Fast campus

The Red.

LLM 마스터클래스 #8

LLM 서비스 운영: The world

Part 8-1.
마켓 이야기

최근 뉴스 – 누가 무엇을 하고 있는가

NVidia

미치듯이 쏟아지는 모델들

AI ROI profitability 모델 - 누가 돈을 벌고 있나?

골드 러쉬때 돈 버는 사람들 은 청바지 회사와 곡괭이 회사

: 엔비디아, Azure/AWS

Big Tech winners and losers

AI 유니콘은 아직 없음

CES, 로보틱스 컨퍼런스와 AI

AI-empowered hardware and robots

- Phones
- Copilot PCs
- Rabbit
- Figure1
- Self-driving softwarae

컨텐츠 마켓

GenAI-powered contents

- Sora
- P*nhub
- Scarlet Johansson
- Adobe
- Copyrights, RAI, fakenews...

그래서 어디로 가는 건가요...? AGI는?

Quo Vadis...?

빅테크들끼리 싸우고, 우리는 결과만 쓰면 되는 걸까?

AGI는?

- AGI란 무엇인가
- AGI가 정말 도래할 것인가
- 강사의 의견

Part 8-2.

무엇을 배워야 할까,
무엇을 해야 할까?

마켓이 원하는 인재가...?

현재 학생으로 AI 인재가 되고 싶다면...?

- AI 인재 뺏기 전쟁?
- PhD?
- AI 인재가 되려면...?
- Deeplearning/ language, Computer vision

무엇을 할 것인가?

현재 직장인이라면 옵션은:

- 스타트업으로서 희망이 있는가?
  - 도메인이 확실해야 한다
  - 남들이 귀찮아 하는 일은 해야 한다 – 데이터 준비
    하지만 누가 돈을 내려고 하지?
- 뭔가 만들어보는 개발자
- 개인으로서 훨씬 더 유용함
- 큰 회사의 개발자 – LLM을 통합하기
- LLM을 도입하는 회사의 관련자
- 도메인 전문가가 되자

Level: Hard ★★★

AI 스타트업으로 돈 벌기

- Foundational model 은 돈이 너무 많이 듦
- Fine-tuned vertical: 기존 기업이 아니라면 데이터를 따라잡기 힘들고,
  제네릭 툴을 제공한다면 빅테크를 따라잡기 힘듦 (예: 의료 AI, 마케팅 툴 만들기)
- GPT Wrapper: OpenAI 가 뭔가 공개할 때마다 쓸려나감
- 내가 할 수 있다면 다른 사람도 할 수 있다

완전히 새로운 하드웨어에 맞는 모델을 만들고 그에 맞는 기능을 만들어 배포하기

- 새로 나오는 피아노 모델 / 태블릿 / 건설 현장에 쓰이는 기계등에 맞는 생성형 AI

Real-time, 틀리면 안 됨. 사용자 수 예측 불가

Level: Medium ★★☆

기존 툴에 생성형 AI 더하기

- 이미 마켓 쉐어가 있다면 가능성이 있음
  - 기존 병원 관리 앱에 AI 기능 더하기
  - 이미지 에디팅 툴에 생성형 AI 기능 더하기
  - 쉽게 sound effect 더하기
  - 백그라운드 쉽게 만들기
- 기존 앱 플러그인 아키텍쳐를 잘 안다면 가능성이 있음
  - 3D 에디팅 툴에 플러그인 더하기 등등
- 기존 아키텍쳐에 익숙하다면 가능성이 있음
  - 이미 workflow management, CI/CD, deploy, scale 등이 되는 스트럭쳐가 있고 거기에 생성형 AI 를 더함
    : 자동 답글 생성, 고객 관리 및 자동 이메일, 재고 관리 및 자동 alerting

Level: Easier ★☆☆

기존 시스템에서 조직이 필요로 하는 생성형 AI 기능 더하기

- 인프라 온리 – 고객 상대하지 않는 부분
- 기존 모델로의 API 콜로 가능
- 파인 튜닝 필요 없음
- 기본적인 언어 이해로 가능함
- 텍스트 기반
- Async 가능
- 배치 가능

Part 8-3.
그 외 이야기

AI 혁명, 어디로 갈 것인가?

