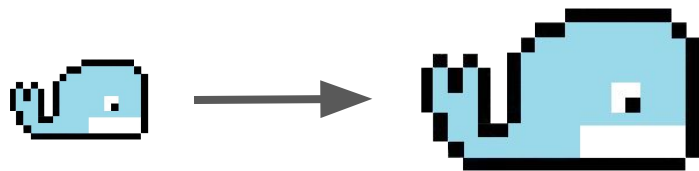


LLM In Production



24.03.22

그랩(이호연)

0. 목차

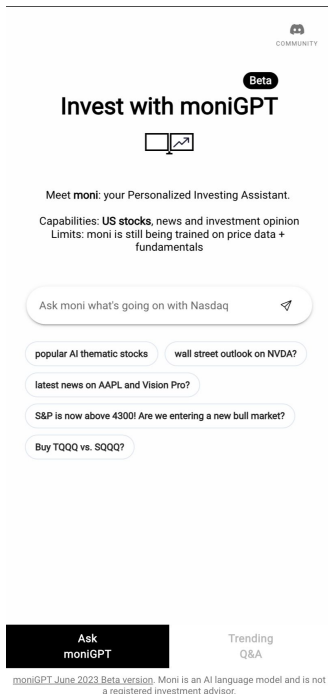
1. Intro
2. Super News (LLM Application)
3. Reducing Hallucination
4. Building Production LLM App
5. LLMOps

1. Intro > 소개



- 그랩(이호연)
- **(現) Tech Lead @ Project Pluto**
- (前) Software Engineer(Data) @ Socar
- (前) Software Engineer @ Class 101
- (前) Growth Engineer @ Market Fit Lab
- ...

1. Intro > 오늘 할 이야기



- **LLM powered Application**을 운영하면서 경험한 것들을 공유합니다.
- **RAG, Hallucination, LLMOps**에 대해 주로 이야기할 예정입니다.
- Model Training, Fine Tuning 과 관련한 이야기는 안할(못할) 예정입니다 (인원 확충이 되서 뭔가를 좀 해본다면..)

2. RAG In Production > Super News

The screenshot displays the 'SUPER NEWS' website. At the top, there's a navigation bar with links like 'Wall Street Live', 'Chart Tracker', 'Headlines', 'Markets', 'Finance Chatbot', and 'Opinion'. A search bar with 'Hoyon' and a magnifying glass icon is on the right. Below the navigation bar, a dark banner shows various market indices: SPX 5,224.52 ▲ 0.02%, NYM 62.961.12 ▲ 0.02%, VIX 13.10 ▼ 0.12%, US3M 4.37 ▼ 0.02%, DXY 93.44 ▲ 0.02%, BTC 61,311.42 ▲ 0.02%, INFLC 91.22 ▼ 0.02%, and GOLD 2,100. Below this, the main content area features a large article titled 'Fed's Dovish Outlook Sparks "Everything Rally"' with a photo of the US Capitol. To the right of this article are three smaller snippets: 'Fed's Rate Cut Forecast Fuels Stocks, Bonds Rally', 'Financials Surge, Oil Rallies as Fed Decision Looms', and 'Luxury Brands Plunge on China Spending Fears'. Below the main article, there are four columns of news: 'Wall Street Live' (listing Fed's Rate Cut, Speculation Fuels Dollar Bears, etc.), 'Macro' (AI Boom and GDP Growth Propel Stocks, etc.), 'Equities' (Market Records Broken as Momentum Shines, etc.), and 'Most Read' (Oman Investment Authority Plans IPO, etc.). At the bottom left, there's a link to 'Get Personalized Newsletter'.

