생성형 AI 기반 RAG 개발자 과정

GPT·ChatGPT 진화와 정렬

Session 3 / 33 — Day 1

GPT-3.5 → 4 → 5, 그리고 RLHF·MoE·추론 모델

박수현 · (사)한국정보공학기술사회 · 2026

LLM 이론 · LangChain · ChromaDB CRUD · Agent — 4일 핸즈온

이번 시간을 마치면

📈 세대별 진화

- GPT-1 → 5 가 **무엇을** 바꿨나
- 규모·정렬·양식의 3축

🎯 정렬(Alignment)

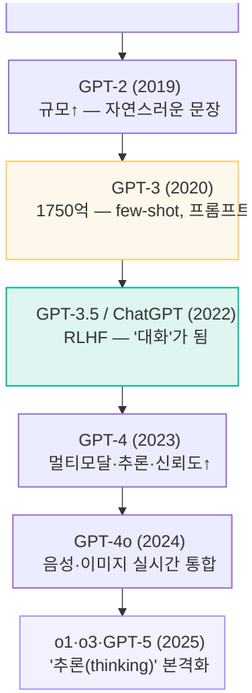
- **RLHF** 3단계 파이프라인
- **Safety** — 안전한 답을 만드는 법

⚙️ 최신 아키텍처

- **MoE** — 큰데 빠른 비결
- **Thinking(추론) 모델** — o 시리즈
- 우리 코스 **모델 선택 가이드**

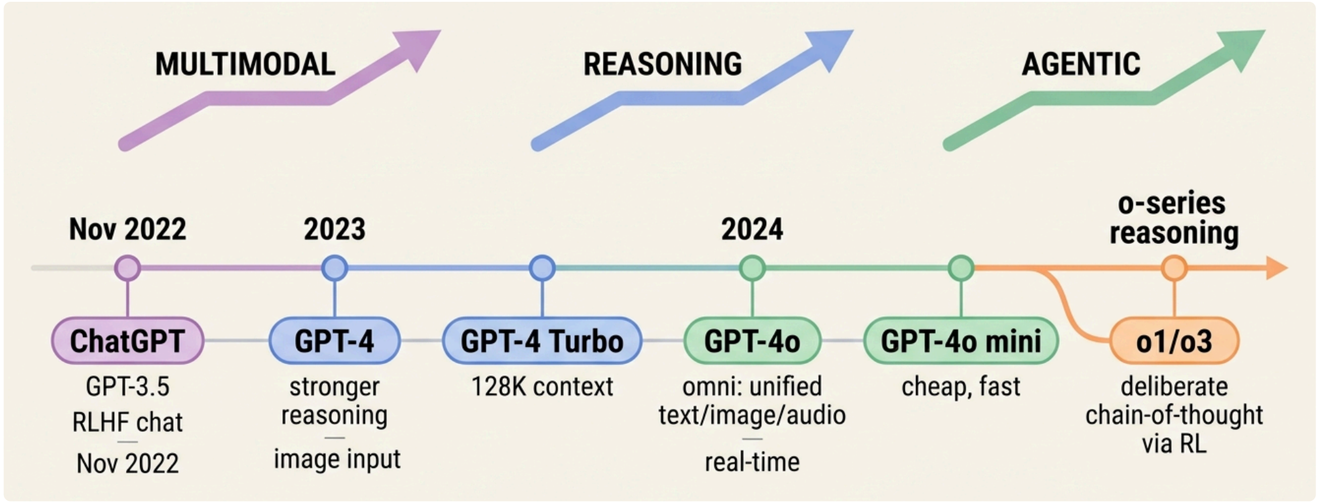
> 🎯 목표: API 로 모델을 **고를 때 판단 기준**을 갖는다.

무엇이 세대마다 바뀌었나



💡 3축 추세: **규모(파라미터)** · **정렬(사람 의도)** · **양식(텍스트→음성·이미지·추론)**.

