

# 공작기계 메뉴얼 특화 생성형 언어 모델 개발

이예원<sup>1</sup> · 박주영<sup>1</sup> · 김시연<sup>1</sup> · 조성우<sup>2</sup> · 엄주명<sup>1,2\*</sup>

<sup>1</sup>경희대학교 산업경영공학과 / <sup>2</sup>경희대학교 인공지능대학원

## Development of a Generative Language Model Specialized in CNC Machine Tool Manuals

Yewon Lee<sup>1</sup> · Ju-yeong Park<sup>1</sup> · Siyeon Kim<sup>1</sup> · Seongwoo Cho<sup>2</sup> · Jumyung Um<sup>1,2</sup>

<sup>1</sup>Department of Industrial & Management System Engineering, Kyung Hee University

<sup>2</sup>Department of Artificial Intelligence, Kyung Hee University

Due to the advance of machine tool technology, operation manuals have been becoming more complex and diverse than before. Accordingly, if machine technicians solve individual troubles happened in own machines, the losses such as human error, increased idle time, and delays still occur because of fast retirements and the lack of expertise in manufacturing companies. In this paper, we propose localized Manual LLM, an AI-based language model that enables real-time problem solving by answering operators' questions on-site. The training data set was developed through rapid engineering using GPT-3.5 Turbo and is divided into two main categories: a question collection manual for general manual topics and an alarm manual for CNC warnings. Llama2 was selected as the pre-trained model and hyperparameters were optimized through fine-tuning through multiple tests, and the final evaluation was based on the training loss and response accuracy. The model can be accessed through a mobile app, allowing operators to receive immediate support.

**Keywords:** Manual LLM, Generative Language Model, Prompt Engineering, Natural Language Processing, Machine Tool, CNC Controller

### 1. 서론

#### 1.1 연구 필요성 및 목적

공작기계 산업의 발전에 따라, 공작기계와 이를 제어하는 CNC (Computer Numerical Control) 컨트롤러의 종류도 다양해지고 있다. 이에 작업자들은 더 많은 공작기계와 CNC의 메뉴얼 가이드북을 숙지해야 한다는 부담을 가지게 된다. 작업자들은 수천 쪽씩 되는 메뉴얼 가이드북을 암기할 수 없으므로 산업 현장에서 문제가 발생할 때마다 두꺼운 메뉴얼 가이드북을 찾아보거나 그동안 쌓은 노하우를 기반으로 이를 해결하곤 한다. 이런 방식의 문제 해결은 또 다른 문제점을 낳게 된다. 책자를 찾으며 문제를 해결하면 그 양이 방대하므로 문제를

해결하는 데 소요되는 시간이 늘어나게 된다. 노하우를 기반으로 문제를 해결할 때는 휴먼 에러가 발생할 수 있고 이에 따른 시간 손실이 발생한다. 문제 해결을 위해 전문 엔지니어를 부르더라도 문제는 곧바로 해결되지 않는다. 관련 논문에 따르면 CNC 기계에 결함이 발생할 때 초급 유지보수 기술자들의 경험 부족과 결함 원인을 신속하게 찾을 수 있는 경험이 풍부한 엔지니어 수의 부족으로 결함 진단에 문제를 겪고 있음을 알 수 있었다(Anzari *et al.*, 2023; Michael *et al.*, 2021).

이러한 문제점을 해결하기 위해 작업자가 공작기계와 CNC 컨트롤러를 능숙하게 사용할 수 있도록 도움을 주는 지원 프로그램 및 시스템이 필요할 것이다. 이에 본 연구에서는 공작기계 산업 현장에서 작업자의 요청에 따라 문제 사항을 파악하고 해결 방법의 즉각적인 제시를 통해 작업의 효율을 극대화하는

본 연구는 과학기술정보통신부 정보통신기술기획평가원의 지원과(No. RS-2022 -00155911, 인공지능융합혁신인력양성(경희대)), 산업통상자원부 한국산업기술평가원의 지원을 받아 수행되었습니다(No.20026431).

\* 연락처 : 엄주명 교수, 경기도 용인시 덕영대로 1732 경희대학교 국제캠퍼스, Tel : 031-201-5443, Fax : 031-202-8106, E-mail : jayum@khu.ac.kr  
2024년 11월 14일 접수; 2024년 11월 24일; 2024년 11월 29일 수정본 접수; 2024년 12월 2일 게재 확정.

AI 기반 생성형 언어 모델, Manual LLM을 개발하고자 한다.

## 1.2 연구 내용 및 방법

- 1) 메뉴얼 텍스트 전처리: 우선, CNC 공작기계의 메뉴얼 가이드북 pdf의 텍스트를 추출하고 전처리한다. 이 단계에서는 텍스트 정제, 토큰화, 및 문장 구조 유지를 위한 작업이 수행된다.
- 2) QA 데이터셋 생성: 전처리한 텍스트를 기반으로 QA(질문-답변) 데이터셋을 생성하기 위해 트랜스포머 모델을 활용한다. 이 데이터셋은 메뉴얼 가이드북의 전반적인 내용과 작업 현장에서 발생할 수 있는 정보를 중심으로 한 QA 데이터로 향후 모델 학습 및 평가에 사용된다.
- 3) 모델 학습 및 평가: 후보 모델(GPT-2, Llama2)은 생성된 QA 데이터셋을 기반으로 학습된다. 학습된 모델은 정성적 및 정량적 평가를 거쳐 성능을 평가하고 최종 모델을 선정한다.
- 4) 프로토타입: 사용자 친화적인 UI를 갖춘 모바일 애플리케이션을 개발하여 최종 모델의 사용성을 향상한다.