Super News

- LLM-Powered Financial News Platform
- 미국을 주요 타겟으로 US Equities, Macro, Crypto 등 각 분야별 콘텐츠 생성중
- 하루 최대 **1천** 개의 기사 생성

2. RAG In Production > Super News

SUPER NEWS

Sign Up Hoyeon

Wall Street Live Chart Tracker Headlines Markets Finance Chatbot 117% Opinion

SPX 5,224.52 ▲ 0.02% NIK 16,245.13 ▲ 0.02% VIX 13.50 ▼ 0.02 USDT 4.87 ▼ 0.02 BNY 10.44 ▲ 0.02 BTC 61,311.42 ▲ 0.02 ETH 3,024.12 ▲ 0.02 GOLD 2,100.00

Fed's Dovish Outlook Sparks "Everything Rally"

7 hr ago

Fed's Rate Cut Forecast Fuels Stocks, Bonds Rally

Financials Surge, Oil Rallies as Fed Decision Looms

Luxury Brands Plunge on China Spending Fears

5:00, 22:40 KST

Wall Street Live

Macro >

2 hr Fed's Rate Cut Speculation Fuels Dollar Bears, Caution Urged

4 hr Fed's Hawkish Dot Plot Threatens EM Bonds, Currencies

4 hr Fed and BOJ Dovish Tones Calm Markets, Yen Falls

5 hr BOJ Moves Hint at Shift, Equities Over Bonds for Japan

6 hr Silver's Surge: Poised to Break \$30 Amid Fed Cuts

Equities >

AI Boom and GDP Growth Propel Stocks to Record Highs

Turkey's Gold Smuggling Hits Record Amid Economic Turmoil

Iberdrola's \$39.3B Bet on US Grids, Eyes Dividend Hike

Market Records Broken as Momentum Stocks with Strong Balance Sheets Shine

Li Auto Adjusts Q1 Delivery Forecast Amid EV Market Challenges, Outperforms Peers

Challenges of Outsourcing in Car Manufacturing Amid EV Sector Downturn

Most Read

- Oman Investment Authority Plans IPO for Asyal Group to Diversify Economy
- FDIC, Rialto Face \$5M Lawsuit Over Signature Bank Loan Fallout
- Nvidia Unveils Blackwell Architecture with Revolutionary R200 Chip
- Sanctions Snarl Russian Crude, Tankers Idle Off India
- Aircraft Insurance Dispute Over Seized Planes Sparks Legal Battle
- USB Unveils High Conviction Picks with Notable Upsides

Get Personalized Newsletter

Super News

플랫폼에서 생성되는 모든 콘텐츠는 **Fully LLM**으로만

- 미국 증시 뉴스
- Chart Tracker(차트, 테이블 해석)
- Finance GPT(Q&A 봇)
- (WIP) Newsletter

2. RAG In Production > Super News



Gemini



다양한 LLM

THE WALL STREET JOURNAL
WSJ



yahoo!
finance

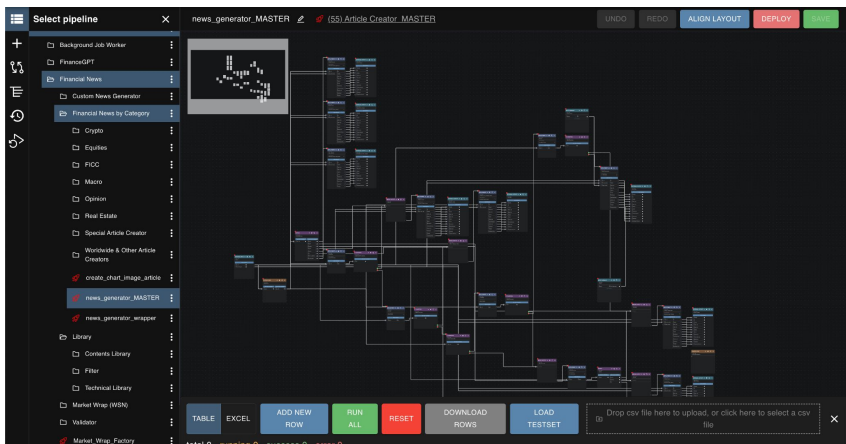
Google

Bloomberg



다양한 소스

2. RAG In Production > Super News



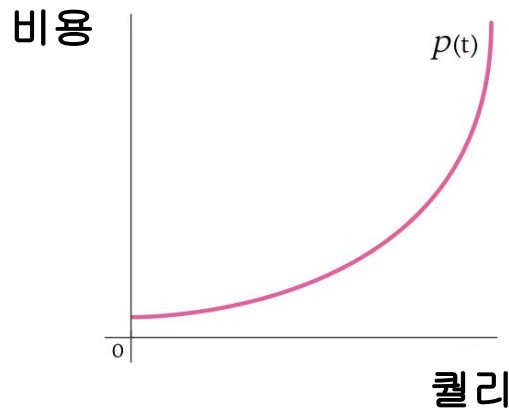
Internal LLMOps Platform

하나의 기사를 생성하기 까지

- 수십 개의 데이터 크롤링 -> 전처리 -> 임베딩 & Ingestion
- Multiple **LLM API Call**
- Multiple **Embedding Vector Lookup** (hybrid)
- External Apis
 - Google Search
 - Yahoo Finance
 - ...