하이프 사이클

몇몇만 제외하면 다들 지는 게임

언제나 빨리 변하고 있는 IT

- 상전벽해
  - 22년 전과 지금 - No front-end, backend
  - Javascript - server side
  - Python
  - No mobile
- 인터넷, 핸드폰, 모바일 앱, 클라우드, 빅데이터
- "뭘 배워야 대박 날까요?" 질문
- 십오년 전에는 데이터
- 몇 년 전에는 암호화폐, NFT, Web3, computer vision, NLP ...
- 이제는 AI

신기술의 사이클

↓ 현재 여기

믿을 수 없는 성과를 보이는
신기술의 POC, 기술 주도자
몇몇의 큰 성공

신기술이 도입되고 나서
첫 1-3년, 하이프 사이클

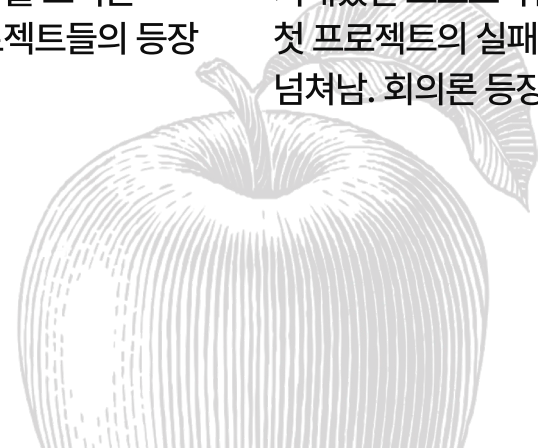
새로운 툴과 프레임워크,
플랫폼등의 춘추전국 시대.
노동시장에서 새로운
직함과 전문가들의 하이프
사이클및 스카우트 경쟁

기대했던 프로토타입이나
첫 프로젝트의 실패가
넘쳐남. 회의론 등장

성과를 보이는
프로젝트들의 등장

좀 더 안정적인 기술 스택,
방법론, 관리방법,
커리큘럼 등의 등장

기존 기술/ 시스템으로의
완전한 편입



사이클을 버텨내는 개발자 - 1

초반

- 우왕좌왕 - headless chicken
- POC, greenfield
- 망함, 이직, 창업
- 회귀자를 원하는 리크루터들
- 엄청난 투자와 스타트업들
- 벼락부자 등장, 뉴스 도배, "인류의 미래는 ..... " 등의 제목 등장

사이클을 버텨내는 개발자 - 2

중반

- No plan survives first encounter with reality
- No plan survives first contact with the enemy (Moltke)
- 중간 매니저들 물갈이
- 부서 구조조정
- 투자 취소
- 승자 독식 패턴이 드러나는 분야
  - 플랫폼, 라이브러리, 새로운 언어, 서비스, 툴셋
  - 예: MS 는 모바일을 완전히 접음, 구글은 SNS 포기

사이클을 버텨내는 개발자 - 3

안정기

- 2-5년 후: 거의 모든 기업이 기술 도입. 상용된 제품 혹은 in-house
- 표준이 된 테크 스택 2-3개가 업계 표준이 됨
- 새로운 직함, 직종, 테크 스택이 구인광고에 현실적으로 등장
  - "모바일 개발자"
  - "자바 백엔드 개발자"
  - "UI/UX 디자이너"
  - "데이터 과학자"

AI 현재 사이클

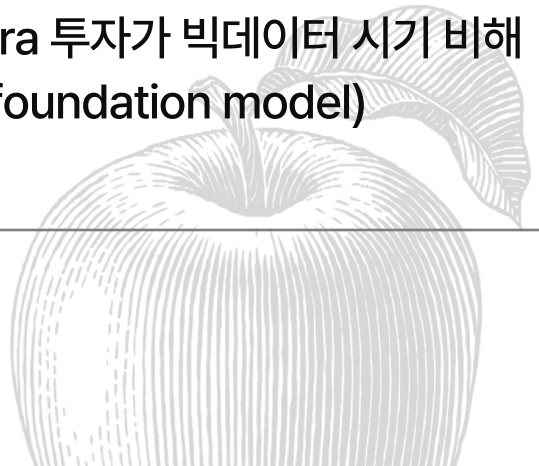
초반

- 우왕좌왕 - headless chicken
- POC, greenfield
- 망함, 이직, 창업
- 회귀자를 원하는 리크루터들
- 엄청난 투자와 스타트업들
- 벼락부자 등장, 뉴스 도배,
  "인류의 미래는 ..... " 등의 제목 등장

+ 중반

이유:

- 간단하게 API로 활용 가능.
  이미 아는 기술
- POC 만들기 정말 쉬움
- Ops/Infra 투자가 빅데이터 시기 비해 적음 (x foundation model)



드루와 지금 드루와

- 안정화까지 걸릴 시간이 훨씬 단축됨
- 개발자로서 앞으로 1-2년 내에 예상할 수 있는 상황
  - 한국에서 주력으로 쓰게 되는 모델과 테크 스택, 플랫폼의 등장
  - 그것으로 간단한 서비스를 만들던지 기존 시스템을 증강할 수 있는 예시 구현이 프로덕션에 디플로이
  - 거의 모든 도메인에서 fine tuning / RAG 를 위한 데이터셋 등장. 그리고LLM 시스템 컴포넌트 도입 시작.
  - LLMOps 의 확장
- 수학, 딥러닝 몰라도 됨. 박사학위 없어도 됨.