

학습을 위한 custom LLM service 개발

1. 목적

- 개인 맞춤형 학습 기회 제공
- 빠르고 간편한 학습 방법 제공
- 다양한 학습 방법 활용 가능

2. 사업 주제

학습을 위한 custom LLM service 개발

3. 일정표

일시	내용	비고
3. 21.	팀 구성 및 팀별 사업계획서 작성	목적, 사업주제 정하기
3. 28.	사업계획서 작성(역할)	도서관 활동 ,개인별 1쪽
4. 4.	발표용 ppt 올리기	사업계획서, SWOT
4. 11.	초레봉 산행	
4. 18.	핵심기술 정리하기	도서관 활동 수업
4. 25.	사업계획서 자료 정리	
5. 16.	체육대회	
5. 23.	계획서 완성하기	
5. 30.	보고서 및 ppt 만들기	
6. 13.	발표 및 평가	
6. 20.	발표 및 평가	
6. 27.	진로활동 생기부 정리, 진로활동 생기부 정리 피드백	
7. 11.	개인별 진로활동 발표	

4. 내용

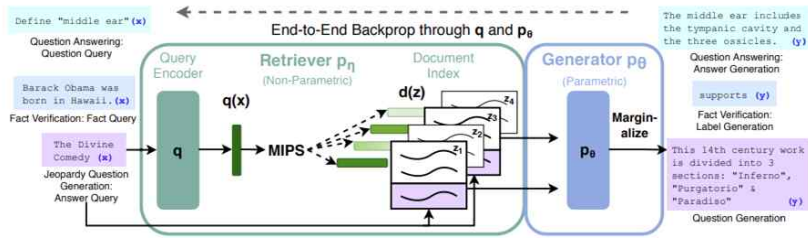
가. 평가주안점: 내용구성력, 창의성, 진로준비성, 발표력, 협동성

나. 참가자 및 역할

연번	학번	이름	역할(구체적으로 상세하게 기술)	비고
1	3102	권호승	<p>본인은 1인 기업의 운영자로서 모든 업무를 혼자 진행하여야 합니다. 우선 사업 아이템을 선정하여야 하나 이미 선정했으므로 건너 뛴겠습니다. 이후 학습용 AI에 관한 수요를 조사하기 위해 기존 시장에 존재하는 다른 서비스(예 : 수학대왕 등)와 내용 중복 여부를 확인해야 합니다. 또한 기존 사업의 사용자 수, 업무실적 규모를 조사하여 나의 사업에 대한 전망을 미리 고민해 볼 수도 있겠습니다. 사업 아이템의 성공 가능성에 대한 확신이 선 후에는 기술 조사에 진입합니다. 기본적으로 LLM model을 만드는 방법을 공부해야하는데, huggingface 등 사이트에서 사전학습된 모델을 사용하고 파인튜닝, RAG 기술을 이용하여 강화학습하는 방법, pytorch 등을 이용하여 직접 모델을 개발하는 방법 등 여러 방법을 강구합니다. 또한 서비스 제공을 위해서는 이 모델이 web 또는 app, 또는 기타 형태로 제공되어야 하기 때문에 만약 웹을 이용한다면 서버컴퓨터를 가동하는 비용, 또는 클라우드 서비스 등을 고려하고, 앱을 만든다면 사용자의 기기에 따라 java, kotlin, swift 등 추가적인 프로그래밍 언어에 대한 탐구가 필요할 것입니다. 서비스를 개발한 후에는 배포하는 방법을 공부해야 합니다. 웹을 개발한 후에도 local 환경에서 여는것과 hosting하는 것이 다르기 때문에 사이트 주소가 IP주소가 되는 상황을 막기 위해서는 도메인 설정 방법도 탐구해야 합니다. 그렇게 모든 것을 정하고 배포까지 완료하면, 유지보수와 경영을 해야합니다. 사용자가 오류로 인해 고통받을 수 있으니 삼성전자 원격 서비스를 본받아 화면공유를 통해 오류를 찾아내는 방법을 탐구할수도 있고, 이외에도 AS에대한 방법을 강구해야 합니다. 또한 재정상황을 살펴야하는데, 무엇을 제조하는 사업이 아니라 초기자금이 크게 들어가는 것은 아니나 수입이 적을 수 있습니다. 이 서비스에서 얻을 수 있는 수익이라고는 광고수익, 유료 서비스 수익 등이 있으나 광고의 경우 너무 많으면 사용자의 불편감을 야기할 수 있고, 유료 서비스의 경우 AI의 성능 조절 등을 통해 만들 수 있으나 본인의 사업 철학은 최대한 양질의 서비스를 제공하는 것이기 때문에 무료/유료 서비스 간 격차를 벌리기보다는 컴퓨터 구동 비용, 인건비, 사전학습 모델 이용료 등 지출을 탄력적으로 관리하여 기업의 재정을 안정화시키는 것이 큰 목표라고 볼 수 있겠습니다. 마지막으로, 기업이 파산할 때를 고려해야 합니다. 제가 대표기 때문에 사업 실패에 대한 책임이 고스란히 스스로에게 돌아와 예비 자금을 축적하고, 그러나 부당한 방법이 아닌 기업의 순이익을 증가시켜 사업 패망에 대한 대처를 해야합니다. 마지막으로 사용자에게 개인정보 이용 동의를 구하고 사용자가 입력하는 학습 데이터를 검토하여 저작권에 어긋나진 않는지 윤리적 측면에서도 살펴야 합니다.</p>	팀장