2. RAG In Production > 운영하면서 느낀점

- 높은 퀄리티의 생성물을 얻기 위해서 정말 **비용**(사람, LLM, 인프라 구성 등)이 많이 듦
- 견고하게 운영하기 위해선 **소프트웨어 엔지니어링**에 대한 전체적인 이해도가 많이 중요함
- Domain Expertise가 필요한 서비스의 경우 **Communication Cost 절감**이 생각보다 중요함
- 퍼포먼스는 Agent < LLM Chaining (Rule Base + Prompt Engineering이 녹아진)
- Prompt Engineering이 중요하지만 이렇게 힘든지 몰랐지



2. RAG In Production > 문제

- 기사 생성 과정 혹은 결과물에서 **LLM Hallucination**이 발생하면?
- 잘못된 소스들을 바탕으로 기사가 생성되면? (**Poor data retrieval**)
- 생성물의 **Quality**를 제대로 정량화하고 싶다면?
- 기사 생성 중간에 **Technical Issues** 들이 발생하면?

3. Reducing Hallucinations > Workaround

Technical

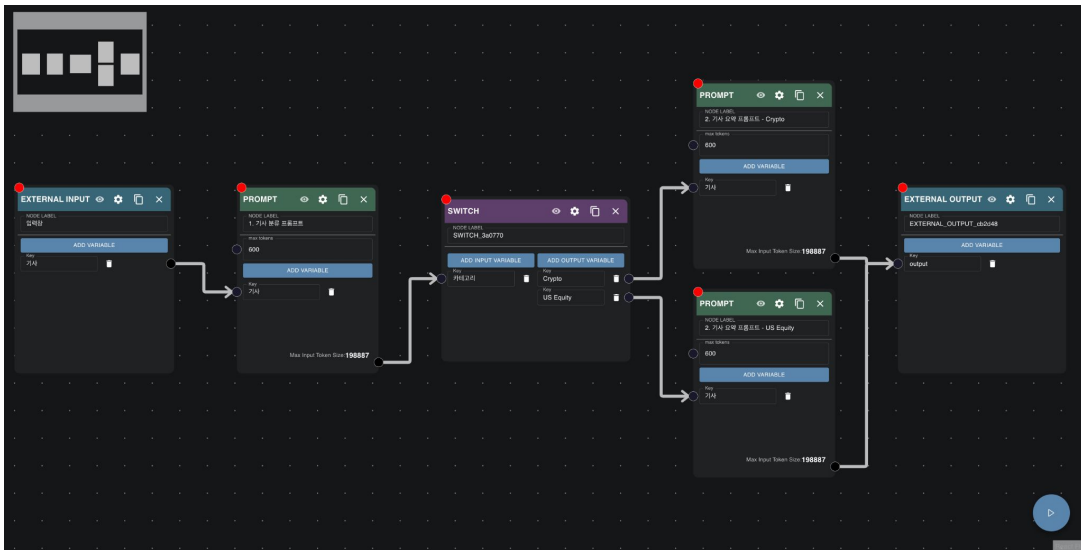
- LLM Chaining
- Retrieval 개선하기
- GuardRail 적용 (Online Evaluation)

