## 



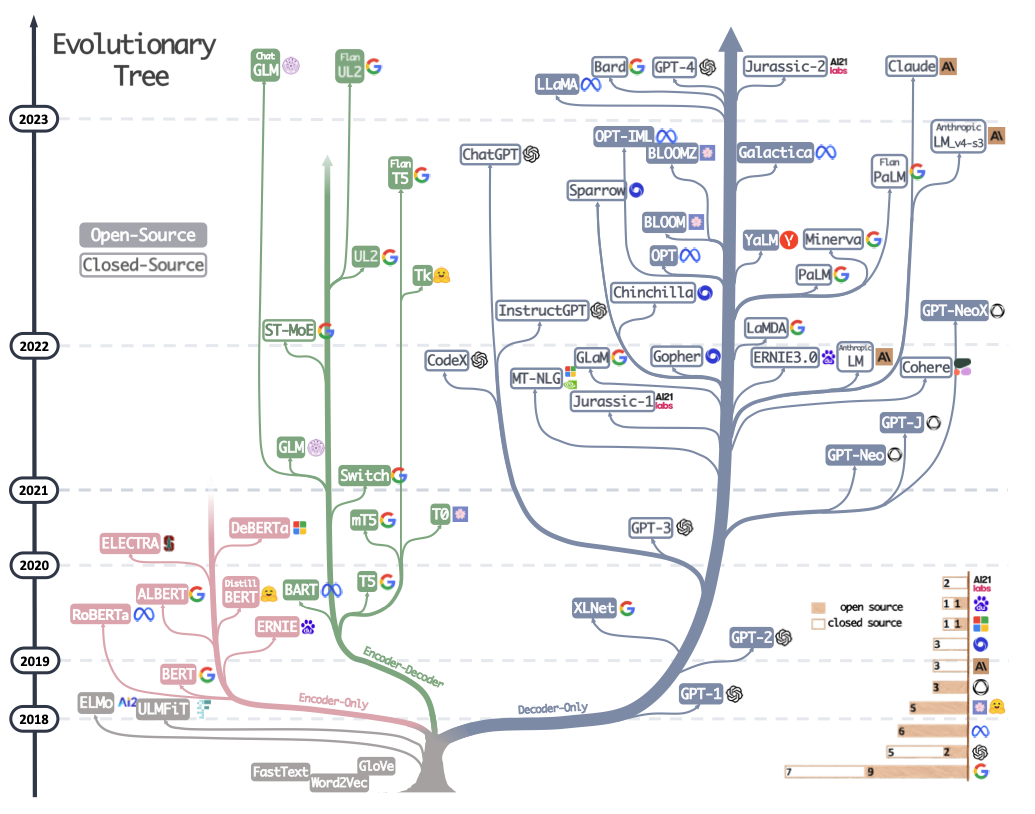
고객 편의를 위한 LLM 솔루션  
 *LLM is All You Need*

2024년 1월 17일 **작성자 이름**전진환 신동원 전대엽 유재현

# 개요 및 주제 **LLM(Large Language Model)의 발전은 다양한 분야에서 맞춤형 솔루션 제공의 가능성을 열었습니다.** 또한 **개별 사용자의 특정 필요와 선호에 맞추어 LLM을 미세 조정(Fine Tuning)하는 것이 큰 관심사가 되고 있습니다. 본 프로젝트는 이러한 기술의 최신 동향을 반영하여, 사용자에게 보다 이해하기 쉽고 편리하며 효과적인 맞춤형 LLM 솔루션을 제공하는 것을 목표로 합니다.**

**동기**

**∙ Bard와 ChatGPT의 등장으로 언어 이해 및 생성 분야에 뛰어난 능력을   
 발휘하며 다양한 분야에서 혁신을 주도하고 있습니다.  
∙ Gemini와 같은 멀티 모달 모델의 시대가 도래하면서, 영상, 음성, 텍스트,   
 이미지 등 다양한 데이터를 인식하는 새로운 가능성이 열렸습니다.  
∙ LLM의 발전이 가져온 사회적 영향과 응용 분야에서의 혁신적 변화는   
 프로젝트의 주요 동기가 되었습니다.**

