

"ChatGPT와 Vibe Coding의 등장 이후, 개발자 생태계는 어떻게 변했는가?"

GitHub & Stack Overflow 데이터 기반 실증 분석



목차

- 목차1 프로젝트 개요
- 목차2 GitHub 데이터 수집
- 목차3 Stackoverflow 데이터 수집 & 데이터 전처리
- 목차4 데이터 분석 & 데이터 시각화
- 목차5 분석 결과 해석



Definitions Summary Synonyms Sen ▶

Definition of 'vibe coding'



vibe coding

in British English

(ˈvaɪb ˌkəʊdɪŋ ⓘ)

noun slang

the use of artificial intelligence prompted by natural language to assist with the writing of

computer code

‘Vibe Coding’,
2025 올해의 단어
선정

♡ 688 💬 1 📌



빅데이터활용사례를통한이해_윤태웅

프로젝트 개요 - 주제 선정

1. ChatGPT 출시(2022.11.30)
2. Vibe Coding 시대의 도래
3. 1억 8000만명의 새로운 개발자?

주제 선정 이유 : 2022년 11월 ChatGPT 출시 이후 개발 시장이 급격히 변화했습니다.

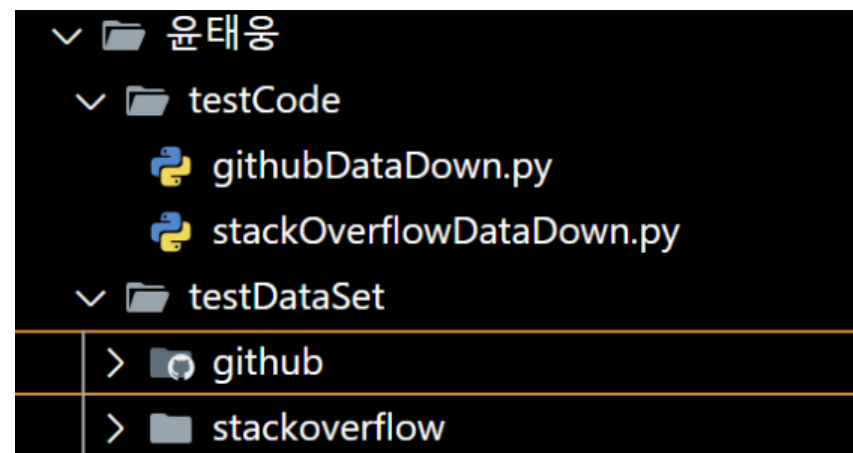
본 프로젝트는 AI 코드 생성 도구가 개발자 생태계에 미친 영향을 github와 stackoverflow 데이터를 통해 분석합니다.



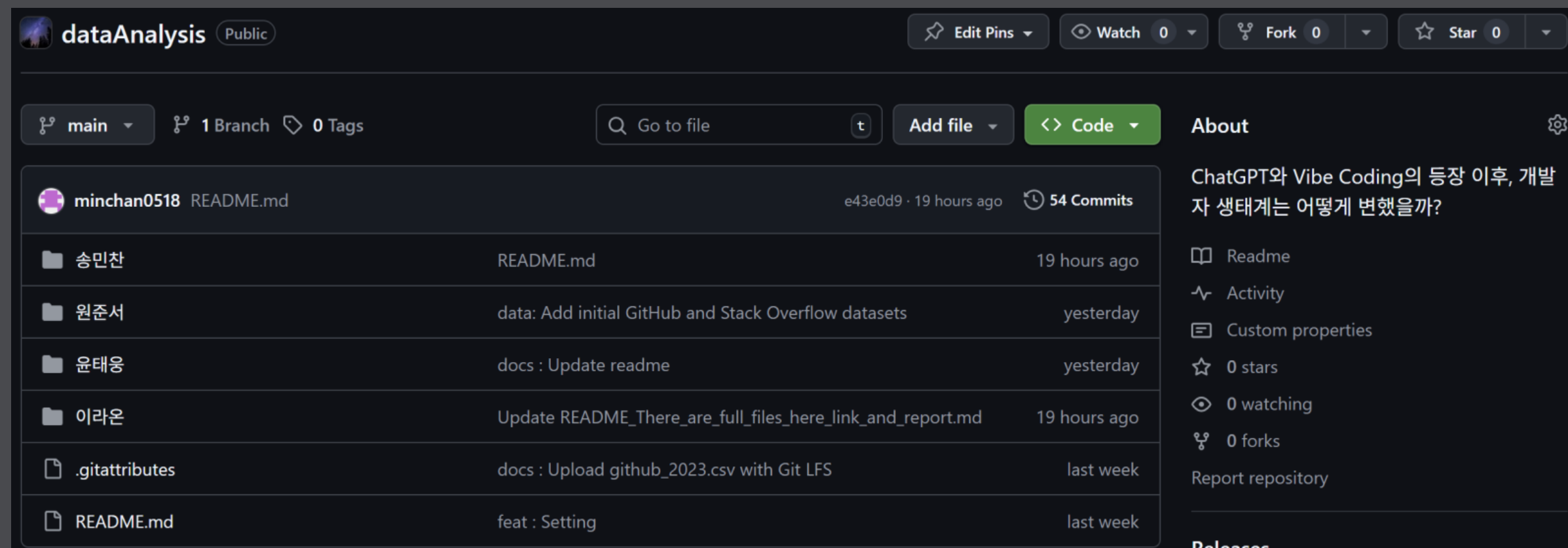
역할 분배

1. PM/분석 결과 해석: 윤태웅
2. GitHub 데이터 수집: 송민찬
3. Stack Overflow 데이터 수집/전처리: 이라온
4. 데이터 분석/시각화: 원준서

데이터 수집 테스트 코드 작성 및 검증



GitHub 조직 기반 협업 환경 구축





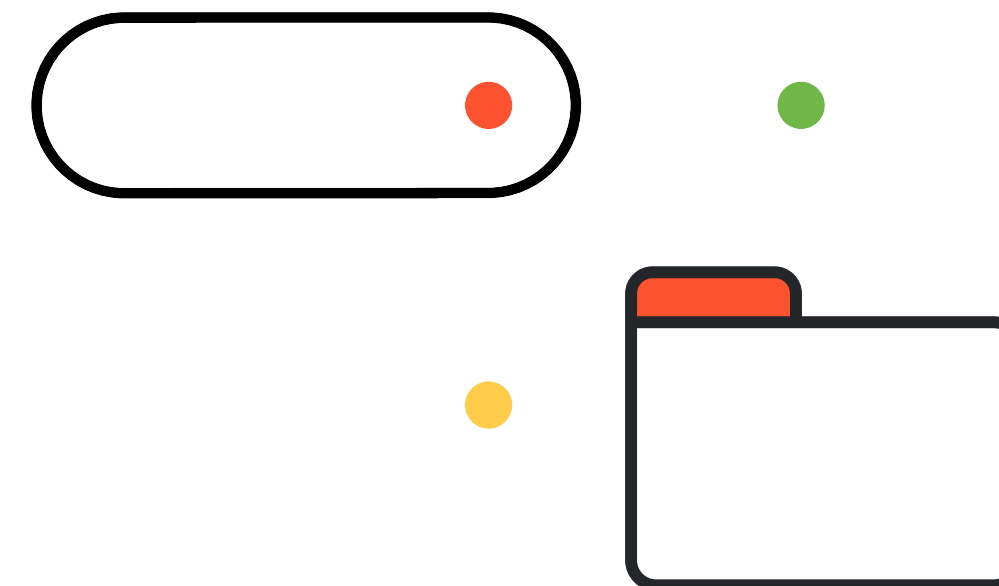
연구 가설

RQ1 : ChatGPT 출시 전후(2021-2022 vs 2023-2025)
GitHub AI 프로젝트는 어떻게 변화했는가?

RQ2 : Github 레포지토리가 2023년 이후 실제로 증가했는가?

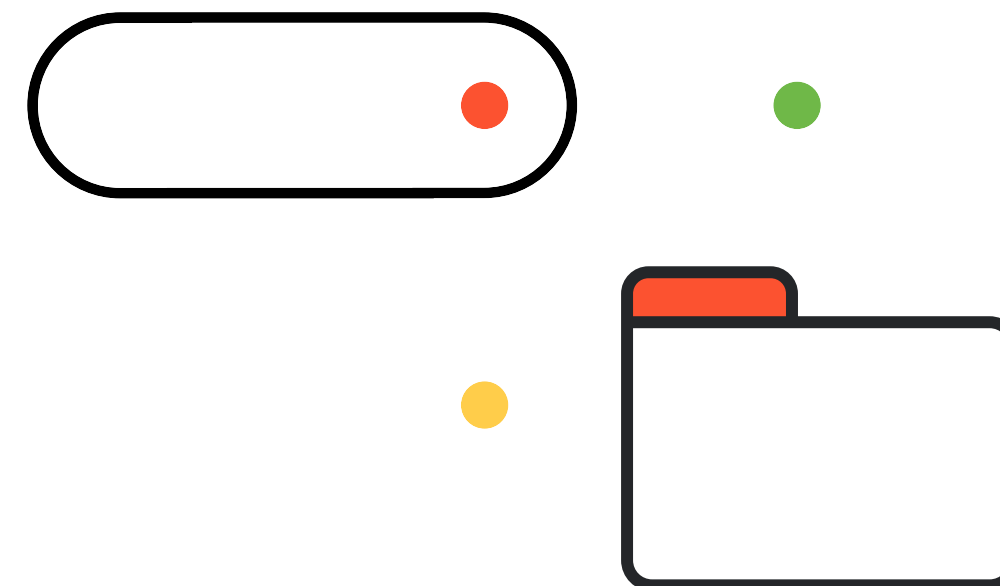
RQ3 : AI 분야별(LLM, 강화학습, 컴퓨터비전, NLP) 프로젝트 증
가 양상이 다른가?

RQ4 : Stack Overflow 질문 수와 GitHub 프로젝트 수 간 상관관
계가 있는가?