5. 사업계획서

사업 이름	나만의 학습 친구			
팀 이름	남방큰돌고래			
참가자 (학번/이름)	3102			
	권호승			
사업 주제	학습을 위한 custom LLM service 개발			
SWOT 분석	Strength	Weakness	Opportunity	Threat
	·전자 AI 시스템을 활용한 접근성 향상 ·학습에 대한 수요 저격 ·커스텀 서비스를 이용한 폭넓은 학습범위	·고기능의 하드웨어 필요 ·custom 자료 인식의 불안정성 ·저작권 문제에 대한 대처 미흡	·꾸준히 상승하는 교육 수요 ·사전학습된 모델 등 더 쉬운 개발을 위한 도구가 나오고 있음 ·개발 과정에서 빠른 피드백 가능	·해킹 ·정전 ·경쟁 제품의 등장 ·네트워크 불안정 ·나태
세부내용	<div>1.기술</div> <div>(1)LLM</div> <div>LLM은 Large Language Model의 약자로, 말그대로 인간의 언어를 이해하기 위해 문자의 형태로 입출력을 하는 AI 모델을 말하는 것이다. 이는 주로 ANN(Artifiical Neural Network, 인공신경망)으로 이루어진다. LLM은 결국 언어 모델이므로 언어 모델에 대한 기술적 이해가 필요하다. 모델은 언어를 덩어리로 이해할 수 없으므로 잘게 기본 단위로 쪼개는 작업을 반복해야 하는데, 이를 토큰화(tokenization)이라 하고 이를 수행하는 것을 토큰라이저(tokenizer)라고 한다.</div> <div>Split on spaces</div> <div><div>Let's</div><div>do</div><div>tokenization!</div></div> <div>Split on punctuation</div> <div><div>Let</div><div>'s</div><div>do</div><div>tokenization</div><div>!</div></div> <div>이후 단어를 모델이 이해할 수 있는 수학적 형태로 변환해야 하는데, 이를 임베딩(embedding)이라고 한다. 대표적 방법은 벡터를 활용한 방법으로, 코사인 유사도(벡터의 길이가 아닌 각도를 바탕으로 유사도 판단)를 이용하는데, 예를 들어 'big'과 'large'는 유사한 의미를 가지고 있기 때문에 수학적으로도 유사해야 한다.(벡터의 경우 사잇각이 작아야 한다)</div>			