## 2. 관련 연구

### 2.1 생성형 언어 모델

관련 논문을 통해 우리는 Long Short-Term Memory 이나 Recurrent Neural Network 보다 더 나은 퍼포먼스를 보여주는 트랜스포머 모델에 관한 내용과 이 트랜스포머가 BERT, GPT 등 자연어 처리 모델에 어떻게 적용되고 활용되는지에 대한 내용을 알 수 있다(Vaswani *et al.*, 2017; Wolf *et al.*, 2020). 또한 현시점 Natural Language Processing의 강자인 OpenAI에 Microsoft의 지속적인 투자가 이루어지고 있으며, 두 회사가 파트너십을 강화했다. 이와 함께, NC소프트의 멀티모달 생성 AI, 네이버의 HyperCLOVA, 카카오브레인의 KoGPT 등 다양한 언어 모델들이 Manual LLM 개발의 필요성을 확보할 수 있었다(Son *et al.*, 2024; Yoo *et al.*, 2024; Weitzmann, 2021).

#### (1) GPT-2

본 연구에서 다루게 될 언어 모델 중 하나인 GPT-2 모델의 구조와 작동 원리를 분석하였다. 공작기계 작업자를 위한 챗봇 개발에는 사내 보안을 위해 공장 내 서버에서 운영 가능한 경량 모델이 필요하여 GPT-2를 선택하였고, 최신 GPT-3.5는 성능은 뛰어나지만 운용 보안 문제로 배제하였다. GPT-2는 OpenAI에서 개발한 자연어 처리를 위한 딥러닝 언어 모델 중 하나로, 2019년에 공개되었으며 자연어 처리 분야에서 혁신적인 성과를 내는 인공 지능 언어 모델이다(Radford *et al.*, 2019). 이 모델의 구조를 이해함으로써 우리는 자연어 생성 및 이해 과제를 보다 효과적으로 해결할 수 있을 것으로 기대한다.

GPT-2는 자연어 처리 작업에서 뛰어난 성능을 보이며, 어텐션 메커니즘을 활용해 문장 내의 단어 간 상호작용을 모델링하는 transformer 아키텍처를 기반으로 한다. GPT-2는 WebText라 불리는 40GB 크기의 거대한 코퍼스와 인터넷에서 수집한 데이터를 결합하여 훈련하였다. 약 117M 개의 가장 작은 GPT-2 모델 파라미터들을 저장하는 데 필요한 저장 공간만 약 500MB이고, 가장 큰 GPT-2 모델 파라미터들을 저장하는 데에는 약 6.5GB의 저장 공간이 필요하다. 다양한 크기와 버전으로 제공되며 작은 모델부터 큰 모델까지 다양한 용도와 요구 사항에 맞게 선택할 수 있다. GPT-2와 같은 pre-trained 모델은 특정 작업에 맞게 추가적인 학습을 통해 성능을 향상할 수 있으며, 이로써 새로운 작업에 많은 양의 데이터가 필요하지 않으면서도 효과적인 모델을 생성할 수 있다.

#### (2) Llama2

본 연구에서 다루게 될 언어 모델 중 다른 하나인 Llama2 모델의 구조와 작동 원리 또한 분석하였다. Llama2는 Meta에서 2023년 출시한 오픈소스 언어 모델이다 (Touvron *et al.*, 2023). 공개된 버전은 7B, 13B, 34B, 70B의 4가지로, GPT-2 모델과 비교하여 훨씬 많은 크기의 파라미터를 가지고 있다. Context의 길이는 Llama1 모델보다 2배 정도 늘어난 약 4K 개, 학습된 토큰의 수는 Llama1 모델보다 약 1.5~2배 정도 늘어난 2T 개이다. Llama2 모델은 파라미터의 크기가 큰 모델일수록 train loss가 토큰 처리에 따라 더 급격히 줄어든다. Llama2 모델 또한 GPT-2와 같은 pre-trained 모델로, 특정 작업에 맞게 추가적인 학습을 통해 성능을 향상할 수 있으며, 이로써 새로운 작업에 많은 양의 데이터가 필요하지 않으면서도 효과적인 모델을 생성할 수 있다.

### 2.2 LLM 도메인 커스터마이징

본 연구는 GPT-2와 Llama2를 활용하여 공작기계 작업자 지원을 위한 챗봇을 개발하는 데 있어, 도메인 특화 학습을 실용적으로 적용한 예시를 제공한다.

금융 문서를 활용한 QA 데이터 증강을 통해 투자 분석에 특화된 LLM을 개발하였다(Li *et al.*, 2023). 제조업 공정 데이터를 활용하여 설비 유지보수를 위한 챗봇을 구축하였다(Wang *et al.*, 2023). 이러한 기존 연구들과 비교했을 때, 본 연구 역시 특정 도메인 데이터를 기반으로 LLM을 커스터마이징하여 실무 활용성을 검증했다는 공통점을 지닌다. 그러나 차별점으로는 기존 연구들이 주로 의료, 금융, 제조 등 비교적 잘 연구된 도메인에 집중된 것과 달리, 본 연구는 공작기계 작업자 지원이라는 상대적으로 미개척된 분야에서 LLM 커스터마이징의 새로운 가능성을 열었다는 점을 들 수 있다. 기존 연구들과의 주요 차별점은 다음과 같다.

- 1) 경량화 모델 적용: Llama2를 선택하여 개인 서버에서 운영할 수 있도록 최적화, 운영 효율성을 중시한 설계한다.