Non-technical

- Prompt Engineering
- Evaluation Process 정립

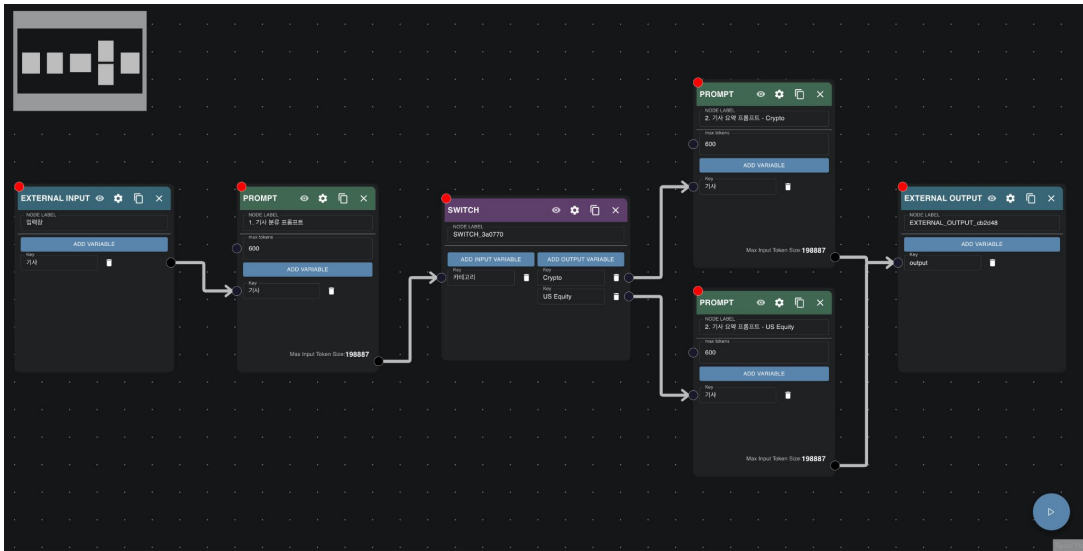


3. Reducing Hallucinations > LLM Chaining



- Chaining = Prompt 쪼개고 이어 붙이기
- 복잡한 명령을 수행할수록 Chaining은 필수
- 장점
 - 프롬프트가 명확해지고 작성이 쉬워짐
 - 재사용이 가능함
- 단점
 - 소프트웨어 복잡도가 높아짐
 - 비용이 더 나감

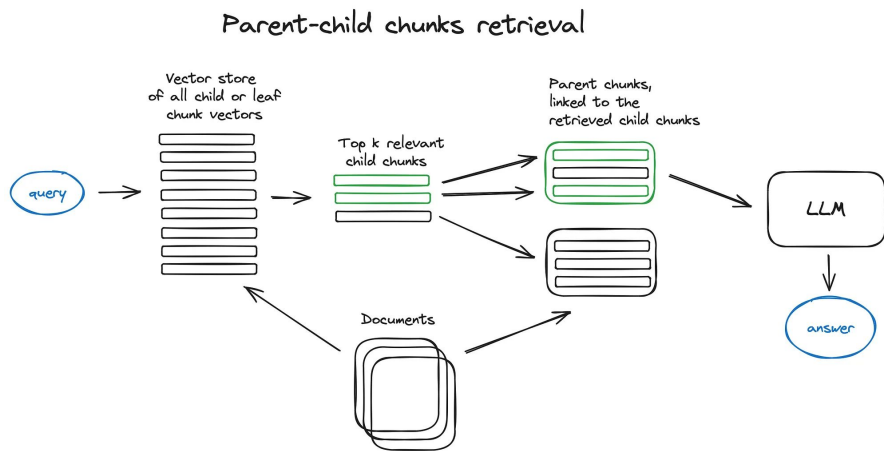
3. Reducing Hallucinations > LLM Chaining



E.g., 테슬라 사야 돼 말아야 돼?