Github 데이터 수집





GitHub 데이터 수집

수집 대상: 12개 AI 키워드,
2021-2025년 레포지토리

✓ 송민찬

- github_2021_01.csv
- github_2021.csv
- github_2022.csv
- github_2023.csv
- github_2024.csv
- github_2025.csv
- GitHub2021~2025.ipynb
- README.md



키워드: 총 12개 AI 관련 키워드

- LLM 관련 : chatgpt, gpt-4, openai-api, llm, langchain
- AI Agent : AI-agent, autonomous-agent, github-copilot
- 전통 AI : machine-learning, artificial-intelligence, deep-learning, neural-network

수집 시간

- 2021-2022년: 10시간
- 2023-2025년: 6시간
- 총 소요 시간: 16시간

데이터를 수집하며 생긴 문제와 해결방법

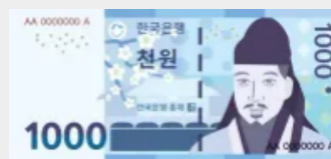
1

 실행 중(9시간 34분 15초)

**문제: 인증된 토큰당
요청 제한**

1. 3개의 GitHub 계정 생성
및 토큰 발급
2. 병렬 수집 파이프라인 구축
3. 결과: 데이터 수집 속도 향상

2



**단일 검색 쿼리당 최대
1000개 레포지토리만 반환**

1. 전체 기간을 일(Day) 단위로 세밀하게 분할
2. 각 날짜별 독립 쿼리 실행
3. 결과: 제한 없이 전체 데이터 수집 성공

3

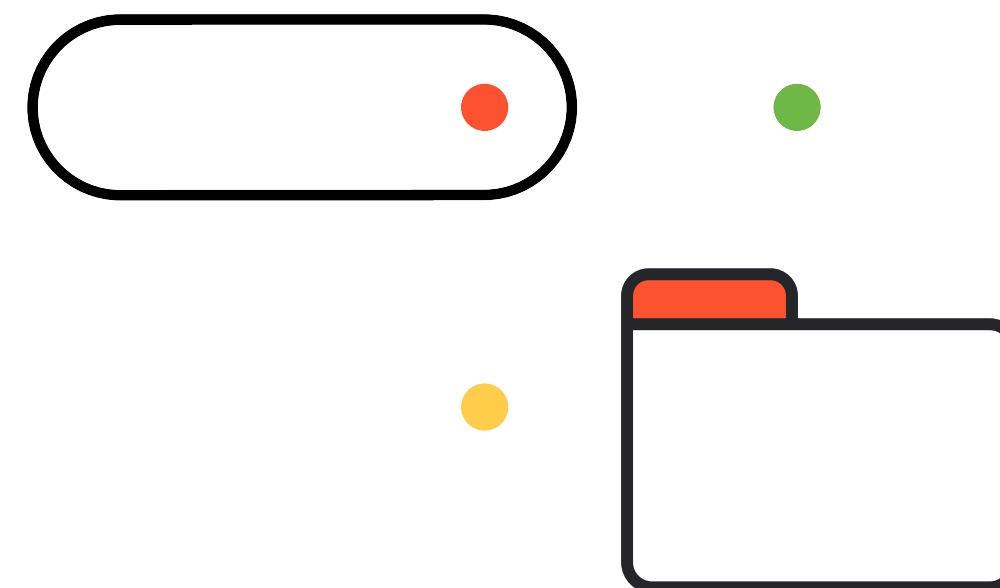
5

하나의 쿼리에서 논리 연산자(OR) 최대 5개 제한

1. 12개 키워드를 5개씩 청크로 분할
2. 다중 쿼리로 모든 키워드 커버
3. 결과: 12개 키워드 검색 결과 완전 반영



Stackoverflow 데이터 수집 & 데이터 전처리



데이터 수집

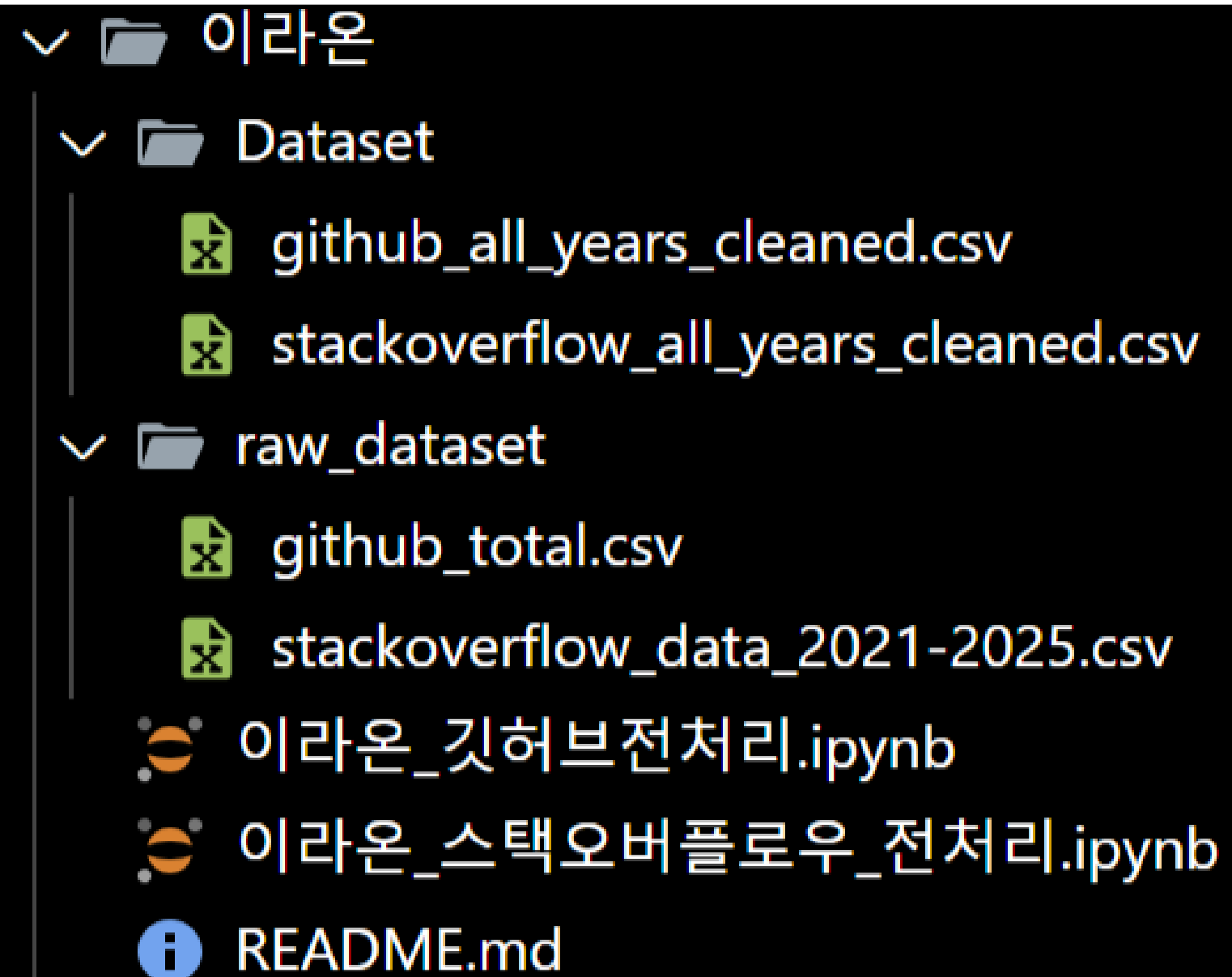
Stackoverflow 데이터 수집

BigQuery 기반 효율적 수집

직접 연동: Google Colab에서 BigQuery API를 통해
공개 데이터셋 접근
서버 사이드 필터링: SQL WHERE 절로 2021-2025년
데이터만 선별 추출

수집 항목: id, title, tags, creation_date

효율적 처리: 전체 테이블 다운로드 없이 필요 데이터만
Pandas DataFrame 변환





GitHub 데이터 병합

연도별 데이터 통합

- 데이터 통합: 5개 연도별 CSV 파일을 pandas.concat으로 단일 DataFrame 병합
- 품질 검증:
 - df.info(): 데이터 타입 일관성 확인
 - df.isnull().sum(): 결측치 분포 검증



GitHub & Stack Overflow 전처리 과정

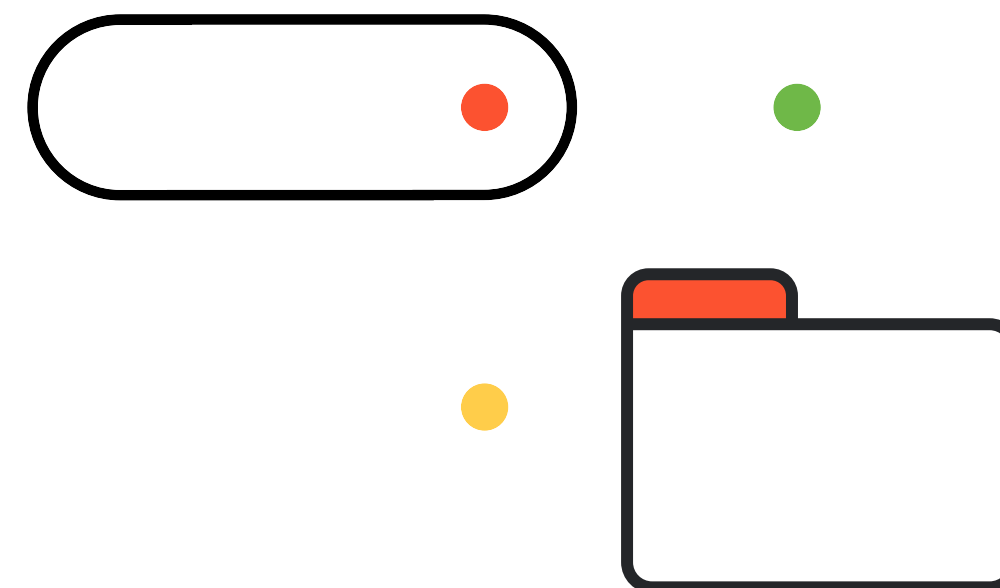
- 중복 제거: GitHub(repo_id), Stack Overflow(id) 기준 drop_duplicates 실행
- 결측치 처리: 핵심 열(ID, 날짜, 태그) 제거, 텍스트 열은 빈 문자열 대체
- 타입 표준화: creation_date UTC 통일, tags 소문자 변환

파생 변수 생성

변수명	생성 기준	목적
period	ChatGPT 출시일(2022-11-30) 기준	Before/After 시기 구분
AI_field	tags 키워드 매칭	AI 분야 대분류 (LLM, Computer Vision, NLP, General ML 등)
키워드 플래그	특정 키워드 존재 여부	'ai-agent', 'vibe-coding' 등 Boolean 값 생성