- 2) 자동화된 데이터 생성: 메뉴얼 북을 기반으로 QA 데이터를 자동 생성하여 학습 비용을 대폭 절감한다.
- 3) 도메인 맞춤형 학습: 공작기계 작업자에 특화된 데이터를 기반으로 성능을 향상시켜 실무 적용 가능성을 검증한다. 이로써 본 연구는 기존의 LLM을 특정 도메인에 최적화하는 기존 흐름을 기반으로, 실제 공작기계 산업 환경에 적합한 모델을 구축하는 데 중요한 기여를 한다.

### 3. 본론

제안한 시스템은 3가지 단계에 의해서 구축이 된다(<Figure 1>). 먼저 메뉴얼 북에서 QA 데이터 셋을 생성하는 단계, Llama모델을 학습시키는 단계 그리고 3) 모바일 사용자와 통신하는 단계로 나눌 수 있다. 해당 단계의 기능에 대해서 설명하도록 한다.

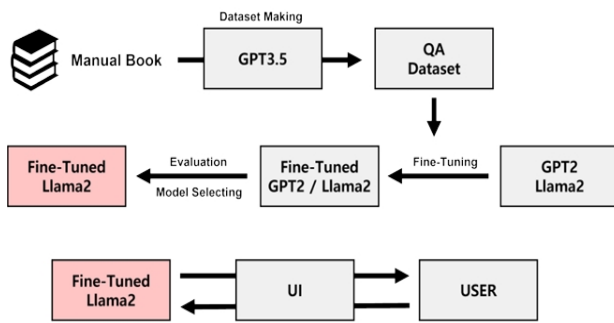


Figure 1. Procedures of Fine-tuning and Operations of the Proposed Manual LLM

#### 3.1 Dataset

##### (1) 데이터 수집과 전처리

본 연구에서는 공작기계 사용자를 지원하기 위해 pdf 형식의 메뉴얼 가이드북으로부터 텍스트를 추출하여 사용하였다. 연구에 사용된 메뉴얼 가이드북은 'FANUC MANUAL GUIDE i OPERATOR'S MANUAL'이다. FANUC MANUAL GUIDE i는 CNC 기계 도구를 조작하는 데 사용되는 소프트웨어이고, operator's manual은 이 소프트웨어의 사용자 메뉴얼로 CNC 작업자나 프로그래머가 FANUC MANUAL GUIDE i를 사용하여 CNC 기계 작업을 수행할 때 필요한 지침과 도움말을 제공한다. 이 pdf에서 텍스트를 추출하는 것은 자동화된 방식으로 이루어졌으며, 추출된 텍스트 데이터를 이후 활용하기 위해 전처리를 진행하였다.

추출된 텍스트 데이터로부터 패턴을 발견하고, 이를 기반으로 문서의 구조를 이해하였다. 텍스트 데이터는 title, heading, content로 나누어 저장하였다. 이때 content는 이후 GPT-3.5 Turbo 모델에 넣을 것이므로 content의 토큰 수는 모델이 처리

가능한 최대 토큰 수인 약 4,000개에 맞게 정제하였다.

##### (2) 메뉴얼 가이드북 QA 데이터 생성

전처리가 끝난 데이터의 title, heading, content를 합쳐 context를 생성하였다. 이후 OpenAI API를 호출하여 GPT-3.5 Turbo 모델을 사용하여 질문을 생성하였다. 질문 생성에 사용한 프롬프트는 <Figure 2>와 같다.

질문 데이터를 생성한 뒤 같은 모델에 또 다른 프롬프트를 사용하여 답변을 생성하였다. 답변 생성에 사용한 프롬프트는 <Figure 3>과 같다. QA 데이터 생성 과정에서는 GPT-3.5 Turbo 모델에 사용한 프롬프트에 FANUC MANUAL GUIDE i를 언급함으로써 현장의 작업자에게 유용한 질문과 답변을 자동으로 생성하였다.

```

You are going to make some questions that
may arise using a machine equipped with the
FANUC MANUAL GUIDE i. A part of the
manual book will be provided. Please write
questions based on the context below.

Context:{context}

Questions:
1.
    
```

Figure 2. Prompt(1)

```

Here are some questions that may arise
using a machine equipped with FANUC
MANUAL GUIDE i. A part of the manual book
will be provided. Please answer to the
questions based on the context below.

Context:{context}

Questions:
{questions}

Answers:
1.
    
```

Figure 3. Prompt(2)

##### (3) 알람 QA 데이터 생성

알람에 대한 부분은 작업자가 가장 많은 오류를 만날 수 있는 곳 중 하나이므로, 더 많은 데이터를 생성하였다. 메뉴얼 가이드북의 'Appendix'의 'Alarms' 섹션에서 알람 번호, 발생원인, 해결 방법을 추출하여 데이터프레임으로 만들었다. 이를 기반으로 두 가지 유형의 QA를 생성하였다.

첫 번째는 알람 발생원인이 주어졌을 때 해결 방법을 묻는 질문에 대해 해결 방법을 답변으로 제공한다.

두 번째는 알람 번호가 주어졌을 때 해결 방법을 묻는 질문에 대해 발생원인과 해결 방법을 답변으로 제공한다. 이때는 첫 번째 경우보다 다양한 질문을 생성하고자 GPT-3.5 Turbo 모델에 prompt one-shot 방식을 사용하였다. 사용한 프롬프트는 <Figure 4>와 같다.

You will be provided with an example that contains a context, a question, and an answer. Please make some questions and answers based on this example.