1. Bullish, Bearish View별 요청
2. 개별 결과들 바탕으로 Embedding lookup 진행
3. 위 Prompt 결과들 + 검색 결과 바탕으로 최종 요청

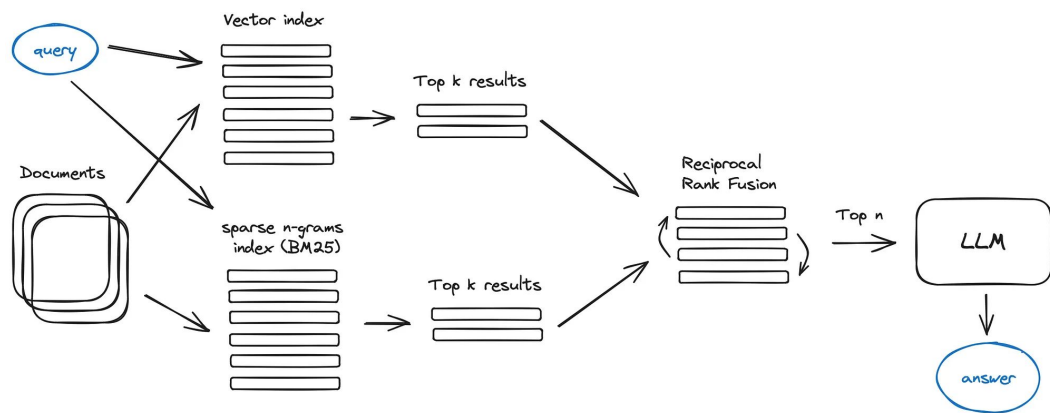
3. Reducing Hallucinations > Retrieval > PreProcessing



- **Embedding**할 텍스트는 명확한 문맥이 있을수록
좋다
- Embedding PreProcessing 개인적인 팁
 - 길이 길게 + 적게 < 길이 짧게 + 많이
Chunking하기 (후처리가 용이함)
 - 텍스트에 **Summary**나 **Tag** 등의 부가
정보를 붙여서 임베딩하는 것도 괜찮음
 - **Hierarchy**를 깊게 두는게 그만한 효용이
있는지는 모르겠음
- **LlamaIndex** 가 Retrieval Data Chunking
관련해서는 가장 많이 지원해주는 듯

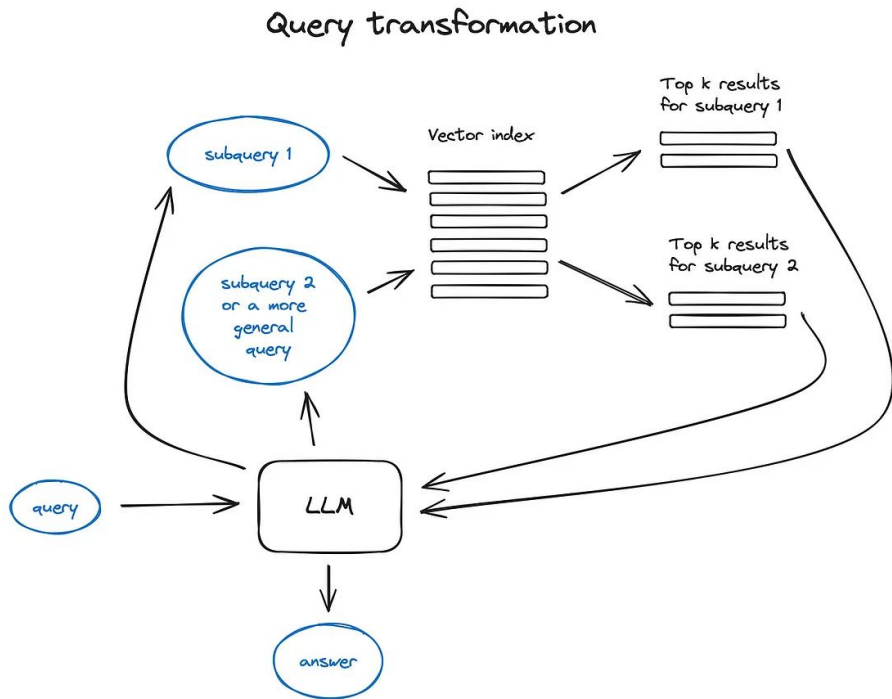
3. Reducing Hallucinations > Retrieval > Hybrid Search