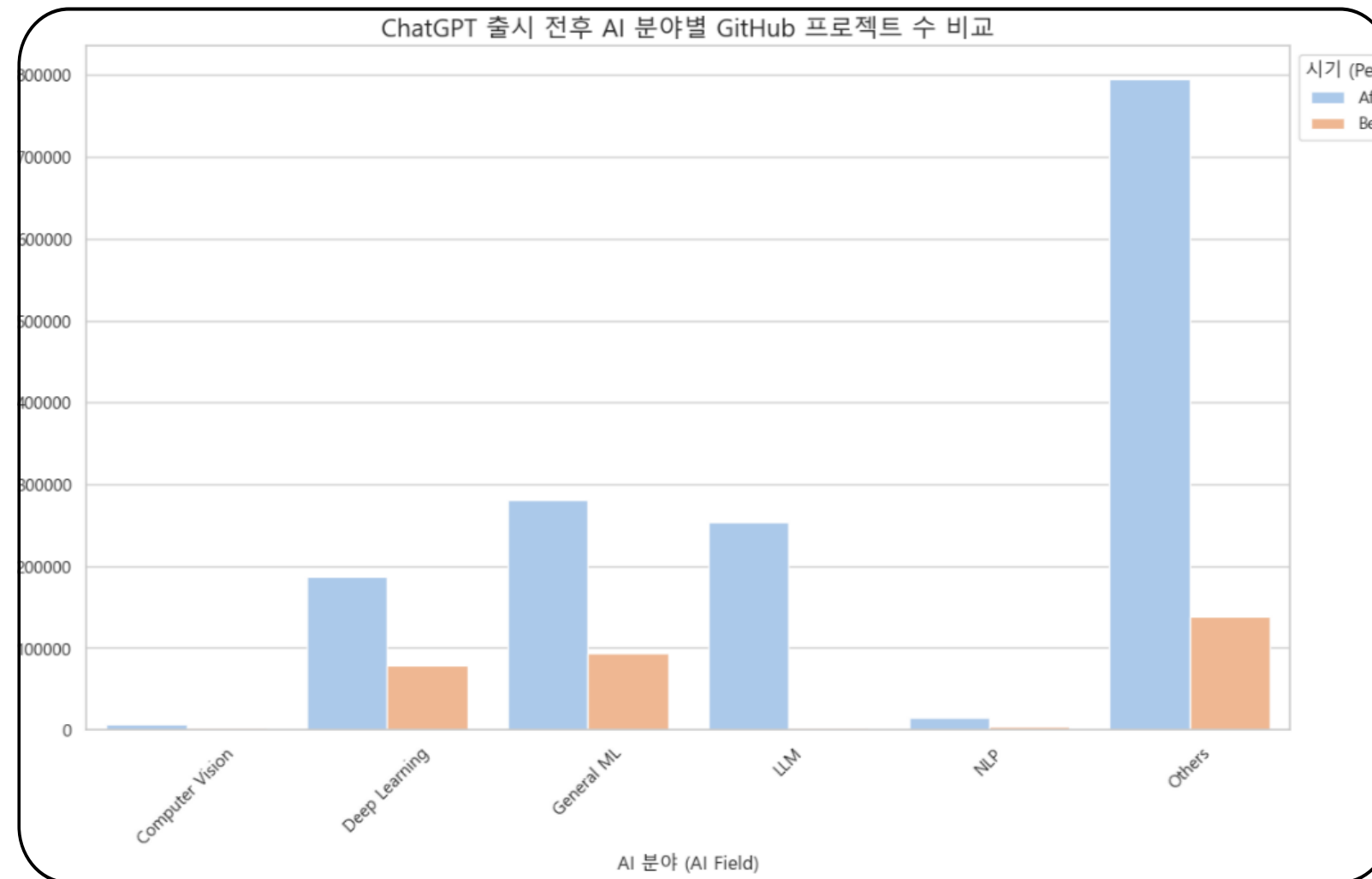


데이터 분석 & 데이터 시각화



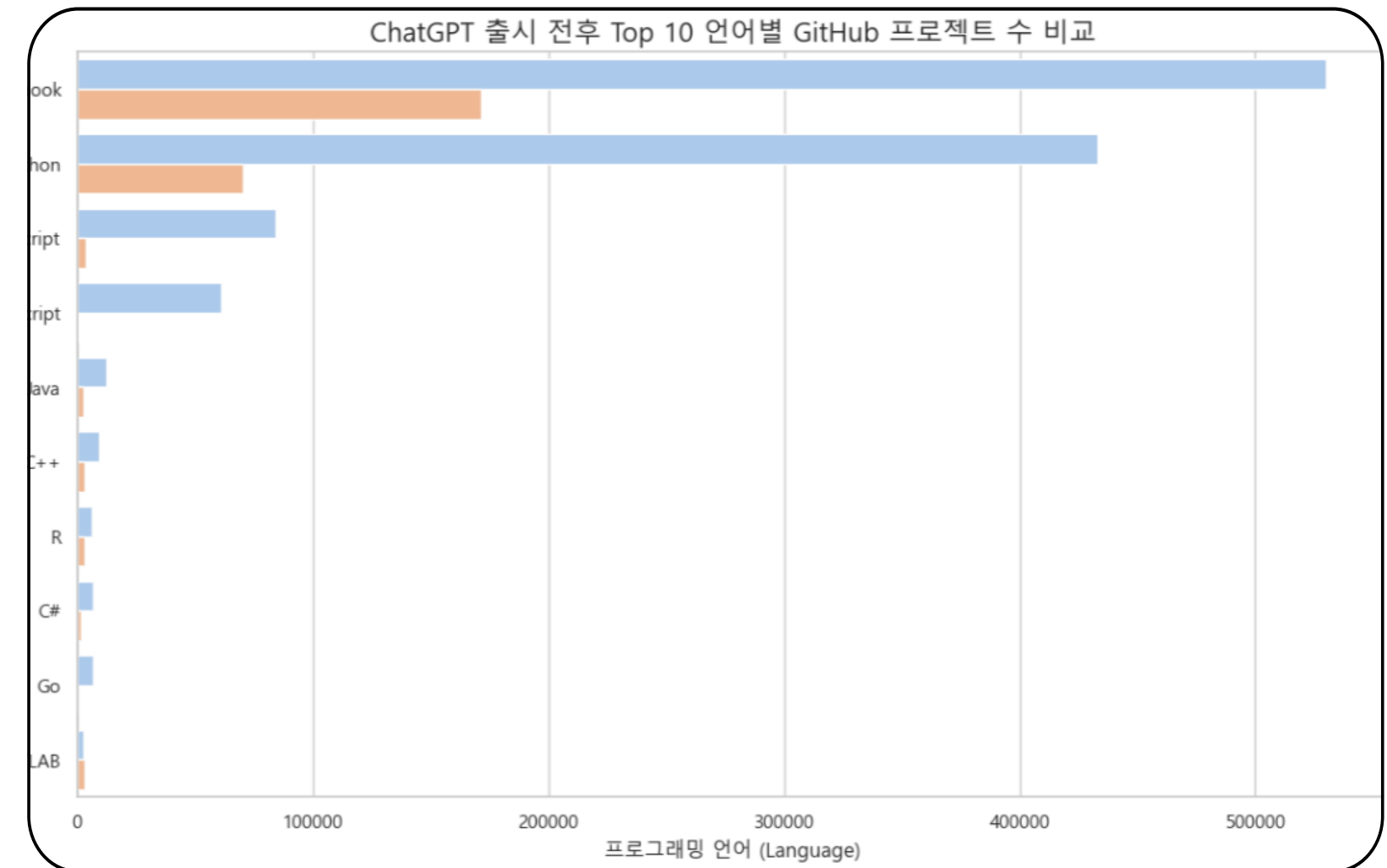
데이터 분석 & 시각화 - 1. ChatGPT 출시 전후(2021-2022 vs 2023-2025) GitHub AI 프로젝트는 어떻게 변화했는가?

막대 그래프를 활용한 비교 분석



모든 AI 분야에서 After(2023-2025)의 프로젝트 수가 Before(2021-2022)에 비해 압도적으로 증가했습니다.

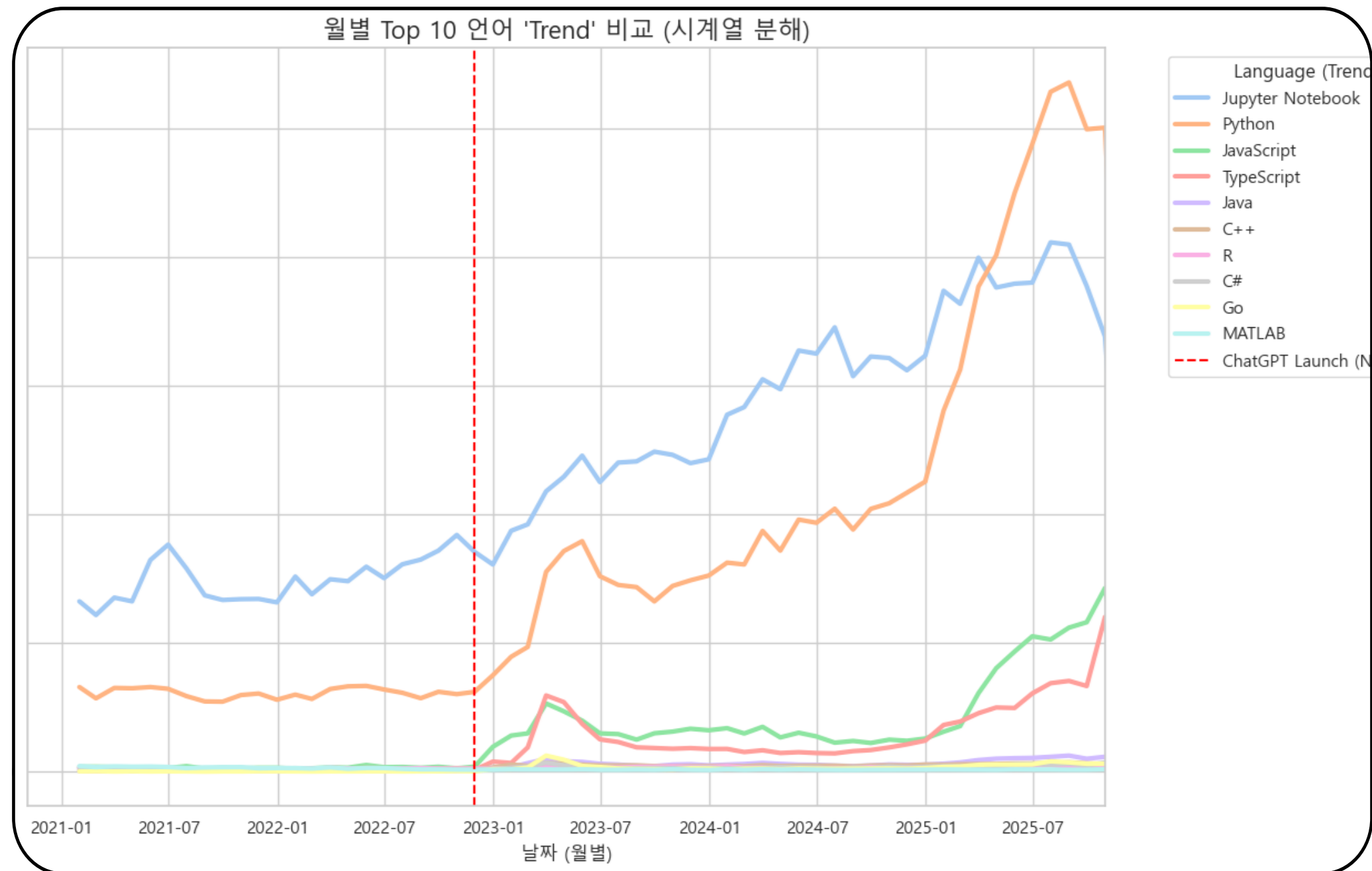
막대 그래프를 활용한 비교 분석



Jupyter Notebook과 Python이 압도적인 1, 2위를 차지했습니다.