The number of questions and answers cannot be less than the number of warning numbers. You will first be provided with an example. Consider this example:

Context:  
Appendix  
Alarms

B. ALARMS APPENDIX Alarm 16i 30i Description Cause The first feed override is invalid. 3062 3562 Action A value not specifiable as turning or other first feed overrides is entered. Modify the machining program to specify an appropriate value.

Questions:

1.What should I do when I receive warning number 3062?  
2.What should I do when I receive warning number 3562?

Answers:

1. Cause: The first feed override is invalid. Action: A value not specifiable as turning or other first feed overrides is entered. Modify the machining program to specify an appropriate value.  
2. Cause: The first feed override is invalid. Action: A value not specifiable as turning or other first feed overrides is entered. Modify the machining program to specify an appropriate value.

Now, please write questions and answers based on the context below.

Context:{context}

Questions:  
1.

Figure 4. Prompt(3)

#### (4) 최종 데이터셋

2.1.(2)와 2.1.(3)에서 만든 QA 데이터를 통합하여 최종 학습 데이터와 검증 데이터를 만들었다. 이때 검증 데이터는 기존 QA 데이터를 특정 비율로 단순히 분리할 경우, 메뉴얼 가이드북의 챗터별 데이터 수나 알람 번호별 데이터 수가 균등하지 않아질 수 있으므로 이 점에 유의하여 학습 데이터와 검증 데이터를 랜덤 샘플링 방식으로 나누었으며, 비율은 6:4로 설정하였다. 학습 데이터는 메뉴얼 가이드북 QA 데이터 약 10,000개와 알람 QA 데이터 약 10,000개로 이루어졌다. 검증 데이터는 메뉴얼 가이드북 QA 데이터 약 4,500개와 알람 QA 데이터 3,000개로 구성되었다. 데이터셋의 검증에 대해서는 문법적 검증 절차를 수행하였으며, 전문가에 의한 컨텍스트 검증 절차를 추후에 수행할 예정이다.

### 3.2 모델 학습

본 연구의 목적은 공작기계 작업자를 위한 메뉴얼에 특화된 챗봇을 만드는 것이기 때문에 특정 도메인에서 최적의 성능을 제공하는 fine-tuning 방법을 선택했다. fine-tuning은 모델 파라미터를 조정하여 특정 도메인에 최적화된 성능을 제공할 수 있는 접근 방식이다. 반면 Retrieval-Based Augmentation은 모델을 업데이트하지 않고 외부 데이터베이스를 활용하여 확장성을 제공하나, 본 연구는 도메인 특화 성능 극대화가 중요하므로 fine-tuning이 적합하다고 판단하였다. 우리는 GPT-2와 Llama2의 2가지 모델을 fine-tuning한다. 해당 2가지 모델은 서로 다른 구조와 학습 방식을 가지고 있으므로 기존에 데이터를 전처리한 것에 추가적인 전처리 과정이 필요하다. 이에 따라 각 모델의 특성과 요구사항에 맞게 데이터를 조정하고 가공하였다. 일반적으로 전처리한 데이터는 question과 answer 열로 구성되어 있다.

#### (1) GPT-2 Fine-Tuning

초기 단계에서는 원본 데이터를 텍스트 파일 형태로 전환해야 한다. 또한 학습 데이터의 구조나 형식을 조정해야 한다. 이를 위해 우리는 질문 데이터와 답변 데이터에 특별한 토큰을 추가했다. 'question' 열에 있는 모든 질문 앞에는 '[Q]'라는 표시를 붙여주었다. 이 표시는 해당 텍스트가 질문임을 모델에 알려주는 역할을 한다. 마찬가지로 'answer' 열에 있는 답변의 시작 부분에는 '[A]'라는 표시를 붙여주었다. 이 표시는 해당 텍스트가 답변임을 모델에 알려주는 역할을 한다. 이러한 방식으로 모델이 질문과 답변을 명확히 구분하도록 하였다.

허깅페이스의 transformers 라이브러리를 사용하여 GPT-2 모델을 불러온 후 학습을 진행했다. 모델 학습은 구글 코랩 환경에서 진행했다. GPU는 google colab의 T4를 사용했다. GPT-2 모델 학습에 사용된 주요 파라미터로는 epoch와 배치 사이즈가 있다. 모델을 학습시킬 때, 여러 번의 epoch를 거치며 학습을 반복하게 되면, 모델의 성능은 일반적으로 점차 개선된다. 하지만 epoch 수를 지나치게 늘리면, 과적합(overfitting)이라는 문제에 직면할 가능성이 있다. 이는 모델이 학습 데이터에 너무 특화되어 실제 환경에서의 성능이 떨어지는 현상을 말한다. GPT-2 모델 fine-tuning 과정에서는 epoch를 4로 설정하였다. 그 이유는 검증 데이터셋을 통해 학습 과정을 모니터링 해보았을 때, 5번째 epoch부터는 과적합 현상이 나타났기 때문이다. 그리고 배치 사이즈는 한 번의 파라미터 업데이트를 할 때 사용하는 샘플의 수를 말한다. 배치 사이즈는 8로 설정하였다.