우리가 쓰는 모델들, 한눈에



> 2022 이후 GPT 계열 · ChatGPT(3.5)→GPT-4→GPT-4o(멀티모달)→o 시리즈(추론). 멀티모달·추론·에이전트로 확장

세대	결정적 변화	개발자 체감
GPT-3.5	**RLHF** — 대화·지시 따르기	ChatGPT 경험, 저렴한 실습 모델
GPT-4	추론·신뢰도·멀티모달	복잡 작업·이미지 입력
GPT-4o	속도·비용·실시간 멀티모달(옴니)	`gpt-4o-mini` = 가성비 메인
o 시리즈	**추론(thinking)** — 풀기 전 '생각'	수학·코딩·다단계 문제
GPT-5 / 5.5	즉답+추론 **통합 라우터**	난도 따라 경로 자동 분기

PART A

PART A
정렬(Alignment) — RLHF

> 유능한 모델을 사람 의도에 맞게 정렬하기 — 약 12분

GPT-3 는 유능했지만 — 실사용과는 거리가 있었다

사전학습만 된 모델의 문제

- 인터넷 텍스트로 **다음 단어 예측**만 학습
- 질문에 **답하기보다 이어쓰기**를 함
- Q: "한국의 수도는?" → A: "일본의 수도는? 중국의..."
- 무례하거나 위험한 출력 가능

핵심 간극

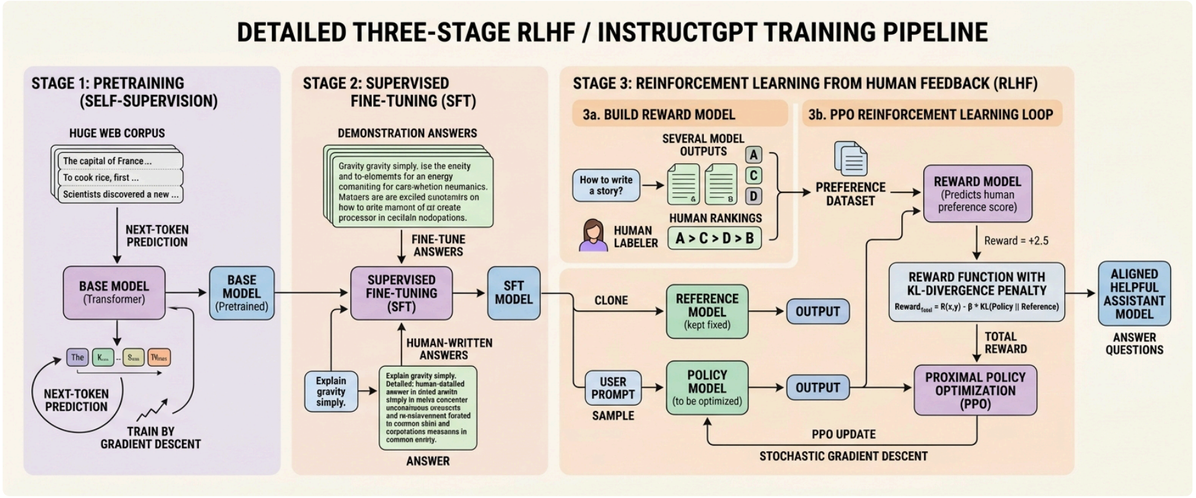
- **능력(can) ≠ 의도 정렬(should)**

정렬(Alignment) 이란

- 모델 출력을 **사람의 의도·가치**에 맞추는 것
- 3가지 목표(흔히 **HHH**):
- **Helpful** (도움이 되고)
- **Honest** (정직하고)
- **Harmless** (해롭지 않게)

> 🔑 이 정렬을 해낸 기법이 **RLHF** — 다음 슬라이드.

RLHF — 사람 피드백으로 다듬기



RLHF 3단계 · 사전학습 → 지도 미세조정(SFT) → 보상모델+강화학습으로 사람 선호에 정렬

단계	입력 데이터	얻는 것
① 사전학습	대규모 일반 텍스트	언어 능력·지식의 바탕
② SFT	사람이 쓴 모범 답변	질문에 답하는 형식·어조
③ 보상모델 + RL	응답에 대한 사람의 **선호 순위**	사람 선호에 맞는 정렬

💡 GPT-3 → ChatGPT 의 결정적 차이는 새 모델이 아니라 **이 RLHF 한 겹**이었습니다. 핵심은 **지식 주입이 아니라 사람 선호 정렬**.