데이터 분석 & 시각화 - 2. Github 레포지토리가 2023년 이후 실제로 증가했는가?

시계열 차트를 활용한 추세 및 이벤트 영향 분석

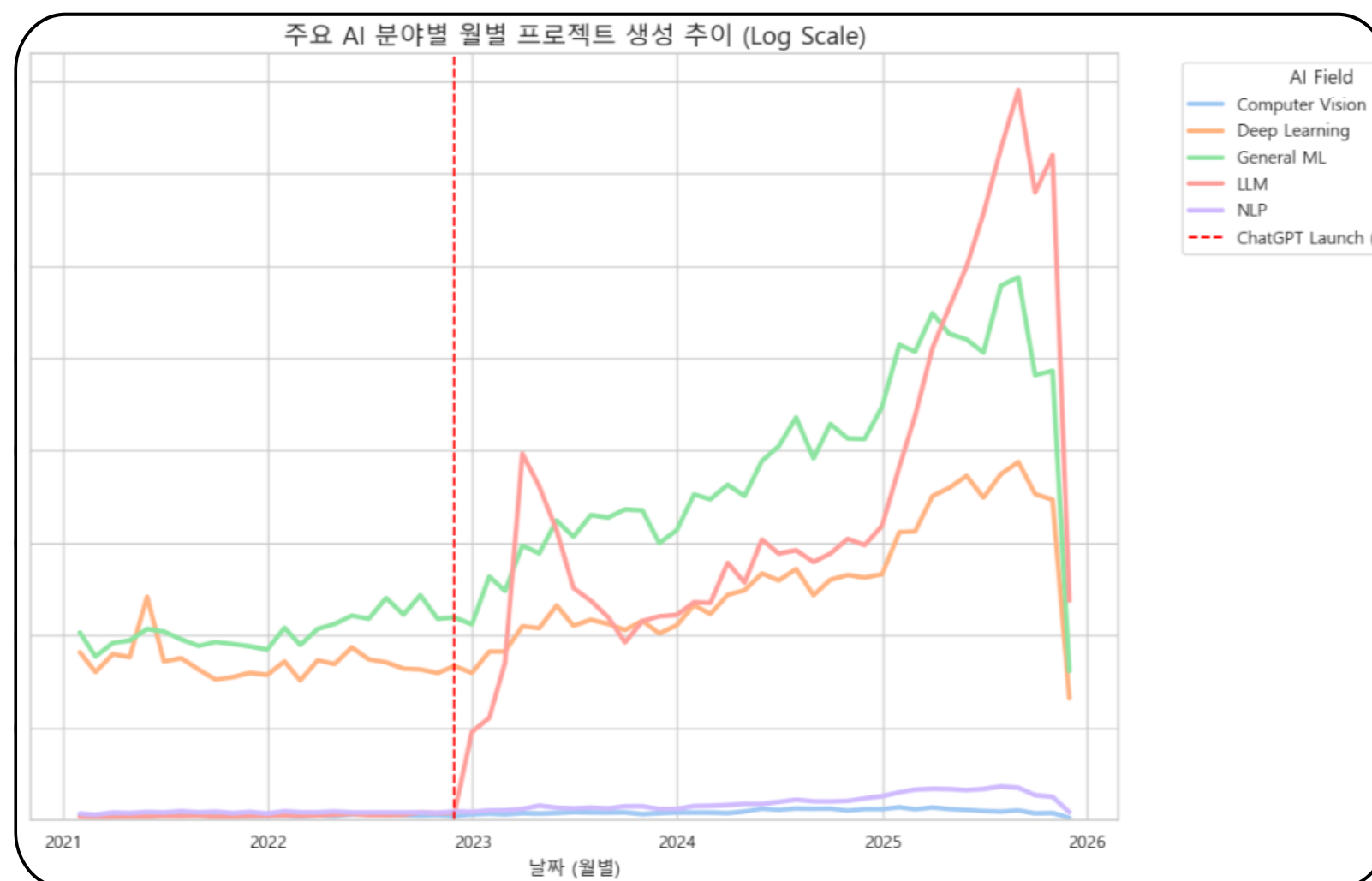


2025년 갑작스럽게 Python이 Jupyter Notebook을 뛰어넘는 모습을 볼 수 있었습니다.



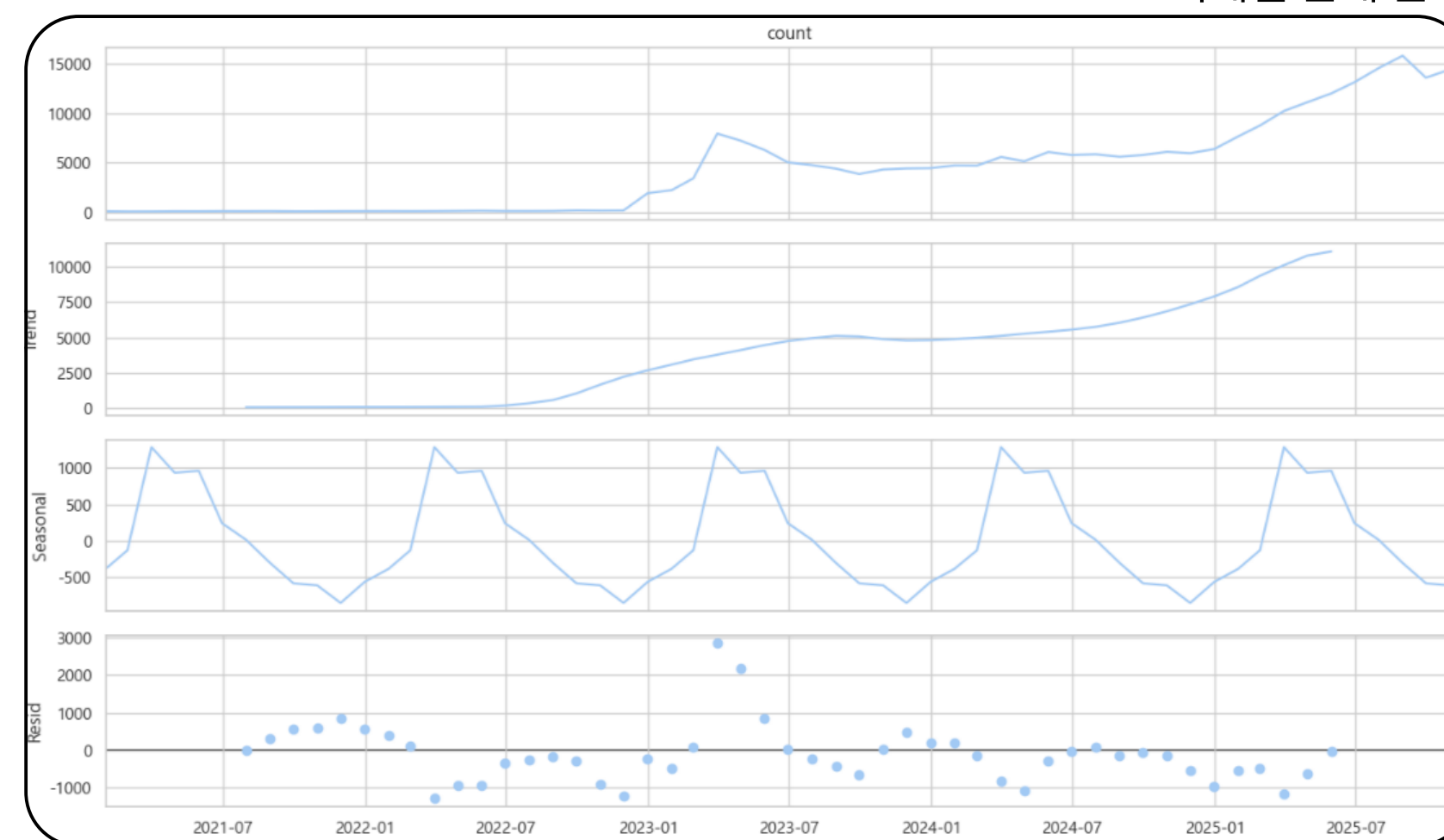
데이터 분석 & 시각화 - 3-1. AI 분야별(LLM, 강화학습, 컴퓨터비전, NLP) 프로젝트 증가 양상이 다른가?

로그 스케일 시계열 차트 및 이벤트 영향 분석



LLM은 ChatGPT 출시 이전에는 거의 바닥에 있었으나 2023년 초를 기점으로 다른 모든 분야를 압도하며 수직으로 올라가는 모습을 보여줍니다.