<p>세부내용</p>	<p>여기까지는 모든 언어모델이 공통적으로 수행하는 작업이고, 이후로는 LLM에 주로 사용되는 트랜스포머의 구조에 대해 설명하도록 하겠다.</p> <p>트랜스포머에는 포지셔널 인코딩(positional encoding)과정이 추가되는데, 문장 내 토큰의 위치에 따라 달라지는 벡터로 단어의 순서를 유지하여 어순을 이해하도록 하는 것이다.</p> <p>이후 벡터들은 Self-Attention Layer로 전달 → Feed-Forward Network→Softmax Layer로 전달되고, Softmax Layer를 통해 얻은 가장 확률이 높은 답변을 출력하는 것이다.</p> <p>하지만 LLM은 여러 단점이 존재한다. 첫째, 학습 이후에 생성되는 최신 정보에 대해서 무지하다. 이는 매우 자명하다. 둘째, 모델이 해당 답을 출력한 인과관계를 파악하기 어렵다. 인공 신경망의 파라미터 개수는 매우 많기 때문에 그 복잡한 구조를 모두 파악하기란 불가능하다. 셋째, 사실과 다른 답변을 내놓을 수 있다. 인공신경망은 결국 수학적 확률에 기반하여 답하기 때문에 인간의 뇌와 같이 명백하게 아닌 사실에 대해 아니라고 답할 수 없다. 논리에 기반해 답하는 것이 아닌, 명백하게 아니라는 것을 사전에 학습하거나 50%만 사실이라면 50%정도의 애매한 답변을 내놓는 것이다. 이를 개선하기 위해 여러 기술들이 고안되고 있다.</p> <p>(2)RAG(Retrieval-Augmented Generation)</p> <p>RAG 기술은 parametric memory와 non-parametric memory를 결합한 기술이다. parametric memory은 parametric이란 단어에서 알 수 있듯 인공신경망이 사전에 학습한 지식을 의미하고 non-parametric memory은 외부 정보를 나타내는 것이다. 이는 텍스트, 문서 등이 될 수 있다.</p>  <p>질문 x를 받아서 외부문서 z를 검색하고 답변 y를 내놓는다. 논문에서 RAG-Sequence와 RAG-Token 모델이 소개되는데 RAG-Sequence는 하나의 문서를 선택하여 전체 시퀀스를 형성하고 RAG-Token 방법은 토큰별로 서로 다른 문서를 활용하는 것이다.</p>
-------------	---