#### (2) Llama2 Fine-Tuning

우선, 질문 텍스트를 처리할 때, 질문의 시작 부분에는 [INST]라는 태그를 붙였고, 끝부분에는 [/INST] 태그를 붙였다. 이렇게 해서 질문의 시작과 끝을 명확하게 구분할 수 있게 했다. 그 후에, 해당 질문에 대한 답변 텍스트를 바로 연결해 하나의 통합된 열을 생성했다. 이렇게 만들어진 통합 열의 시작 부분에는 <s> 태그를, 끝부분에는 </s> 태그를 추가했다. 여기서 <s> 태그는 문장이 시작된다는 것을 의미하며, [INST]는 사용자의 입력 부분을 알려주는 태그다. 예를 들어, <s>와 </s> 태그 사이에 전체 질문과 답변 내용이 포함되어 있고, [INST]와 [/INST] 사이에는 질문 부분만 들어가 있다. 마지막으로, 이렇게 전처리된 데이터는 원래 csv 파일 형식으로 저장되어 있었지만, 뉴럴네트워크 라이브러리와 호환되도록 하기 위해서는 이를 전용 datasets 형태로 변환해야 한다. 요약하면, Llama2 모델 학습을 위한 데이터 전처리는 기존 데이터에 특정 태그를 추가하여 문장의 시작, 끝 및 사용자 입력 부분을 구분하고, 이를 뉴럴네트워크 라이브러리와 호환되는 형태로 변환하는 작업을 포함하고 있다.

먼저, 'datasets'라는 라이브러리는 허깅페이스에서 제공하는 것으로, 주로 학습 데이터를 처리하고 변환하는 데 사용된다. 학습 데이터를 효과적으로 다루기 위한 다양한 도구와 기능들이 포함되어 있어, 데이터의 형식 변환이나 적재 과정에서 유용

하다. 다음으로, ‘peft’는 모델 학습에 있어서 효율적인 방법을 제공하는 라이브러리다. 전체 모델의 모든 매개변수를 미세 조정하는 대신, 일부 매개변수만을 조절하게 되면서 계산량이나 저장 공간 등의 자원을 크게 절약할 수 있게 된다. 이로 인해, fine-tuning 과정이 더 빠르고 경제적으로 진행될 수 있다. ‘Bitsandbytes’는 모델의 크기 최적화를 위한 도구로써, 4bit 양자화 기술을 활용해 모델의 크기를 줄이는데 효과적이다. 이렇게 함으로써 메모리 제약이 있는 환경에서 매우 유용하다. 마지막으로, ‘trl’ 라이브러리는 특별한 학습 방법론을 제시한다. ‘Transform’, ‘retrieve’, ‘label’의 세 가지 과정을 거쳐 학습을 진행하는데, 이는 원본 텍스트 데이터를 변환하고, 그 중에서도 가장 중요하고 관련 높은 정보를 추출하여, 그 정보에 적절한 라벨을 부여하는 과정을 통해 진행된다. epoch는 5로 설정하고 배치 사이즈는 4로 설정했다. 학습률(learning\_rate)은 2e-4로, 가중치 감소율(weight decay)은 0.001로 설정하였다. 이 두 파라미터는 모델 학습의 방향성과 속도를 조절하는데 영향을 미친다. 최적화 알고리즘으로는 AdamW를 사용하였다. AdamW는 Adam 알고리즘의 변형으로, 가중치 감소가 적용된 알고리즘이다. 이는 모델의 일반화를 도와 과적합을 방지한다. 학습률의 변화를 조절하는 방식으로는 코사인 기반의 스케줄링을 선택하였다. 코사인 기반 스케줄링은 학습률이 훈련 과정에서 코사인 함수를 따라 감소하게 하여, 초기에는 전역 최적점을 찾기 위해 빠르게 탐색하고, 나중에는 지역 최적점에 더 세밀하게 접근하는 방식이다. 이 방법은 모델이 좋은 성능에 도달하면 학습률을 줄여, 성능의 향상을 안정화시킬 수 있다. Llama2는 GPT-2와 마찬가지로 구글 코랩 환경에서 진행하였고 GPU는 T4를 사용했다.

## 4. 프로토타입 개발 (UI)

### 4.1 모바일 애플리케이션 제작

사용자는 모바일 애플리케이션(application, 이하 앱)을 통해 공작기계에 관한 질문을 하고, 그에 대한 답변을 받을 수 있다. 앱 디자인은 안드로이드 앱 프로토타이핑 툴을 이용해서 진행했다. 앱에 접속하면, 공작기계 3사의 로고가 표시된다. 사용자가 원하는 공작기계에 해당하는 회사의 로고를 클릭하면, 안내 페이지로 넘어간다. 안내 페이지에서는 인사말과 Manual LLM가 어떤 앱인지에 대한 간단한 설명, example 부분의 예시를 참고해서 Manual LLM에 질문을 넣으면 된다는 설명이 제공된다. 이후 ‘확인’ 버튼을 누르면 채팅창이 뜨고, 모델에 질문을 넣었을 때 답변을 받는 채팅이 진행된다. Fig. 6은 Manual LLM 프로토타입 이미지이다.

### 4.2 서버 컴퓨터 연동

사용자가 앱 입력창에 질문을 입력하면 해당 값을 서버 컴퓨터로 전송해서 fine-tuning한 Llama2 모델을 통해 답변을 생

성한다. Llama2 모델이 답변을 생성하면 앱에 전송해 사용자의 화면에 보여준다. 이를 진행하기 위해 Fast API를 사용하여 서버 컴퓨터에 웹 애플리케이션을 만들었고, 안드로이드 앱 프로토타이핑 툴에서 구현하였다.