Fusion retrieval / hybrid search



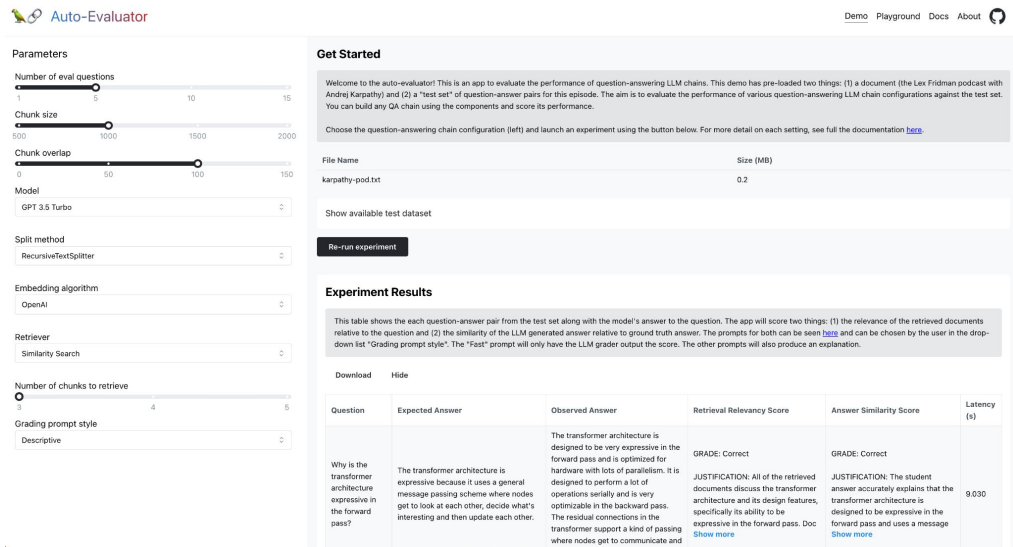
- Hybrid Search: **Keyword & Semantic** 검색을 결합
 - Keyword 검색(BM25, SPLADE, etc)
 - Semantic (Embedding method)
- 일반적 RAG는 **Keyword < Semantic < Hybrid**
- Elastic Search, Pinecone, Weaviate 등이 지원

3. Reducing Hallucinations > Retrieval > Query Transformation



- 대부분의 Q&A 봇에 QT는 적용하면 좋다고 생각함
- 대표적으로 Hyde, Decompose, Step-back prompting 등이 있음
- **Decompose + Step-back prompting**을 결합하는 방식이 제일 괜찮다고 생각함

3. Reducing Hallucinations > Retrieval > Experiment



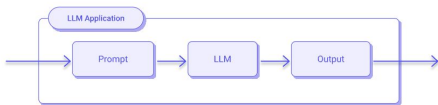
The screenshot displays the Auto-Evaluator web application. On the left, the 'Parameters' section includes sliders for 'Number of eval questions' (set to 5), 'Chunk size' (set to 1000), and 'Chunk overlap' (set to 100). It also features dropdown menus for 'Model' (GPT 3.5 Turbo), 'Split method' (RecursiveTextSplitter), 'Embedding algorithm' (OpenAI), 'Retriever' (Similarity Search), 'Number of chunks to retrieve' (set to 3), and 'Grading prompt style' (Descriptive). The main area is titled 'Get Started' and contains a welcome message, a 'Re-run experiment' button, and an 'Experiment Results' section. The results section includes a table with columns: Question, Expected Answer, Observed Answer, Retrieval Relevancy Score, Answer Similarity Score, and Latency (s).

Question	Expected Answer	Observed Answer	Retrieval Relevancy Score	Answer Similarity Score	Latency (s)
Why is the transformer architecture expressive in the forward pass?	The transformer architecture is expressive because it uses a general message passing scheme where nodes get to look at each other, decide what's interesting and then update each other.	The transformer architecture is designed to be very expressive in the forward pass and is optimized for hardware with lots of parallelism. It is designed to perform a lot of operations serially and is very optimizable in the backward pass. The residual connections in the transformer support a kind of passing where nodes get to communicate and	GRADE: Correct JUSTIFICATION: All of the retrieved documents discuss the transformer architecture and its design features, specifically its ability to be expressive in the forward pass. Doc Show more	GRADE: Correct JUSTIFICATION: The student answer accurately explains that the transformer architecture is designed to be expressive in the forward pass and uses a message Show more	9.030