<p>세부내용</p>	<p>출처 : Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.</p> <p>Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.</p> <p>https://wikidocs.net/166796 https://kevin-rain.tistory.com/170</p> <p>중간고사 기간동안 시중 LLM model을 이용하여 테스트를 진행하였다. 한국사, 경제 과목 학습지를 주제로 문항을 제작하고 답변을 받아 평가하는 방식이었다. 그러나 몇가지 문제점이 있었다.</p> <p>1.학습지의 구성 문제 - 수업시간에 활용하는 학습지의 경우 본문이 있고 빈칸에 필기를 하는 방식으로 진행된다. 이 경우 전개가 일목요연하지 않고 필기내용이 중구난방하게 있기 때문에 이해하기 어려울 수 있다. 나의 경우 필기자료를 포함해 새로운 파일로 정리하였기 때문에 이해의 부족은 나타나지 않았으나 원본파일을 그대로 올렸을 때의 부작용은 아직 충분히 실험해봐야 한다.</p> <p>2.문제의 난이도 - LLM이 제시하는 문제의 난이도가 매우 천차만별이다. 단순히 O/X퀴즈를 내는 것에서부터 사소한 것을 서술형 문항으로 출제하는 등 난이도가 일정하지 않다. 이 경우 사용자가 여러 차례 명령을 내려(프롬프트 엔지니어링을 하여) 난이도를 조절해야 하는데, 이 경우 조금씩 발전되긴 하지만 안그래도 바쁜 시험기간에 프롬프트 엔지니어링에 시간을 투자할 학생은 없을 것이기 때문에 시스템 개발자 차원에서 해결해야 할 문제라고 본다.</p> <p>3.문제의 형식 - 어떤 문제는 객관식, 어떤 문제는 서답형, 어떤 문제는 서술형으로 나온다는 것이 학생의 입장에서는 감이 잡힐 것이다. 그러나 모델은 그것을 모른다. 물론 가장 좋은방법은 학습지를 백지에 그대로 베껴쓸 수 있을 정도로 외우는것이기 때문에 모든 문제를 서술형/서답형으로 내달라고 하면 문제가 없다. 그러나 이런 경우 학습의 의미가 퇴색된다. 학습 과정에서는 객관식 --> 서답형 --> 서술형으로 이어지며 점진적으로 이해도와 암기내용을 늘려나가야되는데, 서술형만 요구하게 된다면 그것은 학습 도우미가 아닌 단순 사실 모의고사에 불과하게 된다는 단점이 있다.</p> <p>4.모델의 지능문제 - 모델이 기본적으로 효과적이지 못한 부분도 있다. OpenAI라는 세계적인 기업의 수많은 사용자가 사용하는 ChatGPT를 사용했음에도 불구하고 문제의 형식을 자기 멋대로 내거나 채점을 대충 진행하는 등(애매한 표현도 맞다고 해준다) 모델이 사람의 지능을 따라오지 못해 생기는 문제가 존재한다.</p> <p>5.모델의 사적 개입 - 학교 시험은 선생님이 말씀하신대로 공부하는 것이 중요하다. 그러나 LLM은 사전지식과 검색기능을 보유하였다 보니 역사나 경제의 경우 본인의 지식을 첨가하여 말하는 경우가 많았다.</p> <p>6.텍스트 이외 형식에 취약 - 공부를 하다보면 표나 그래프형태로 나타나는 자료도 있을 것이다. 나의 경우 세금의 종류를 표 형태로 외웠다. 그러나 모델에게 표에 빈칸을 뚫어 채우는 문제를 내달라고 하면 잘 못낸다. 수십번 수정하고 지적하고 나서야 정상적인 문항 제작이 가능해졌고, 이마저도 여러번 반복하면 다시 오류를 일으켰다.</p>
-------------	---

세부내용

이를 통해 나는 앞으로의 방향성을 잡게 되었다. 이들은 공통적으로 LLM model의 성능향상보다는 원본 데이터와 그 가공방법에 대해 더 심혈을 기울여야 한다는 것이다.

1.키워드 중심의 채점기준 제공 - 실제 시험에서도 채점기준을 제공하여 특정 단어 또는 수식이 들어갈때마다 부분점수를 받는 경험이 있을 것이다. 인공지능의 경우 주관식 서술에 특히 취약하므로 채점기준을 제공하여 보다 객관적 척도를 제공한다는 것이다. 이는 텍스트 형식으로 입력할수도 있으며 학습지 파일에 형광펜을 칠하는 등의 방식으로 표기할 수 있다.

2.기본 틀 제공 - 표를 직접 생성하라 하지 않고 기본 틀을 제공한 뒤 글자만 채우도록 한다. 이 경우 사용자가 보는 것은 표지만 실제 컴퓨터가 다루는 자료구조는 텍스트이므로 더 정확한 작업이 가능할 것이다.

3.답지 제공 - 그래프의 경우 AI가 출제하기도 어렵고 채점하기도 매우 어렵다. 따라서 이 경우 모델이 캡처를 해 놓은 뒤 문제를 내고 답을 사진으로 제공하는 것이다. 이 경우 본인이 직접 채점하기 때문에 정확성 문제가 해결되고, 그래프 문제를 아예 내지 않는다는 기존 문제도 해결한다.

4.예시 문제 제공 - 모델이 문제의 난이도를 조절할 수 있도록 예시 문제를 제공해 학습시킨다. 사실 이 방법은 별로 효과적이지 않을 것이라는 느낌이 든다. 예시를 제공해줘도 맥락을 못잡고 따라하지 못하기 때문이다. 하지만 시도해 볼 가치는 있다고 판단된다.