****[LLM history]

**구현 서비스**

1. PaML2 / Llama2 + LangChain + RAG + PEFT 기능 구현
2. Gemini + LangChain 기능 구현
3. 실험 및 성능 비교
4. PaML2 / Gemini / Llama2를 활용한 맞춤형 LLM 제작
5. 개인별 개인화된 서비스 제공

**진행상황 중 극복 과제**

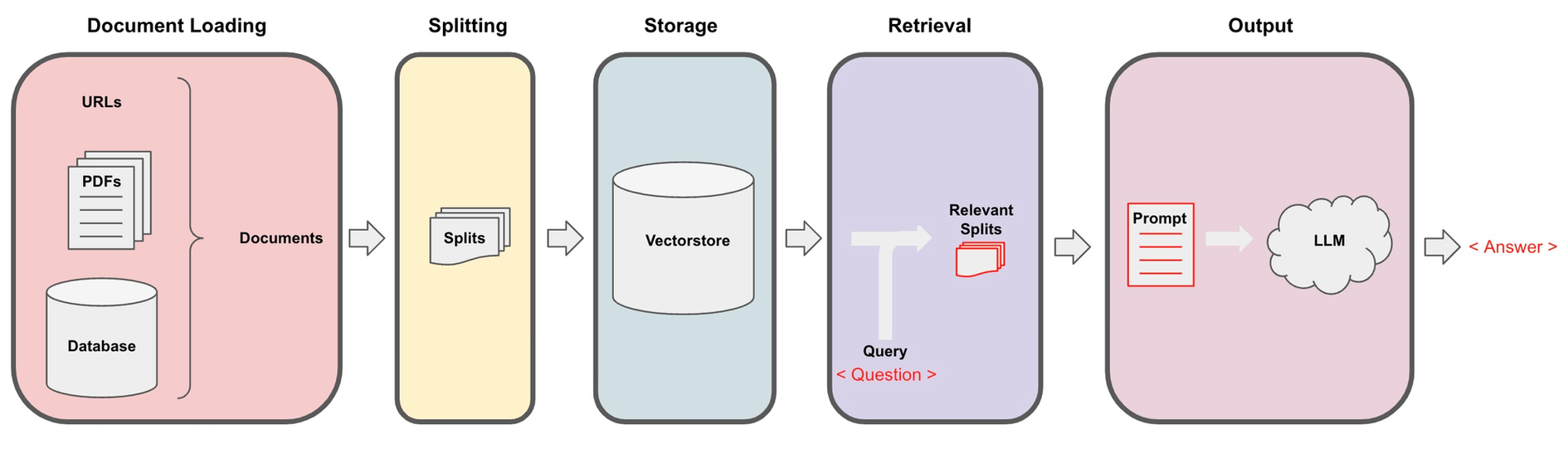
Gemini pro는 2024년 1월 현재 Release 초기 버전으로써 한국어 인식, LangChain 및 PEFT(Parameter-Efficient Fine Tuning) 적용이 어려운 기술적인 상황에 있습니다.

또한 LLM 모델의 PEFT 적용 및 운영을 위해 상당한 컴퓨팅 자원과 비용이 필요합니다.

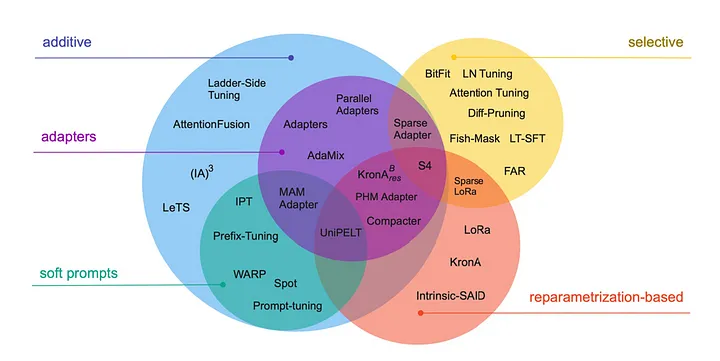
Gemini와 PaLM2에 필요한 라이브러리인 GCP vertex ai, generative AI, LangChain 등의 잦은 버전 변경 및 샘플 코드의 희박성으로 인해 프로젝트를 진행함에 있어 기술 구현에 어려움을 겪고 있습니다.