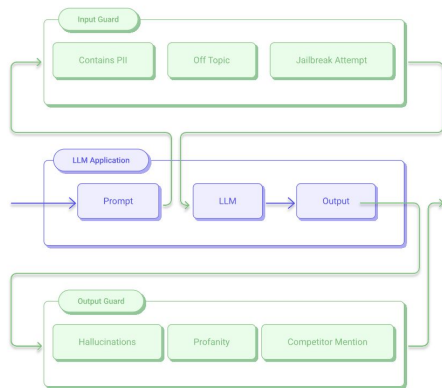
- 사실 Silver Bullet은 없어서 계속 실험이 필요
- Embedding Evaluation 작업도 LLM Evaluation 만큼이나 가치가 있다고 생각함
- score 뿐만 아니라 다양한 평가 지표 필요 (LLM, Answer Sheet 등)

3. Reducing Hallucinations > GuardRail

Without Guardrails



With Guardrails



생성에 필요한 Input을 받는 단계나 생성 결과물에 대한 배포 여부를 결정하는 단계에 GuardRail 심기

E.g.,

- 특정 Regulations Violation 여부 확인 (formatting 같은)
- 'I'm sorry' 같은 특정 hallucination에 많이 발생하는 text기반 equal check하기
- Fact Check LLM 돌리기(소스 데이터를 Ground Truth로)
- ...

3. Reducing Hallucinations > Prompt Engineering

- MODEL * AI Settings * Template 조합으로 계속 돌려보기
- 항상 간결하게 작성하려고 하는 게 제일 중요함
- 말을 잘 듣는 선까지 Instruction을 키워보기 -> 잘 안되면 쪼개기 (Chaining)
- 사내에서 템플릿을 만들어두고 관리하는 중
- 거인 형님들(Model)의 어깨에 잘 올라탈 수 있도록 유연하게 구조 가져가기
 - Model API는 적극적으로 추가 후 실험
 - Gemini 1.5는 1백만 토큰까지 입력이 가능하다고?
 - > 기존 PDF Reader RAG는 빠르게 손절 후 테스트

3. Reducing Hallucinations > Prompt Engineering

인사이트

- SNS에 공유되는 핫한 프롬프트 테크닉을 수십개 돌려서 **Evaluation** 해보면 그닥?
- 어느정도 퍼포먼스가 잘 나오는 시점부터는 프롬프트 자체를 수정하는 것보다 **data retrieval**에 더 신경쓰는 게 비용이 절감되는 효과가 있음
- 항상 최고 사양 모델이 좋은 건 아님
 - **GPT 4**는 더 똑똑하지만, 답변이 느리고 비쌈
 - 복잡한 추론이 필요하지 않다면 **Claude 3 Haiku**나 **GPT 3.5**로 대체 가능(비용, 속도 측면에서 이점)
- **Model API**들도 내부적으로 버저닝을 하고 비공개 업데이트도 여러번 진행하는 듯. 종종 안되던 게 잘되고, 잘되던게 안됨..
- 모든 팀원들의 **Prompt Engineering**의 수준을 어느정도 가지는 것도 중요한듯
 - 엔지니어도 가벼운 프롬프트는 직접 만듦

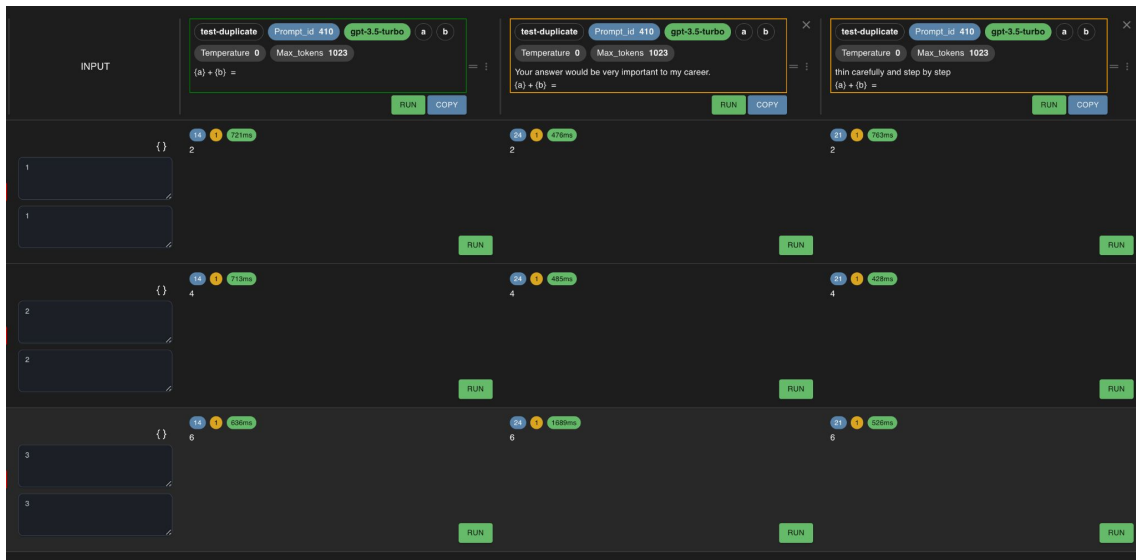
3. Reducing Hallucinations > Evaluation Process