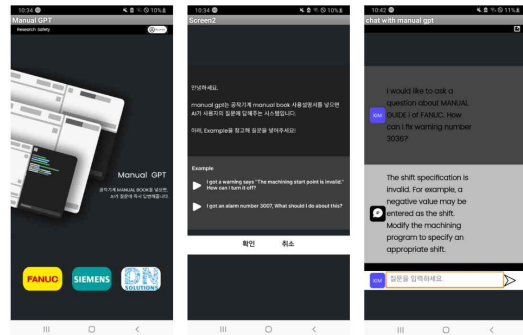


Figure 5. Manual LLM Prototype

Fast API는 파이썬 기반의 웹 프레임워크로, Fast API를 사용한 서버 컴퓨터와 앱 간의 연동은 실제 산업에서 다수의 이점을 제공한다. 먼저, 앱 내에서 모델을 탑재하는 경우 데이터가 변경될 때마다 파인튜닝을 해야 하는 데 시간이 많이 소모된다는 한계점을 극복할 수 있다. 서버 컴퓨터에 파인튜닝된 모델을 탑재하면 작업이 용이해지고 시간을 절약할 수 있게 된다. 또한 복잡한 모델의 경우, 사용자 기기에서 실행하기 어렵기 때문에 서버 컴퓨터에서 실행함으로써 최소한의 리소스 사용이 가능하다. 이렇게 중앙 집중적으로 모델을 관리한다면 데이터를 보호할 수 있고, 운영 및 유지보수가 더 효율적으로 이루어질 수 있다.

이러한 빠른 응답, 데이터 보안, 리소스 최적화, 효율적인 운영 및 유지보수 등의 장점을 가진 Fast API를 사용하여 서버 컴퓨터에 웹 애플리케이션을 구축하였다. Fast API를 이용하여 서버를 설정하고 앱인벤터에 입력된 데이터를 받아온다. 받아온 데이터는 Llama2 모델에 입력되어서 출력하도록 설계하였다. 사용자의 질문을 입력으로 받아, 이를 json 형식으로 변환한 후 구축한 서버 컴퓨터 웹페이지로 전송한다. 이후, 서버에서는 Llama2 모델을 활용하여 입력 값을 처리하고, 모델의 출력 값을 가져와 앱 화면에 보여주도록 구현하였다.

## 5. 성능 평가 결과

공작기계에서 주요 Alarm을 중심으로 우선 총 30개의 QA 데이터를 평가 대상으로 선정하였다. 두 가지 모델을 각각 fine-tuned 버전과 baseline 버전으로 분리하여 총 네 가지의 모델에 대한 성능 평가를 진행하였다.

기존에는 baseline GPT-2, baseline Llama2, fine-tuned GPT-2, fine-tuned Llama2의 네 가지 버전을 모두 같은 기준으로 평가하려고 하였다. 그러나 평가를 위해 뽑은 30개의 데이터를 모

**Table 1.** Example Answers and Evaluation Results for GPT-2 and Llama2 Models

Question #1	I would like to ask a question about MANUAL GUIDE i of FANUC. I received an alarm says, "G codes that cannot be combined together. A figure block not available to the machining type block is specified." What should I check when this warning occurs?	
Answer about Q #1 (fine-tuned GPT-2)	G codes that cannot be combined together. A figure block not available to the machining type block is specified. Modify the machining program to specify a correct combination of machining type block and figure block.	O
Answer about Q #1 (fine-tuned Llama2)	To resolve this warning, check the machining program to ensure that the necessary machining type block is available. Modify the program by replacing the machining type block with an appropriate one.	O
Question #2	I would like to ask a question about MANUAL GUIDE i of FANUC. What should I do if alarm number 3535 is raised?	
Answer about Q #2 (fine-tuned GPT-2)	The groove depth specification is invalid. A value not specifiable as turning or other cut directions is entered. Modify the machining program to specify an appropriate cut direction.	X
Answer about Q #2 (fine-tuned Llama2)	The thread lead specification is invalid. A value not specifiable as the thread lead of a screw thread is entered. Modify the machining program to specify an appropriate value. (Note) The thread lead is the angle under which a thread is made. It is a positive value.	O

델에 입력하여 답변을 얻은 결과 baseline의 두 가지 모델은 학습된 모델들에 비해 답변의 질이 떨어지고, 모델의 성능이 부족한 결과가 나온다. 따라서 제안한 fine-tuned GPT-2와 fine-tuned Llama2의 두 가지 버전을 평가하였다.

### 5.1 정성적 성능 평가

fine-tuned GPT-2 모델과 fine-tuned Llama2 모델을 비교 평가하는 방법으로는 예시 질문에 대한 답변이 모범 답변과 비교하여 얼마나 정확한지에 대한 정성적 평가 기준을 부여하여 팀 구성원 4명이 각각 평가했다. 해당 평가에 대한 과반의 의견을 점수로 환산하였다. 평가 기준은 다음과 같다. 1) 질문에 대하여 정확히 맞는 답변 또는 해결 방안을 제시하는 경우: O, 1점 부여. 2) 질문에 대하여 정확히 맞는 답변을 포함하지만 관련 없는 내용을 함께 포함한 경우, 질문에 대한 답변이 틀리지는 않았으나 표현이나 용어 선택이 애매하여 작업자의 판단에 의존해야 하는 경우: △, 0.5점 부여. 3) 질문에 대한 답변이 잘못된 경우: X, 0점 부여.

<Table 1>은 질문에 대한 fine-tuned GPT-2, Llama2 모델의 답변 예시 및 평가 결과이다. <Table 2>는 평가 결과를 종합한 값이다.