5.답변의 형식 생성 - 모델이 자신의 상식을 투입시킨다는건 굉장히 큰 문제다. 그렇다고 모델을 순전히 개인 데이터로만 학습시키기에는 일반 언어능력이 부족할 수 있고 학습데이터가 너무 부족하다는 단점이 있다. 따라서 기존 훈련된 모델을 사용해야 하는 것은 변함이 없다. 이 경우 답변의 형식을 제한하는 것이 대안이 될 수 있다. 답변의 형식을 제한하는 것은 두 가지로 나눌수 있는데, 첫 번째로 형식을 제공하는 것이다. 예를들어 이름과 업적을 맞추는 문제를 낼 때 ----- : -----의 형식을 제공한다면 모델은 다른 수식어 없이 틀만 채울 것이다. 두 번째는 금지하는 것이다. 형용사 금지, 감정표현 금지, 심지어는 문장 금지(날말만 사용) 등을 통해 사적 개입을 효과적으로 막을 수 있다고 기대한다.

6.학습지 정리 시스템 - 학습지에 필기가 중구난방하게 되어있어 AI가 이해하기 어렵다는 단점이 있지만, 이 경우 의외로 쉽게 해결한다.



이렇게 학습지의 틀을 근본적으로 바꾸는 것이 아닌 여백 사이에 필기내용을 텍스트형태로 정리하는 것만으로도 기계가 쉽게 이해할 수 있다. 이것은 나의 경험담이다. 1번 페이지와 2번 페이지의 내용을 번갈아 보는 등의 작업은 기계가 잘 하므로 여러군데 퍼져있는 내용을 모아오기만 하면 되는 것이다. 따라서 이것을 쉽게 해결할 수 있는 시스템을 만들어야 하는데, AI에게 맡기면 주관이 들어갈 수 있으니 사람이 수작업으로 하거나 기계적으로 해결하는 것이 좋아보인다. 가급적 이 정도는 스스로 하자. 공부 차원에서.

참여 동기
사업 전망

나는 오래전부터 소위 “하브루타” 또는 그 외 토론식 공부에 대한 로망을 가지고 있었다. 로망을 가진 것 뿐 아니라 중2때부터 꾸준히 실천해왔으며 나름 쓸쓸한 효과를 보기도 했다. 그러나 고등학교에 올라오고 나니 친구들이 생각보다 바빠 시간적, 공간적으로 하브루타를 할 약속을 맞추기 어렵다는 것을 깨달았다. 또한 친구가 없는 아이의 경우 애초에 할 사람이 없다는 것도 단점이다. 따라서 인간을 닮은 LLM에게 하브루타 상대가 되기를 요청하는 마음에서 이 프로젝트를 시작하였다.

이 사업이 성공한다면 전망은 더할 나위 없이 좋을 것으로 예상된다. 안그래도 칸아카데미, 쿠팡, 수학대왕 등 온라인 학습 프로그램과 AI교육이 꾸준히 등장하고 있고, 교육에 대한 수요는 감소할 일이 거의 없기 때문이다. 또한 기존 프로그램은 수학 문제 풀이 등 보편적이고 공통적인 주제만 다뤘다면 이 프로그램은 한국어, 경제, 화법과작문 등 개인적 과목, 특정 선생님의 학습지 등을 학습시킬 수 있다는 점에서 매우 유용한 면모를 보인다. 하지만 역시 문제는 기술력이다. AI의 부족함이 몇주간의 실험에서 여실없이 드러났기 때문에 이를 해결하는 것이 사업의 전망을 가를 것이다. 하지만 가능성은 무한하다. 왜냐하면 기술력이라는 것이 단순히 복잡한 코딩과 고성능 컴퓨터만을 의미하지는 않았기 때문이다.



이를 보면 이미 문제 출제 기술은 현재 수준으로 충분히 가능하다는 것이 확인된다. 채점도 해준다. 하지만 내용적으로 실패한 것이므로 이는 프로그램의 형식, 틀, 원본 데이터의 수집과 처리가 중요하다는 것을 의미한다. 따라서 충분한 시간과 인력, 아이디어만 있으면 성공적인 사업이 될 수 있다고 판단한다.