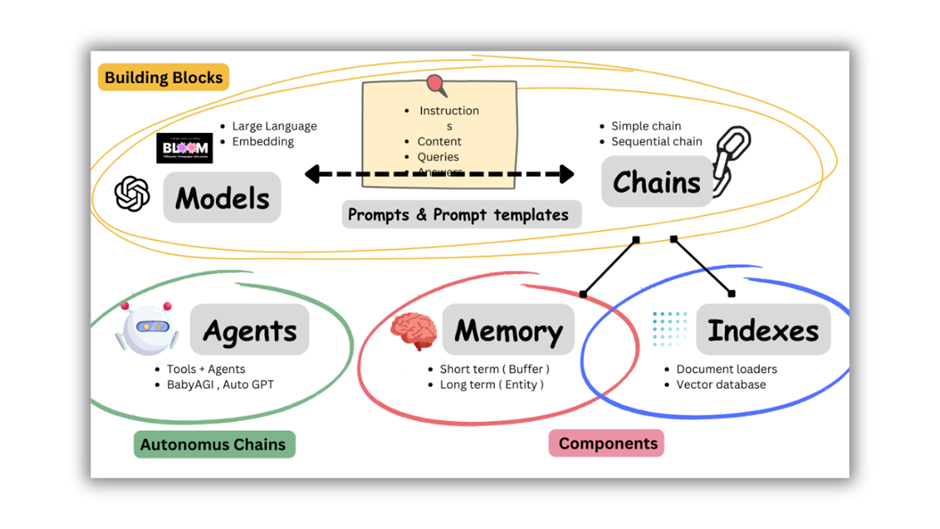
# 관련 서비스 동향

최근에는 LLM을 사용하여 여러 Tuning 기법이 적용 된 사용자 편의성을 고려한 개인화 된 LLM 서비스와 이것을 공유하는 형태의 플랫폼이 등장하여 수익 창출이 가능한 형태로서 서비스 중입니다.