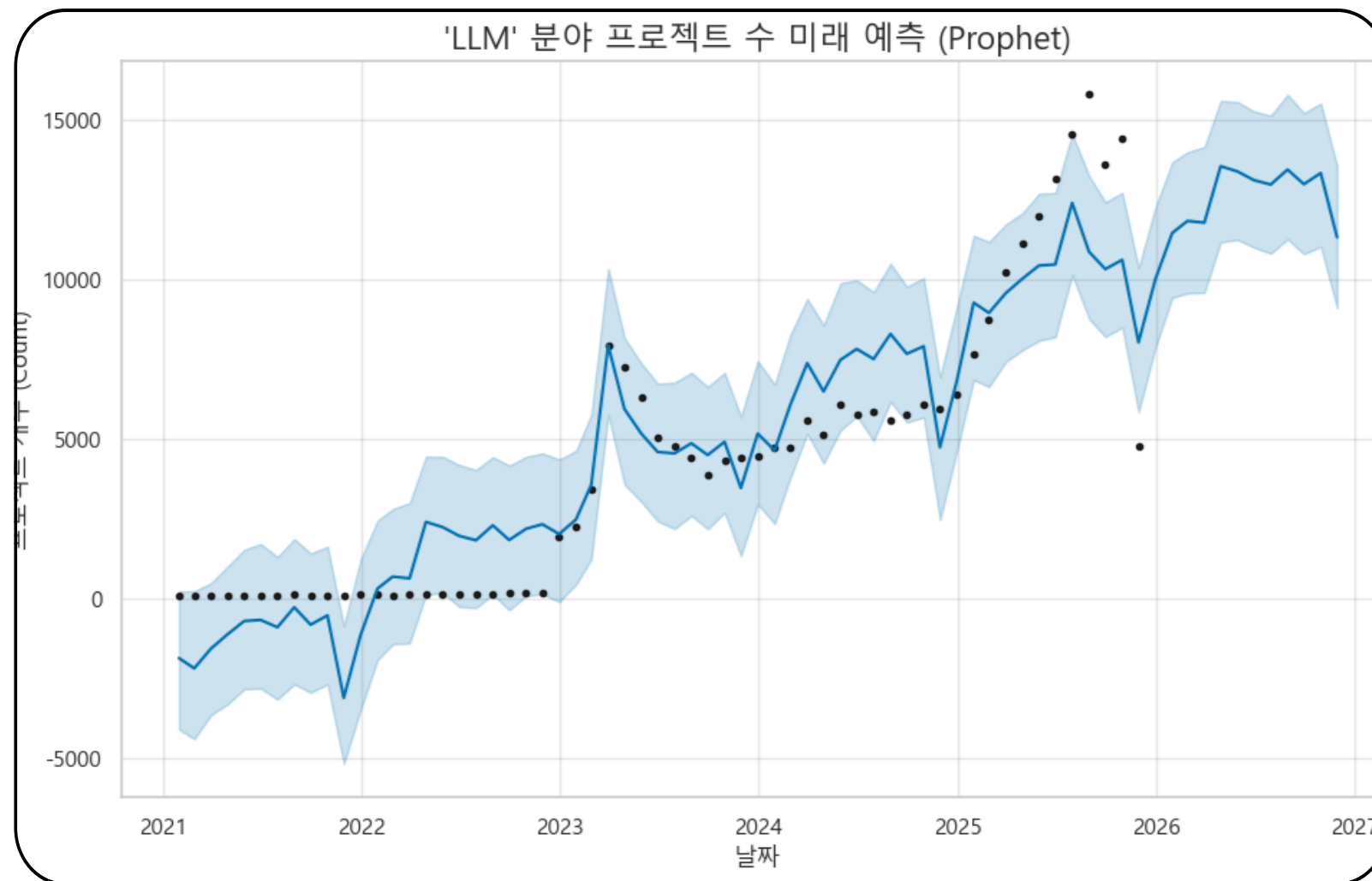
시계열 분해 분석



2022년 말부터 시작된 폭발적인 상승 추세와 함께 뚜렷한 계절적 주기가 반복되는 것이 특징입니다.

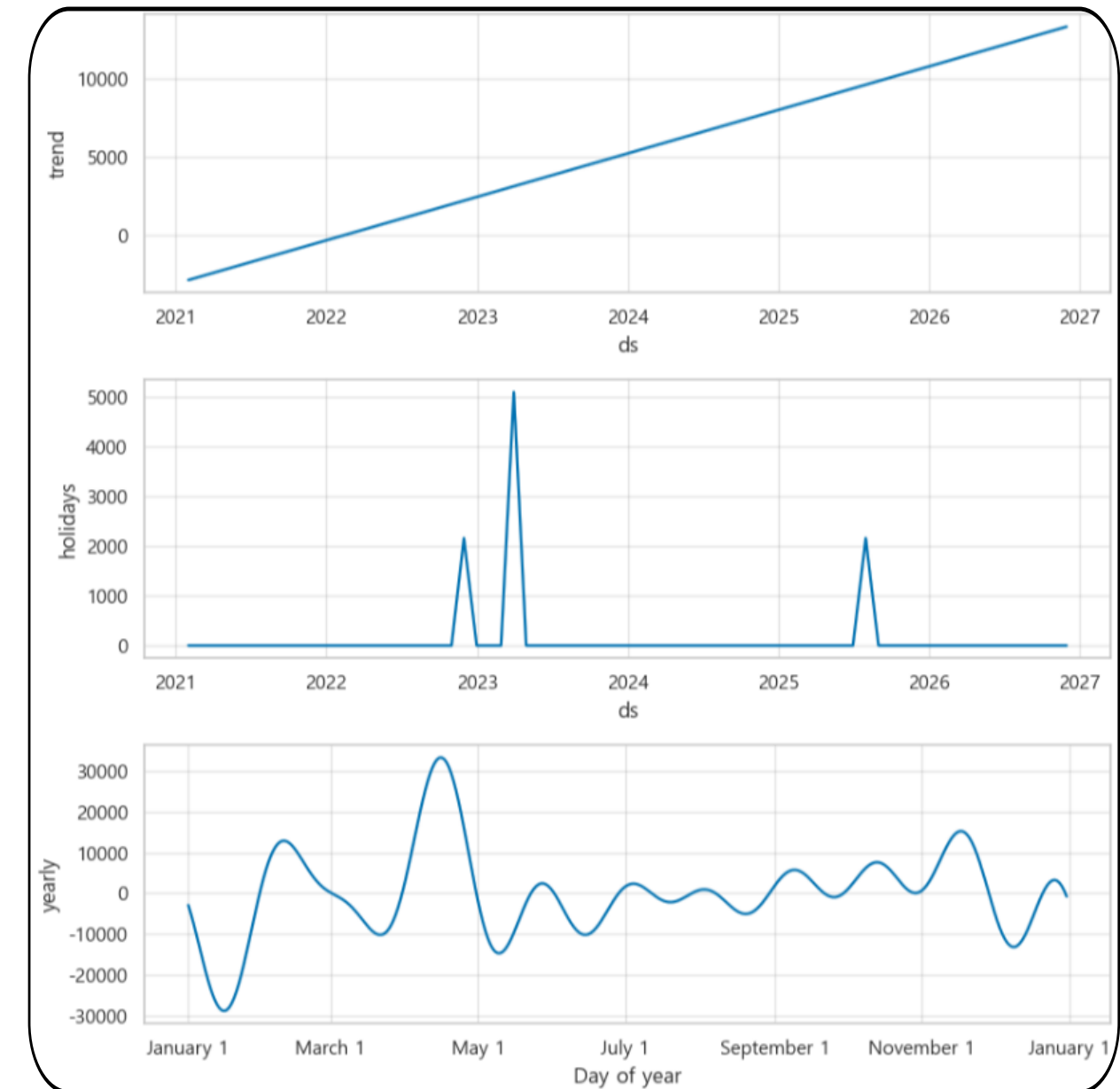
데이터 분석 & 시각화 - 3-2. AI 분야별(LLM, 강화학습, 컴퓨터비전, NLP) 프로젝트 증가 양상이 다른가?

Prophet 모델을 이용한 시계열 미래 예측



2023년 초 폭발적인 성장 이후 꾸준한 우상향 추세가 이어져, 2026년 이후에는 프로젝트 수가 최대 1만 3천 건 이상에 도달할 것으로 예측됩니다.

Prophet 모델 구성 요소 시각화

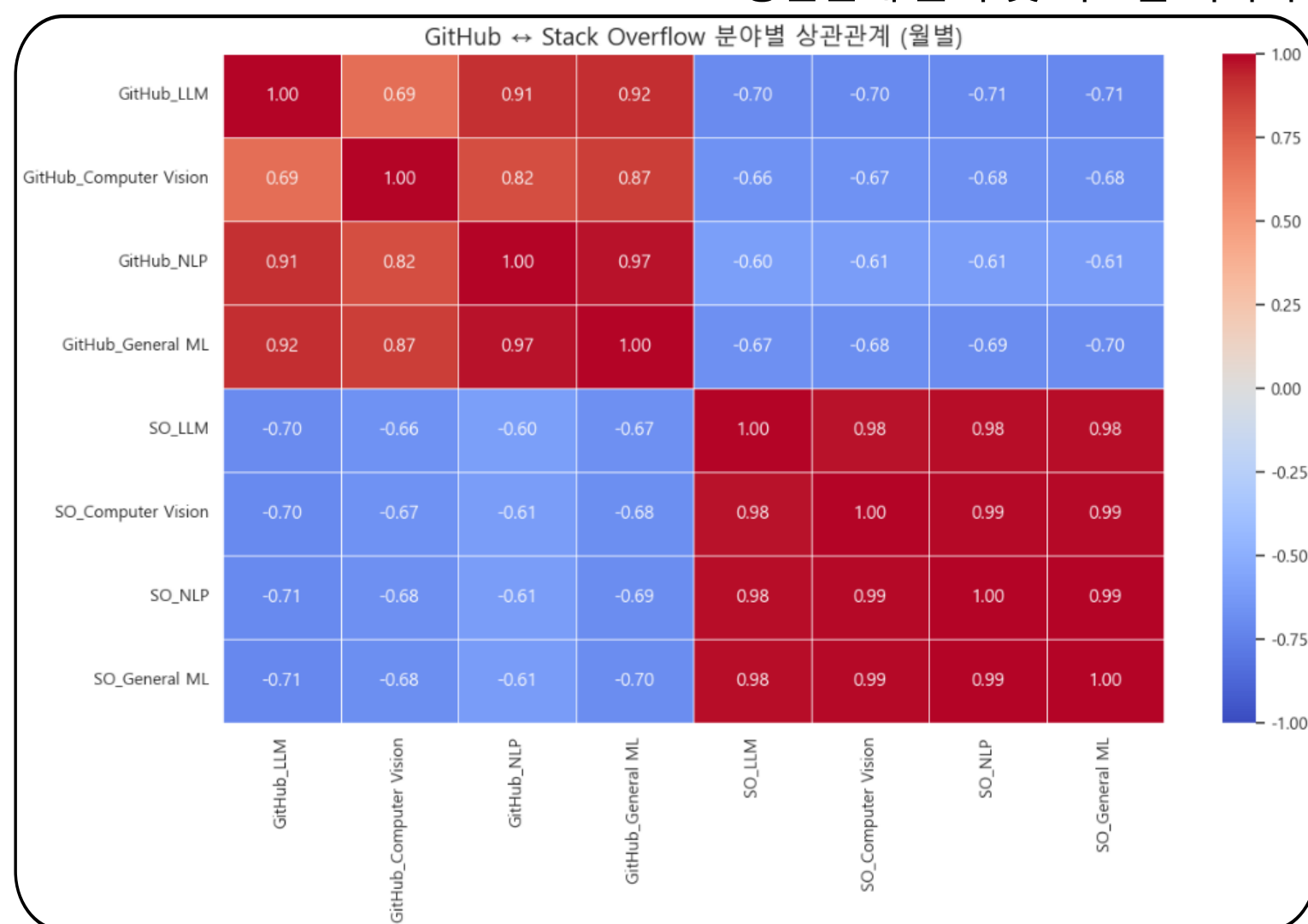


지속적인 상승 추세(Trend)를 바탕으로, 매년 4월과 11월에 활동이 급증하는 연간 패턴(Yearly)이 뚜렷하게 나타납니다.