**Table 2.** Example Answers to Questions by Model

GPT-2		Llama2	
O	8	O	18
△	3	△	6
X	19	X	6
Score	9.5	Score	21

### 5.2 정량적 평가

GPT-2 모델 답변, Llama2 모델 답변에 대해서 Sentence Bert 모델을 사용하여 유사도를 계산했다. Bert 모델은 텍스트의 의미론적 유사성을 파악하는 데 사용된다. 정성적 평가에 사용한 30개의 질문에 대해서 QA 데이터 세트에서 답변과 각 모델의 답변 유사도를 계산했다. Sentence Bert의 구조는 <Table 3>과 같다. 먼저 문장들이 Bert 모델을 통과하고 풀링(pooling)이 진행되면 임베딩 벡터  $u, v$ 가 생성된다. 이후  $u$ 와  $v$  벡터의 코사인 유사도를 계산한다.

정량적 평가에서 GPT-2 유사도는 0.72, Llama2 유사도는

**Table 3.** Quantitative Evaluation Results of Model-Specific Responses

Question	GPT-2 (Base)	GPT-2 (fine-tuned)	Llama2 (Base)	Llama2 (fine-tuned)
#1	0.12337531	0.864207268	0.870586753	0.91112673
#2	0.18033087	0.817527831	0.579422951	0.76273024
...	...	...	...	...
#29	0.1883748	0.281591296	0.341123939	0.46806327
#30	0.2558555	0.728332758	0.529266596	0.89345062
Average	0.18950539	0.719869162	0.515347299	0.79055304

0.79로 Llama2 모델이 선정되었다. 추가적으로 QA 데이터셋을 만들지 않고 학습한 Llama2 모델에 대해서 유사도를 계산한 결과 평균 유사도가 0.56이 나왔다. 하지만 실험 조건이 다르기 때문에 QA 데이터 세트로 학습시키지 않은 모델을 QA 데이터로 학습시킨 모델들과 비교할 수 없다. 하지만 base-line 모델들과 비교했을 때 더 높은 것으로 보면 QA 데이터를 만들지 않고도 학습시킬 수 있다는 것이다. 정성적 평가와 정량적 평가를 한 결과 모두 fine-tuned Llama2 모델의 성능이 제일 좋아 최종 모델을 Llama2 모델로 선정했다. 향후 QA 데이터를 만들지 않고 모델 정확도를 높일 수 있는 방법을 개선하는 방향으로 연구를 진행할 것이다.

## 6. 토 론

본 논문은 공작기계 개별 메뉴얼 지원용 커스터마이징 LLM을 개발과정을 소개하였으며, 두 가지 LLM모델의 성능비교와 Fine-tuning 성능 결과를 비교하였다. 먼저 Manual LLM 서버-앱 구조를 통해서 다양한 공작기계 모델에 하나의 유저가 가변적으로 접속 가능함을 확인하였다. 또한 Local LLM을 통해 제조기업의 보안 요구에 적합하고, 개별 장비에 맞게 운영가능성을 확인하였다. 한편 성능측면에서는 Llama2가 정성적, 정량적 평가 모두 우수한 결과를 보였으며, Fine-tuning을 통해서 두 모델의 차이가 32%에서 7%까지 줄어두는 것을 확인하였다. 이로써 모델보다는 Fine-tuning의 중요성이 비교되었다. 또한 학습에 필요한 메뉴얼로부터의 QA데이터 구축이 커스터마이징 LLM 구축 성능에 영향을 많이 미치는 것을 볼 수 있다.

## 7. 결 론

### 7.1 기대 효과

Manual LLM이 상용화 이후 보급되면 공작기계를 비롯한 산업 현장에서 발생하는 문제점에 대한 빠른 지원 및 문제 해결로 생산성을 향상할 수 있을 것이다. 또한 전문 지식의 공유 및 보존의 장점이 있으며, 특히 새로운 작업자에 대한 교육 및 훈련시간을 절감할 수 있다는 기대효과가 있다. 앞서 서론에서 언급한 다양한 상황대처의 어려움, 노하우 기반의 문제 해결, 휴먼 에러 발생, 유희시간과 대기 손실의 발생 및 증가에 대한 문제를 해결할 수 있을 것이다.

### 7.2 한계점 및 향후 연구 방향

본 연구에서 사용한 메뉴얼 가이드북은 약 1,000쪽의 pdf로 FANUC의 CNC 기계에 한정되어 있다. 더 다양한 pdf를 수집하여 모델을 학습한다면 모델의 범용성을 높일 수 있을 것이다. 또한 다양한 pdf를 사용하는 만큼 pdf로부터 텍스트를 추

출했을 때 그 패턴을 자동으로 파악하여 QA 데이터를 만들 수 있도록 하는 것도 좋을 것이다.

장비나 기계를 다룰 때는 가장 작은 오류나 부정확한 정보조차도 큰 문제와 손실을 가져올 수 있다. 따라서 정보 제공 시 그 내용의 정확성과 신뢰성은 매우 중요한 요소로 작용한다. 질문과 답변 데이터를 생성할 때 사용한 모델은 해당 메뉴얼 가이드북을 학습한 모델이 아니다. 그래서 기본적인 정보나 지식을 가지고는 있지만, 특정 메뉴얼에 대한 구체적이고 세부적인 지식은 부족한 상황이다. 이로 인해 모델 답변에 대한 신뢰성 문제가 생길 가능성이 있다.