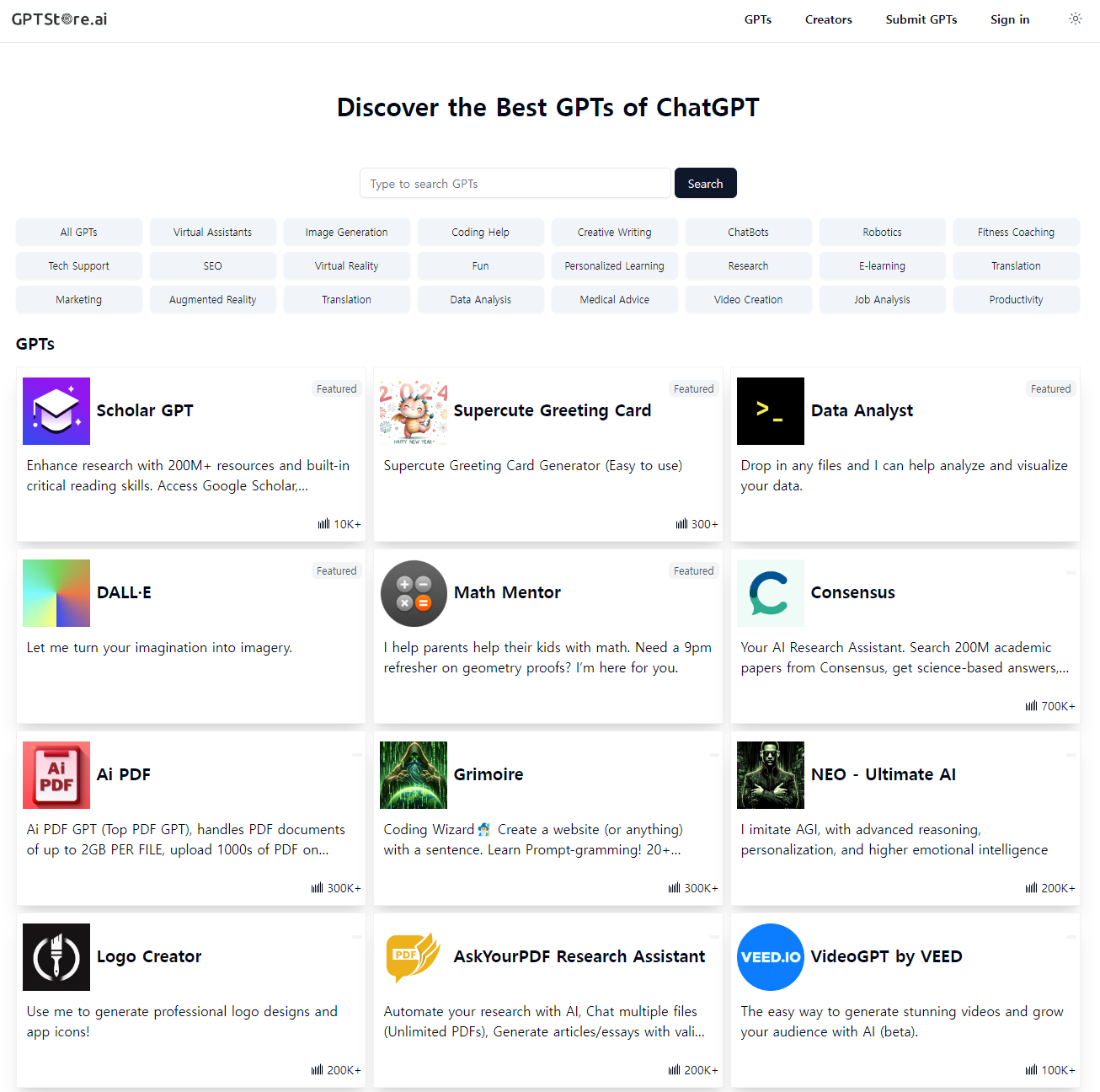
- RAG(Retrieval-Augmented Generation)를 적용한 LLM 구조 

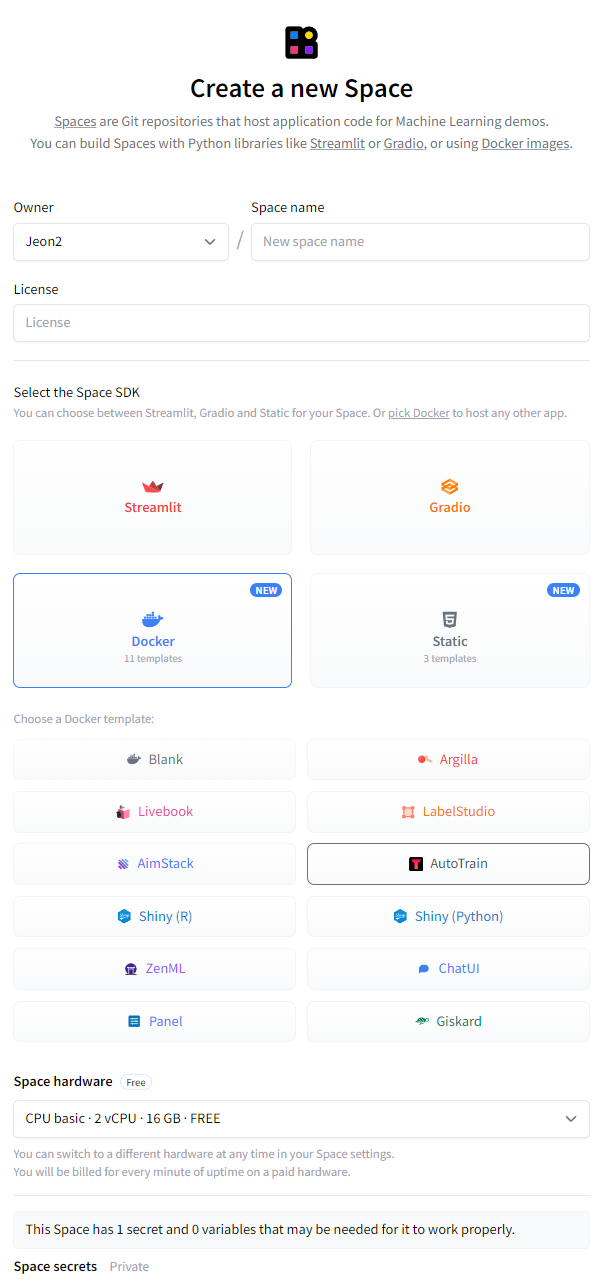
pdf, text, image, video 등의 다양한 타입을 가진 데이터들의 입력값을 참고하여 Retriever(문서 검색)의 결과와 vector store 상의 유사도 비교를 통해 도출된 tokenizer(토크나이저) 된 단어들이 LLM에 입력됩니다. 그 후에 사용자에게 필요한 관련 도메인 지식으로 도출 결과를 한정할 수 있는 Fine tuning(미세조정)을 할 수 있다는 장점이 있습니다.