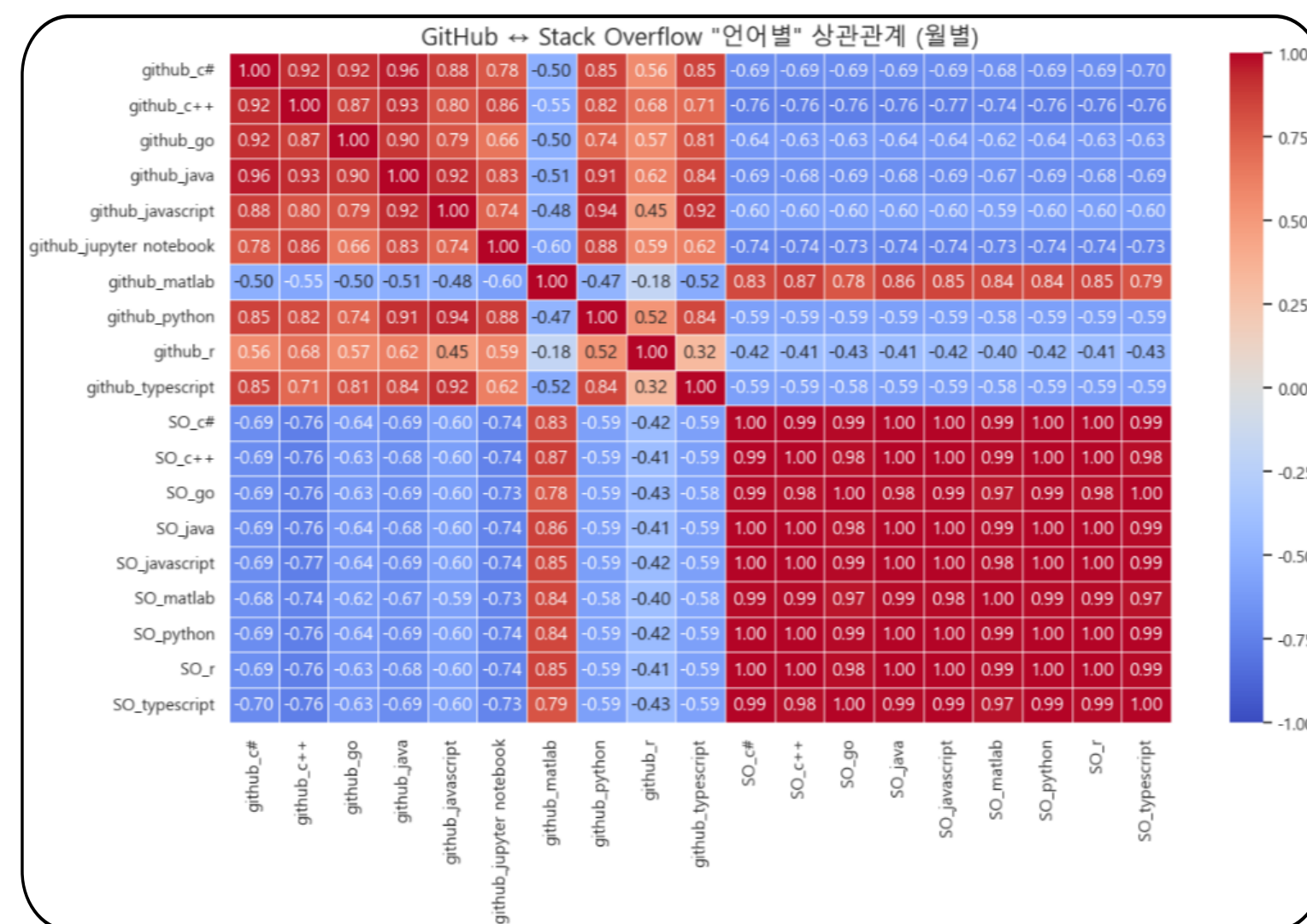
데이터 분석 & 시각화 - 4. Stack Overflow 질문 수와 GitHub 프로젝트 수 간 상관관계가 있는가?

상관관계 분석 및 히트맵 시각화



GitHub 프로젝트가 증가하는 동안, Stack Overflow의 AI 관련 질문은 오히려 줄어들었습니다.

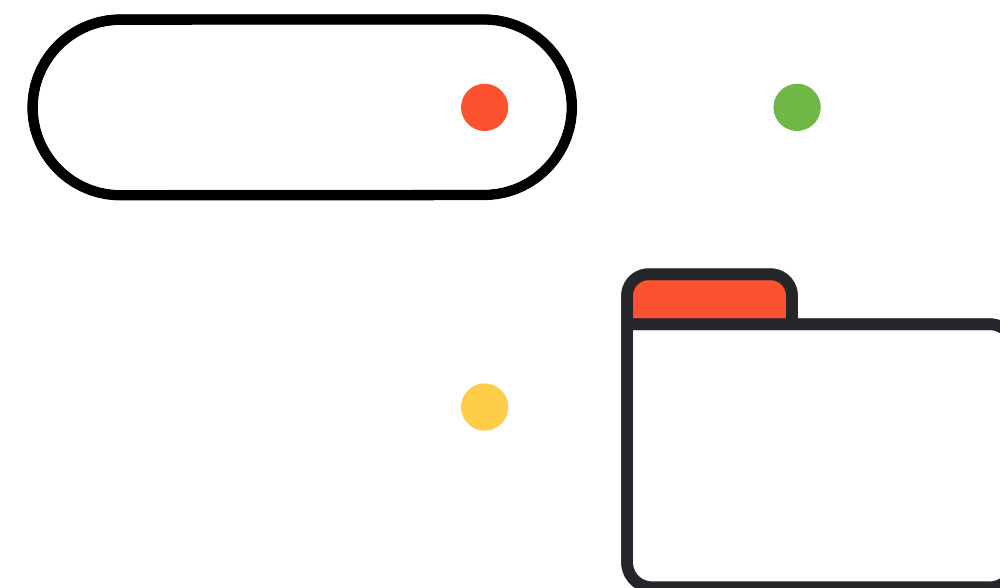
상관관계 분석 및 히트맵 시각화

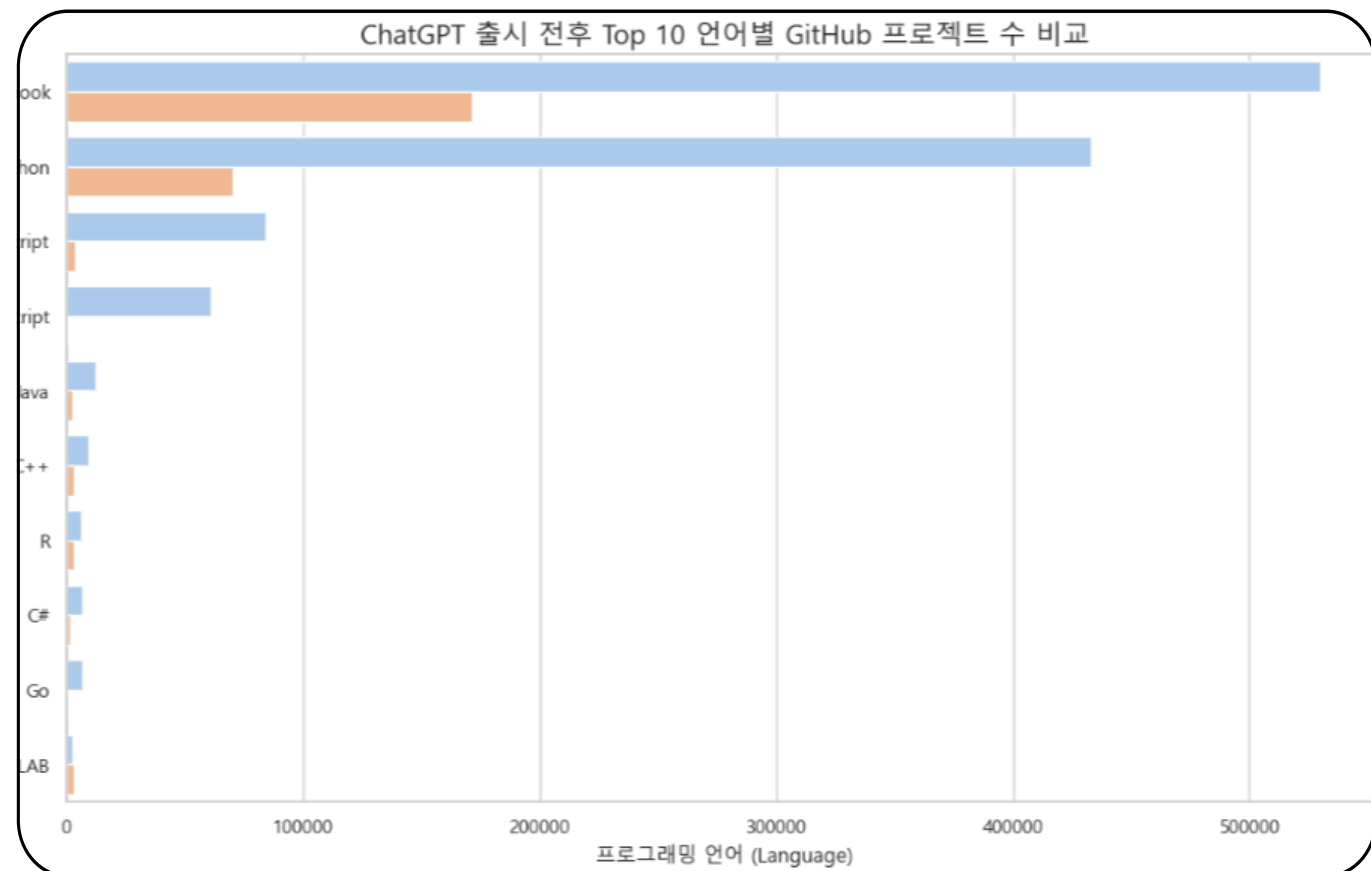
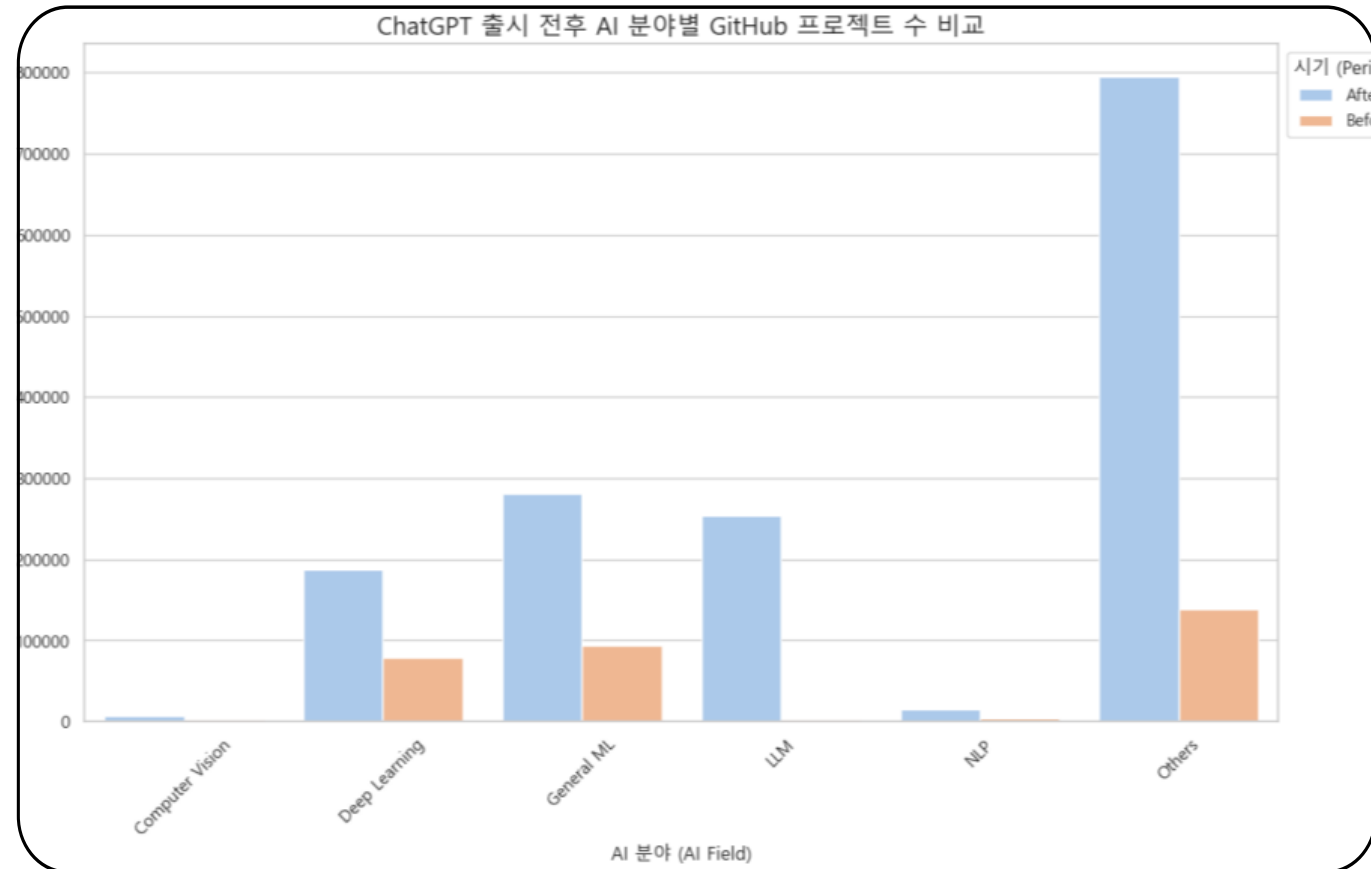