생성형 언어 모델은 자연스러운 문장을 만들어내는 데에는 능숙하지만, 너무 긴 문장을 출력하는 경향이 있다. 이런 특성은 일반적인 상황에서는 유용할 수 있으나, 실무 환경에서는 정보의 정확성과 함께 간결함도 중요하게 여겨진다. 긴 설명보다는 짧지만 분명하고 정확한 답변이 필요한 경우가 많기 때문이다. 이를 고려하여, 현재 모델의 한계를 극복하고, 답변의 길이뿐만 아니라 내용의 신뢰성도 높이려고 시도할 예정이다. 또한 추후 연구로 전문적인 평가를 위해서 공작기계 업체 전문가와 일관성 통계분석을 통한 평가방법을 개발하고 반복 query에 대한 평가 분석도 수행할 예정이다.

## 참고문헌

- Anagnoste, S., Biclesanu, I., D'Ascenzo, F., and Savastano, M. (2021), The role of chatbots in end-to-end intelligent automation and future employment dynamics, *Business Revolution in a Digital Era: 14th International Conference on Business Excellence, ICBE 2020*, 287-302, Springer International Publishing.
- Anzari, F., Sugiarto, D., and Luis, M. (2023), A fault diagnosis system for CNC hydraulic machines: A conceptual framework, *SINERGI*, 27(1), 65-72.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... and Amodei, D. (2020), Language models are few-shot learners, *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018), BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805.
- Khan, J. Y. and Uddin, G. (2022), Automatic code documentation generation using GPT-3, *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering*, 1-6.
- Li, C., Park, J., Kim, H., and Chrysostomou, D. (2021), How can I help you? An intelligent virtual assistant for industrial robots, *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 220-224.
- Li, Z., Huang, T., and Wang, L. (2023), Customizing large language models for financial question-answering tasks using augmented datasets, *Journal of Financial Data Science*, 5(2), 101-120.
- Liu, R. and Shah, N. B. (2023), ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing, arXiv preprint arXiv:2306.00622

- Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. (2021), Self-supervised learning: Generative or contrastive, *IEEE Transactions on Knowledge and Data Engineering*, **35**(1), 857-876.
- Michael, G., Peng, J., Marinov, M. B., and Ovtcharova, J. (2021), Research on fault diagnosis expert system of CNC machine tool based on expert knowledge, *2021 XXX International Scientific Conference Electronics (ET)*, 1-4.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018), Improving language understanding by generative pre-training, *OpenAI Blog*, **1**(8), 9.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019), Language models are unsupervised multitask learners, *OpenAI Blog*, **1**(8), 9.
- Son, S.N., Oh, J. M., Jin, H. G., Jang, C. H., Jeong, J. B., and Kim, K. T., (2024), Varco Arena: A Tournament Approach to Reference-Free Benchmarking Large Language Models, arXiv preprint arXiv:2411.01281 (2024).
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... and Lample, G. (2023), Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., ... and Scialom, T. (2023), Llama 2: Open foundation and fine-tuned chat models, arXiv preprint arXiv:2307.09288.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... and Polosukhin, I. (2017), Attention is all you need, *Advances in Neural Information Processing Systems*, **30**.
- Wang, Y., Zhang, X., and Chen, H. (2023), A chatbot system for predictive maintenance in manufacturing based on LLMs, *International Journal of Advanced Manufacturing Technology*, **127**(5), 987-1003.
- Weitzman, T. and Pyo, H. (2021), CloneBot: Personalized Dialogue-Response Predictions, arXiv preprint arXiv:2103.16750.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... and Rush, A. M. (2020), Transformers: State-of-the-art natural language processing, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38-45.
- Yoo, Kang Min, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim *et al.* (2024), HyperCLOVA X Technical Report, arXiv preprint arXiv:2404.01954.
- Yoo, K. M., Han, J. G., In, S. K., Jeon, H. W., Jeong, J. S., Kang, J. W.,

Kim, H. W. (2024), HyperCLOVA X Technical Report, arXiv preprint arXiv:2404.01954.

## 저자소개

**이예원** : 경희대학교 산업경영공학과에서 2024년 학사학위를 취득하고 현재 경희대학교 스마트팩토리연구소 연구원으로 근무 중이다. 연구 관심 분야는 스마트팩토리, 자연어 처리, 생성형 AI, Digital Twin, CAD/CAM이다.

**박주영** : 경희대학교 산업경영공학과에 재학 중이다. 2023년에는 경희대학교 스마트팩토리 연구실에서 학부연구생으로 활동하며 자연어 처리와 생성형 AI 관련 연구 및 프로젝트를 진행했다. 현재는 에드워드 코리아에서 인턴으로 근무하고 있다.

**김시연** : 경희대학교 산업경영공학과와 소프트웨어융합학과에서 2024년 학사학위를 취득하고 Universität Heidelberg에서 석사 과정에 재학 중이다. 연구분야는 머신러닝, 생성형AI, 자연어 처리이다.

**조성우** : 경희대학교 산업경영공학과에서 2024년 학사학위를 취득하고 현재 경희대학교 인공지능학과 대학원 석사과정에 재학 중이다. 연구 관심 분야는 자연어 처리, 인간-기계 인터페이스, 대규모 언어 모델 및 이미지 생성 모델과 같은 생성형 AI를 포함한다.

**임주명** : 성균관대학교 기계공학과에서 2003년 학사를, POSTECH 산업경영공학과에서 2012년 석박통합을 취득하였다. 스위스 로잔연방공대, 영국 케임브리지대학교, 독일 인공지능연구소 SmartFactoryKL에서 연구원으로 근무하였으며, 2018년부터 산업경영공학과 교수로 재직하고 있다. 연구분야는 스마트팩토리, 제조AI, CAD/CAM이다.