- 적용 가능한 PEFT 기법  
Full Fine-Tuning은 PEFT에 비해 더 많은 시간과 비용이 소요됩니다. 하지만 PEFT기법은 시간대비 효율이 좋아 모델의 성능을 개선하는데 큰 역할을 하고 있습니다. 그 중 LoRA, prefix tune, prompt engineering, P-tuning 등의 미세조정(Efficient Fine Tuning) 방법이 선호 되고 있습니다.

- LangChainLangChain은 기업에서 상업적으로 서비스 하는 유료 LLM과 오픈소스로 배포된 무료 LLM의 서로 다른 스펙을 하나로 맞춰 주는 것을 목표로 하고 있습니다. 이에 현 프로젝트에서 진행하고 있는 다양한 LLM을 개발하는데 있어 복잡성을 크게 줄여줌으로 전처리부터 학습, 배포, 최적화 및 타 API를 chain이라는 개념을 통해 개발 가능하게 해 주는 편의성을 제공하고 있습니다.

**- Open AI ChatGPT Store**



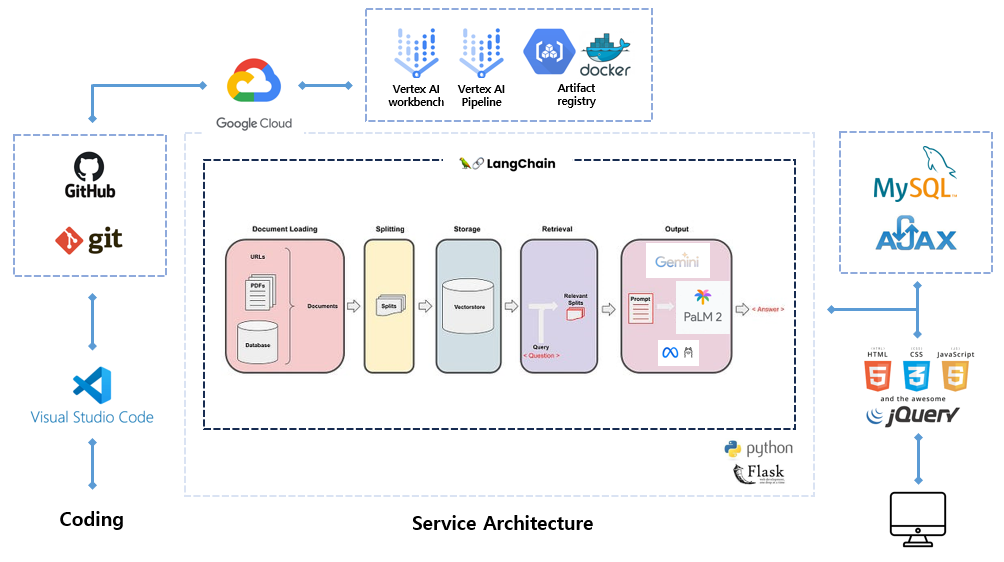
**- HuggingFace Autotrain**

노코드(No-code) 플랫폼으로 코드 사용이 어려운 일반인들도 일정 구독료를 지불하면 개인이 원하는 형태의 도메인과 스타일로 customize된 LLM을 사용할 수 있습니다.

sLLM과 vLLM의 등장으로 온디바이스의 독립된 형태로 사용자가 이용할 수 있을 것이며, 이와 관련된 시장이 생성될 것으로 예상이 됩니다.