모든 AI 분야에서 GitHub 개발 활동이 활발해질수록 Stack Overflow의 질의응답 활동은 급격히 줄어드는 강력한 역의 상관관계가 나타납니다.



분석 결과 해석





1. ChatGPT 출시 전후(2021-2022 vs 2023-2025) GitHub AI 프로젝트는 어떻게 변화했는가?

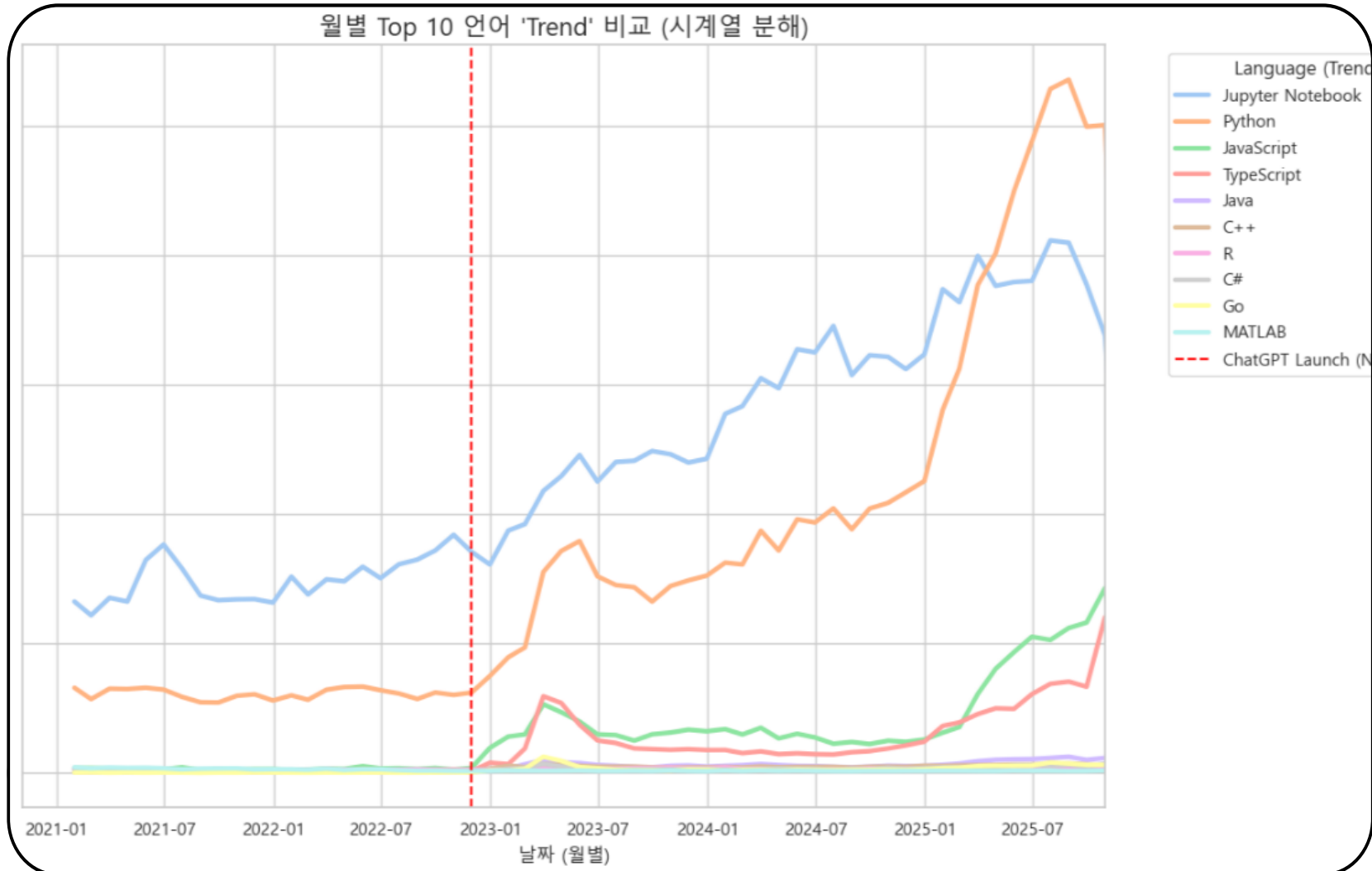
가설 : ChatGPT 출시 이후 AI 레포지토리는 이전대비 500% 이상 증가했을 것이다

1. AI 분야별 프로젝트 수 비교

- 양적 성장
- LLM의 부상
- 다양성 확대

2. 언어별 프로젝트 수 비교

- Python & Jupyter의 독주
- JS&TS 기술의 약진
- 전반적 개발 수요 증가



2. Github 레포지토리가 2023년 이후 실제로 증가했는가?

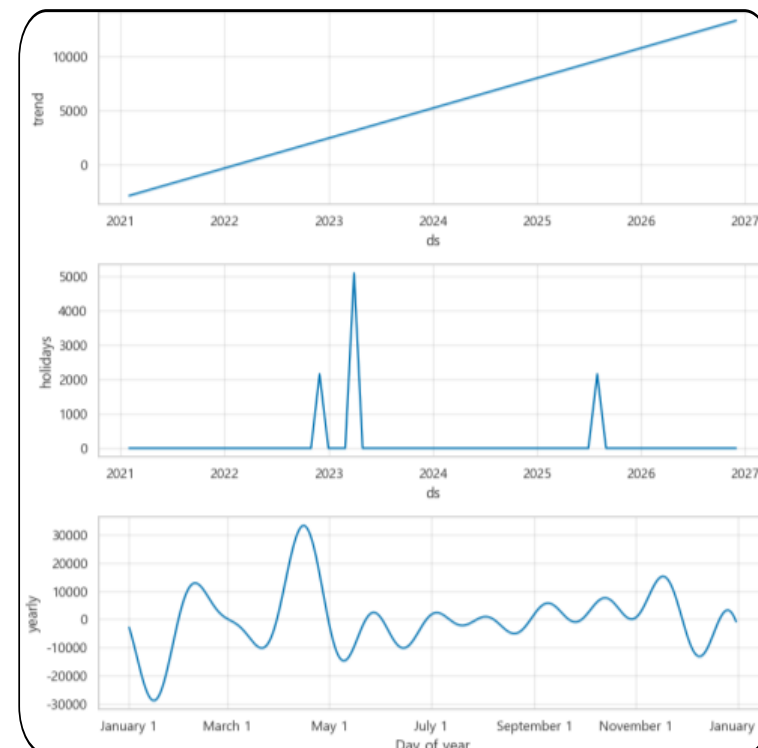
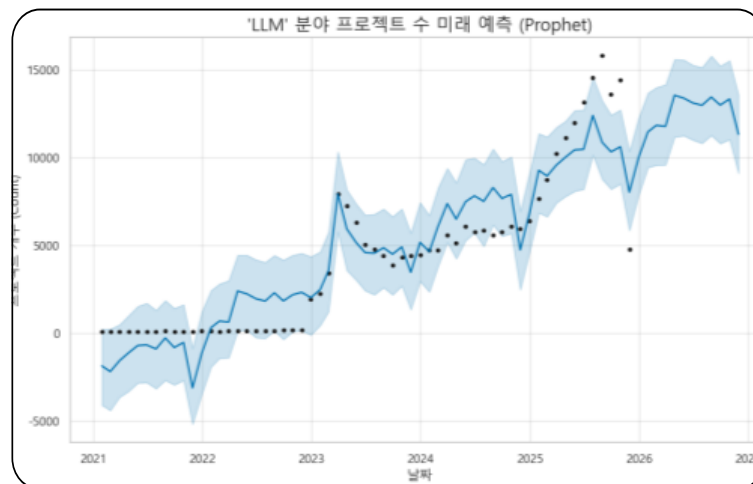
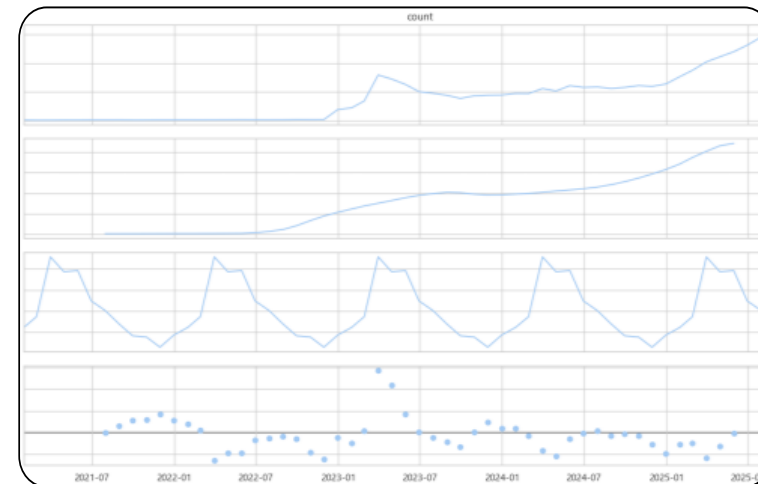
가설 : Python과 Jupyter Notebook은 23년 이후 꾸준히 상승했을 것이다.

