the necessary parts, deploying specialized engineers, and managing the impact on connected facilities or equipment during the downtime. In practice, proactive maintenance activities are not systematically managed and rely heavily on the individual skills of maintenance personnel. While preventive maintenance is emphasized when evaluating overall equipment effectiveness (OEE), in reality, the majority of time is spent on reactive activities. As a result, many manufacturers have developed sophisticated maintenance manuals to increase the survivability of their production facilities. This complexity makes it difficult to conduct research to improve productivity or survivability based on preventive maintenance practices.

In this study, we propose a method to improve the efficiency of quadratic warfare during equipment maintenance activities in SMEs by utilizing AI technology. This study utilizes a large language model (LLM) to solve the problem of knowledge sharing in SMEs based on historical

data of equipment maintenance including failure causes, failure diagnosis, environmental factors, and response time. This study aims to improve the overall efficiency of facilities and provide effective solutions to the practical problems faced by SMEs to increase the productivity and competitiveness of the manufacturing industry.

The research method was to collect preliminary data from the production equipment of the subject company, and to conduct frequency analysis and standard statistical evaluation. In the preliminary experiment, the feasibility of the system based on the super-large language model was verified by querying specific contents of the smartphone manual and receiving relevant responses. In this research experiment, the equipment failure symptoms, failure types, and downtime for each failure component were comprehensively investigated to analyze the impact on the overall equipment efficiency.

This study aims to contribute to the improvement of productivity and competitiveness of manufacturing companies by presenting an AI-based model for improving the efficiency of quadratic warfare through empirical research results. This will play an important role in solving the practical challenges faced by SMEs and promoting sustainable industrial growth.

【Key words】 equipment maintenance, quadratic warfare, emergency maintenance, LLM, AI, RAG