# Programing Language & APIs

**Project Specs  
Backand :** Python / Flask / LangChain **Frontend :** HTML / CSS / jQuary **Models:** Google AI Gemini / Vertex AI PaLM2 / Llama2 **Cloud:**  Google Cloud Platform  
**SCM:**  git  
**DB :** MySQL  
**Etc:**  Docker

****

# 환경설정

1. Local Linux 서버 환경 구성 및 프로토타입 제작

* ubuntu 22.04 / rtx4090 24 gb x 2 / Ryzen Threadrippper Pro 5955WX 16 core
* Nvidia Container Toolkit 셋팅
  + [https://github.com/NVIDIA/nvidia-container-toolki](https://github.com/NVIDIA/nvidia-container-toolkit)
* 프로토타입 결과
  + <https://github.com/DONGWEONSHIN/llm_is_all_you_need>

1. GCP Vertex AI

* Vertex Ai Pipeline등을 활용한 LLMOps (2단계) 배포 자동화 (계획 중)

**관련 자료**

* **프로젝트 진행상황** 
  + <https://wjs545.notion.site/LLM-is-All-You-Need-6258d59d36e44c79832cde148f2934e5?pvs=4>
  + <https://github.com/DONGWEONSHIN/llm_is_all_you_need>
* **RLHF** 
  + <https://tech.scatterlab.co.kr/luda-rlhf/>
  + <https://colab.research.google.com/github/GoogleCloudPlatform/vertex-ai-samples/blob/main/notebooks/official/generative_ai/rlhf_tune_llm.ipynb>
* **HuggingFace PEFT ( LoRA / prefix tune / Prompt tuning / P-tune …)**
  + <https://pypi.org/project/peft/>
  + <https://medium.com/international-school-of-ai-data-science/optimizing-performance-with-peft-a-deep-dive-into-prompt-tuning-2b9a17bc9851>
  + <https://velog.io/@mmodestaa/GPT-3-%EB%93%B1%EC%9E%A5%EA%B3%BC-%EA%B7%B8-%ED%9B%84-Prompting-and-Promt-Tuning-Prefix-Tuning-P-Tuning>
  + <https://www.promptingguide.ai/kr/techniques/>
  + <https://learn.deeplearning.ai/chatgpt-prompt-eng/lesson/1/introduction>
* **LangChain tutorial**
  + 공식 docs : <https://python.langchain.com/docs/additional_resources/tutorials>
  + GCP Blog : <https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-applications-with-vertex-ai-palm-2-models-and-langchain?hl=en>
  + Gemini : <https://github.com/sugarforever/LangChain-Tutorials/blob/main/LangChain_Google_Gemini_API.ipynb>
  + vertex AI : <https://python.langchain.com/docs/integrations/llms/google_vertex_ai_palm>
  + DeepLearning AI : <https://learn.deeplearning.ai/langchain/lesson/7/agents>
* **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks**
  + <https://jeonsworld.github.io/NLP/rag/>
  + <https://arxiv.org/pdf/2005.11401.pdf>
* **llama**
  + <https://github.com/jmorganca/ollama>
  + [https://dajeblog.co.kr/%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-llama2](https://dajeblog.co.kr/%EB%85%BC%EB%AC%B8-%EB%A6%AC%EB%B7%B0-llama2/)/
  + <https://scontent-gmp1-1.xx.fbcdn.net/v/t39.2365-6/10000000_662098952474184_2584067087619170692_n.pdf?_nc_cat=105&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=PCPX3Lee5ooAX91gFNY&_nc_ht=scontent-gmp1-1.xx&oh=00_AfDXTMEmu2plrfsh-vYRkhNtMOZgkYKWeE6LutWEW2AuYw&oe=65AFA97F>