1. ChatGPT 출시 직후 (2022.11 ~ 2023.5)

- Python, Jupyter Notebook, JavaScript, TypeScript 4개 주요 언어 급격한 증가
- 약 6개월간 가파른 상승 곡선 관찰

2. 바이브코딩 등장 이후 (2025.2 ~)

- Python이 다른 언어 대비 압도적으로 급증



3. AI 분야별(LLM, 강화학습, 컴퓨터비전, NLP) 프로젝트 증가 양상이 다른가?

가설 : LLM 분야가 2023년 이후 가장 높은 성장률을 보일 것이다

1. LLM 성장 추이

- ChatGPT 출시 전후로 확연한 차이
- 다른 AI 분야 대비 압도적 성장

2. 시계열 분해 분석

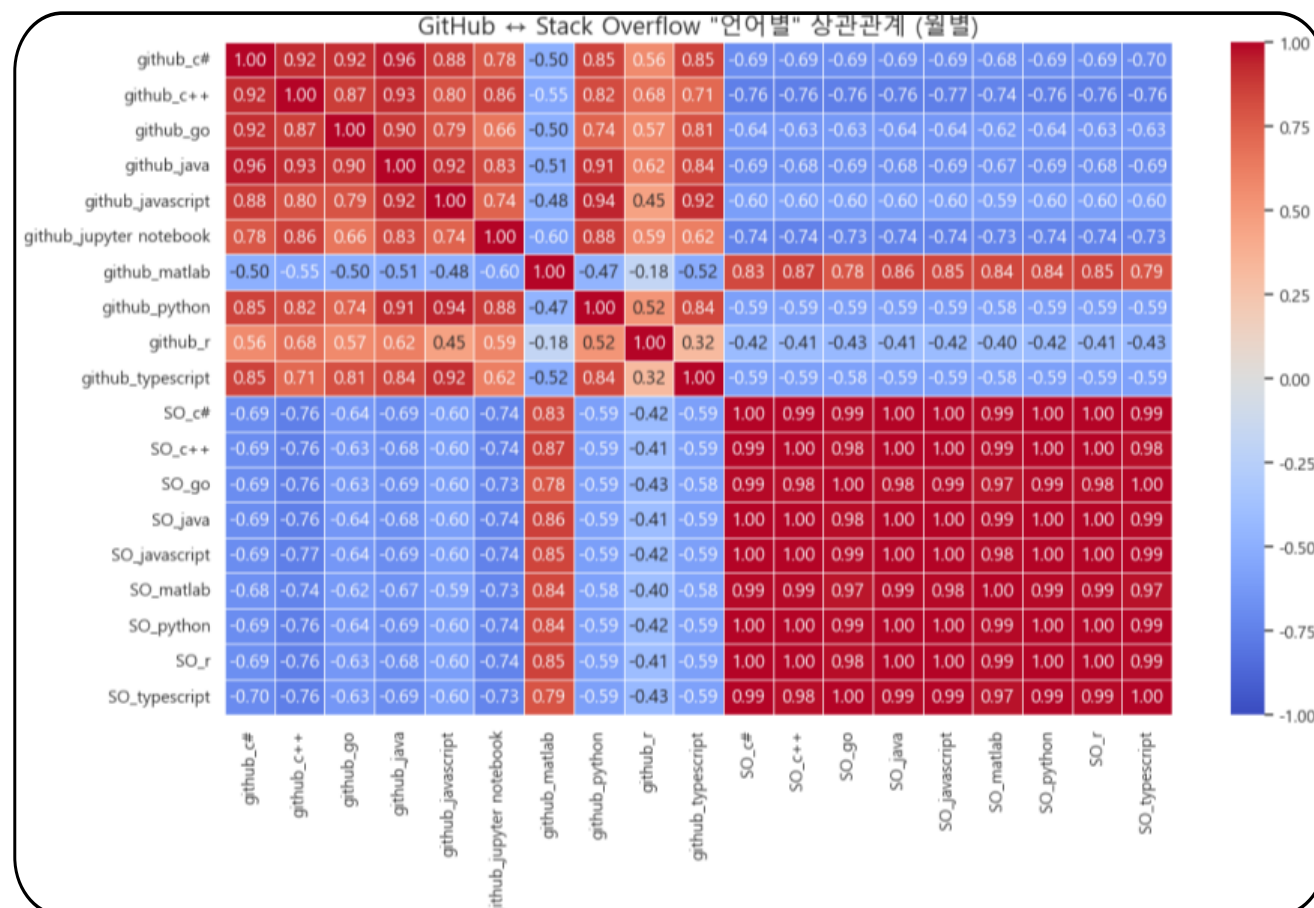
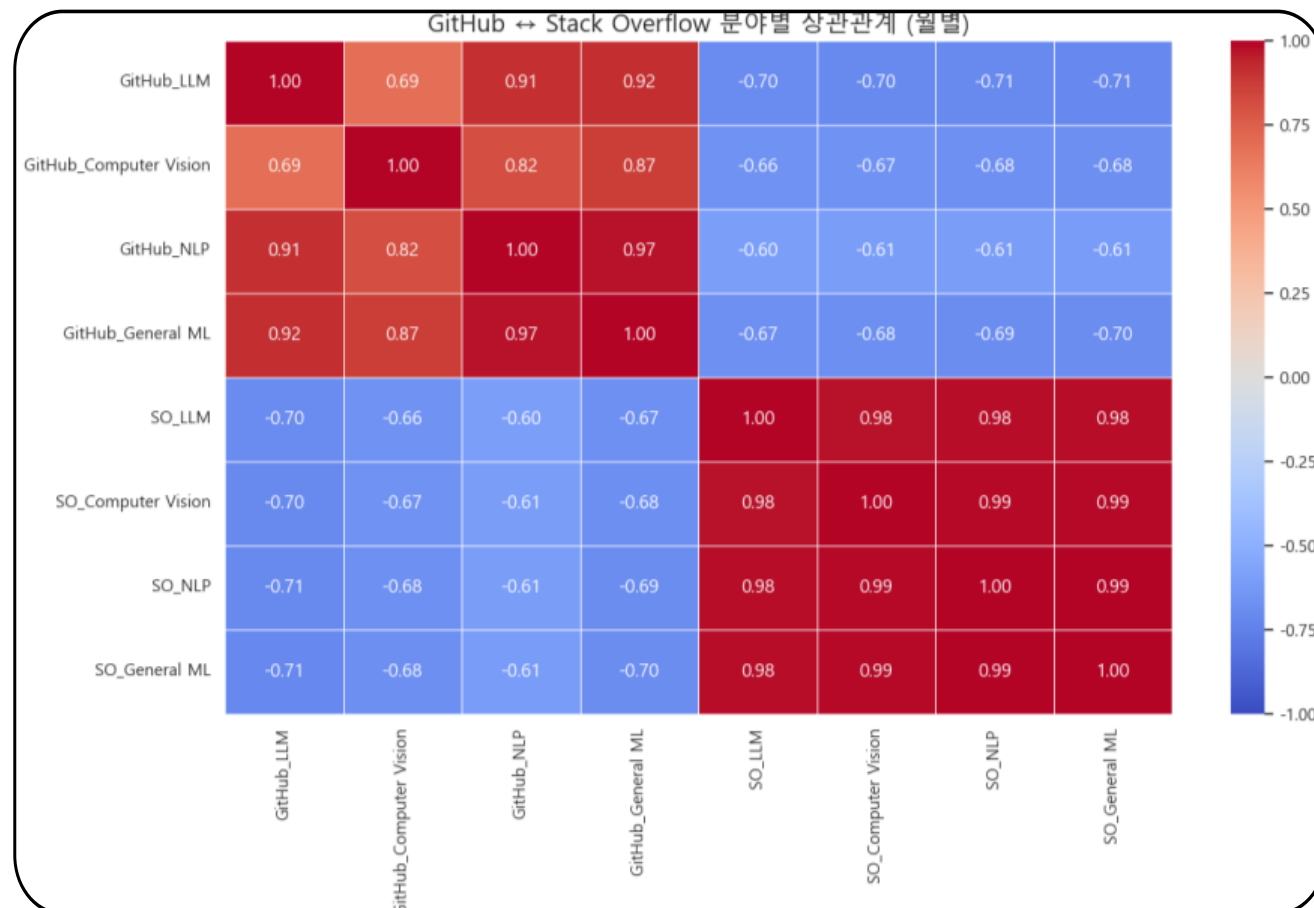
- 계절성: 상반기↑ 하반기↓ 반복 (IT기업 활동 주기 추정)
- 잔차: 2023년 초 스파이크 → 개발자들의 본격적인 ChatGPT 사용 시작

3. Prophet 예측

- 2023년 급등 이후 안정적 성장세
- 2027년까지 지속 증가 예측

4. Prophet 모델 분해

- 추세: 우상향 성장률 정확히 포착
- 이벤트 분리: 2023년, 2025년 스파이크를 계절성과 별개로 처리
- 순수 주기: 이벤트 제거 후 규칙적인 연간 패턴 확인



4. Stack Overflow 질문 수와 GitHub 프로젝트 수 간 상관관계가 있는가?

가설 : GitHub 프로젝트 증가율과 Stack Overflow 질문 증가율 간 강한 양의 상관관계($r > 0.7$)가 존재할 것이다

1. 강한 양의 상관관계 (0.7~1.0)

- GitHub 내부: AI 분야들끼리 함께 성장
- Stack Overflow 내부: 거의 동일한 패턴 (0.98~0.99)

2. 강한 음의 상관관계 (-0.6~-0.7)

- GitHub ↔ Stack Overflow: 반대로 움직임
- GitHub 활동 증가 → Stack Overflow 질문 감소

ChatGPT 이후 개발자 행동 변화 :
커뮤니티 의존도 감소

특이케이스 : Matlab은 특수 목적 언어로 양의 상관관계 유지

마무리

ChatGPT는 GitHub 오픈소스 생태계에 패러다임 전환을 가져왔다!

접근성 향상: Python으로 다양한 프로젝트 생성 가능
자율성 증대: 커뮤니티 의존도 감소, AI 도구 활용 자가 해결
LLM 중심 전환: 기존 AI 분야를 압도하는 새로운 생태계 형성
지속 가능한 성장: 일시적 유행이 아닌 구조적 변화로 확인

→ AI 도구는 개발 문화를 근본적으로 재편하고 있으며 이는 프로젝트의 결과로 명확히 입증되었다.



감사합니다

윤탈웅, 송민찬, 이라온, 원준서