- Prompt Engineering의 꽃은 Evaluation!
- Production 환경에서 LLM 서비스 운영하기 위해 절대적으로 중요함
- 실제로 우리 팀도 Prompt를 배포하기 까지 Evaluation(비교, 채점)하는 시간이 가장 많은 비중 차지
- HoneyHive, Deep Checks 등 LLM Evaluation에 초점을 둔 SaaS 서비스들이 요새 꽤 보임

3. Reducing Hallucinations > Evaluation Process

1. 자동화할 수 있는 것과 아닌 것을 분리하기
2. 주요 메트릭 선정하기
 - Hallucination
 - Answer Relevancy
 - Source Data Relevancy
3. **Evaluation** 방식 선정하기
 - Code (Regex, ...) - format같은 hard check에서 필요
 - LLM - 가장 많이 사용함
 - Embedding Similarity - 크게 유효한지는 모르겠음
 - ...
4. **Evaluation Sheet**(정답지) 구성하기

3. Reducing Hallucinations > Evaluation Process



Prompt 비교/채점 실행 테이블 (Internal LLMOps)

- Prompt Engineering에 생각보다 많은 시간을 소요함
 - Prompt 작성 -> 결과 확인 -> 이전 결과와 비교 & 채점
 - Spread sheet같은 table view를 많이 활용하게 됨
- 프롬프트 수정 후 쉽게 결과를 비교하고, 채점하는 일련의 과정을 지원하는 것은 중요

3. 마무리 홍보 - 1



글로벌 LLM Powered News Platform을 만드는 **Project Pluto**에 조인하세요!

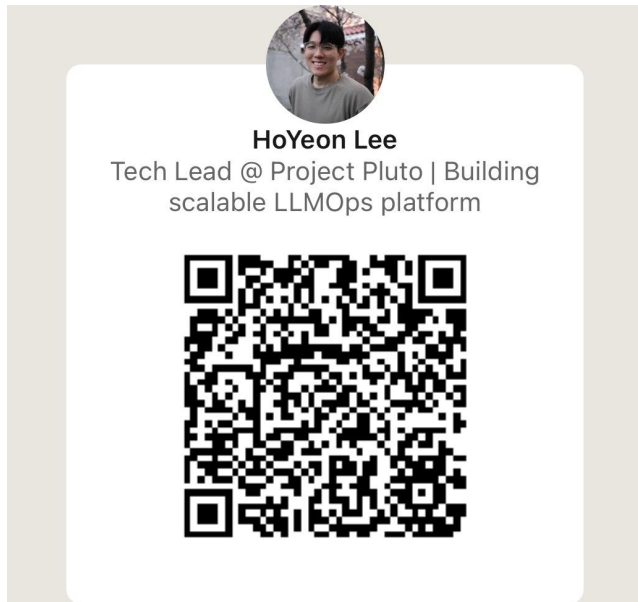
Software Engineer, ML Engineer 애타게 찾습니다

소수의 인원으로 복잡한 RAG, LLM Ops, Model Tuning 등 이것저것 다 하고 있음!

런칭 한 달 만에 **MAU 10만** 달성!

구글 Tech Lead 출신 **CTO**와 뉴욕 주민 **CEO** 그리고 그랩과 함께!

3. 마무리 홍보 - 2



시간 관계상 다 하지 못한 이야기들을 조만간 **개인 공유회**로 **자세하게** 다룰 예정입니다.
위 QR코드로 **링크드인** 추가를 해두면 소식을 바로 팔로업할 수 있습니다.

감사합니다