출처: Day 1 GPT·ChatGPT 진화

RLHF — 만능은 아니다

✅ 얻은 것

- 지시 따르기, 대화 자연스러움
- 거절해야 할 요청 거절
- 위험·편향 출력 감소

🆕 변형들

- **RLAIF** — 사람 대신 **AI 피드백**
- **DPO** — 보상모델 없이 선호 직접 최적화 (더 단순)

⚠️ 한계와 부작용

- **아첨(sycophancy)** — 사용자에게 듣기 좋은 답에 치우침
- **과잉 거절** — 멀쩡한 요청도 막음
- 사람 라벨러의 **편향**이 스며듦
- 정렬해도 모델이 **모르는 최신·사내 정보**는 못 만듦 → **환각** 잔존

> 🔑 정렬(RLHF)은 **선호를 맞출 뿐 지식을 못 넣는다.** 그래서 정렬된 모델 위에 **근거(RAG) + 검증**을 얹는다.

PART B

PART B
Safety / Alignment

> 안전하게 만드는 장치들 — 약 8분

출처: Day 1 GPT·ChatGPT 진화

Safety — 모델 안과 밖, 다층 방어

모델 '안' 정렬

- RLHF 로 위험 요청 거절 학습
- **System Prompt** 로 역할·금지선 지정
- **Red-teaming** — 출시 전 공격 테스트

모델 '밖' 가드레일

- 입력/출력 **모더레이션 API**
- 민감정보 **필터링·비식별화**
- 사용량·권한 **제한**

> 💡 개발자 책임 구간 = **System Prompt + 입출력 필터**. 모델만 믿지 말 것.

출처: Day 1 GPT·ChatGPT 진화

우리가 코드로 다루는 안전 레버

레버	어디서	코스 연결
System Prompt	매 호출 messages[0]	Session 4·5
temperature 낮춤	파라미터	Session 5 — 사실성↑
RAG 근거 주입	검색+프롬프트	Session 6~ — 환각↓
출처 표시·검증	RAG 체인	Session 16·22 — 신뢰도
민감정보 처리	전처리	Session 6 — 비식별 개념

PART C

PART C
최신 아키텍처 — MoE · Thinking

> 큰데 빠른 모델, 생각하는 모델 — 약 10분

MoE — "전문가 여럿, 매번 일부만 호출"

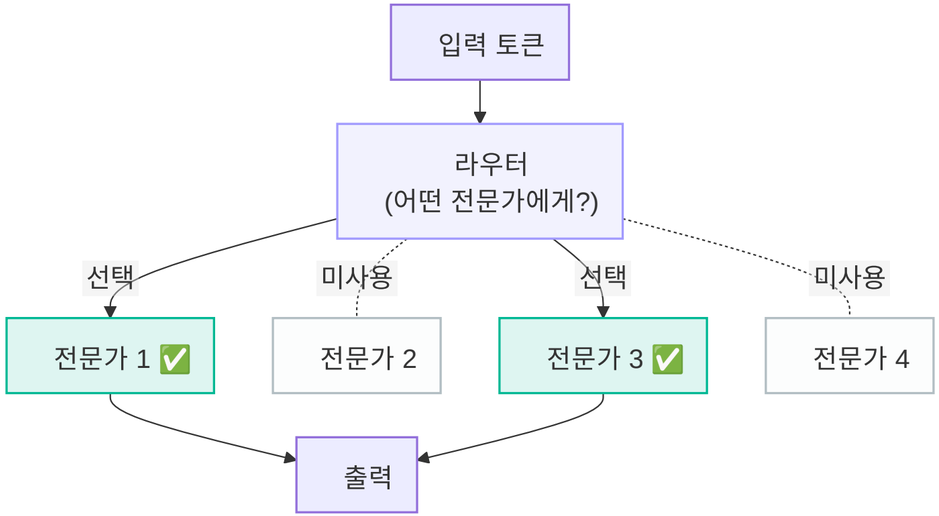
아이디어

- 거대한 신경망을 **여러 전문가**로 분할
- **라우터**가 토큰마다 **일부 전문가만** 활성화

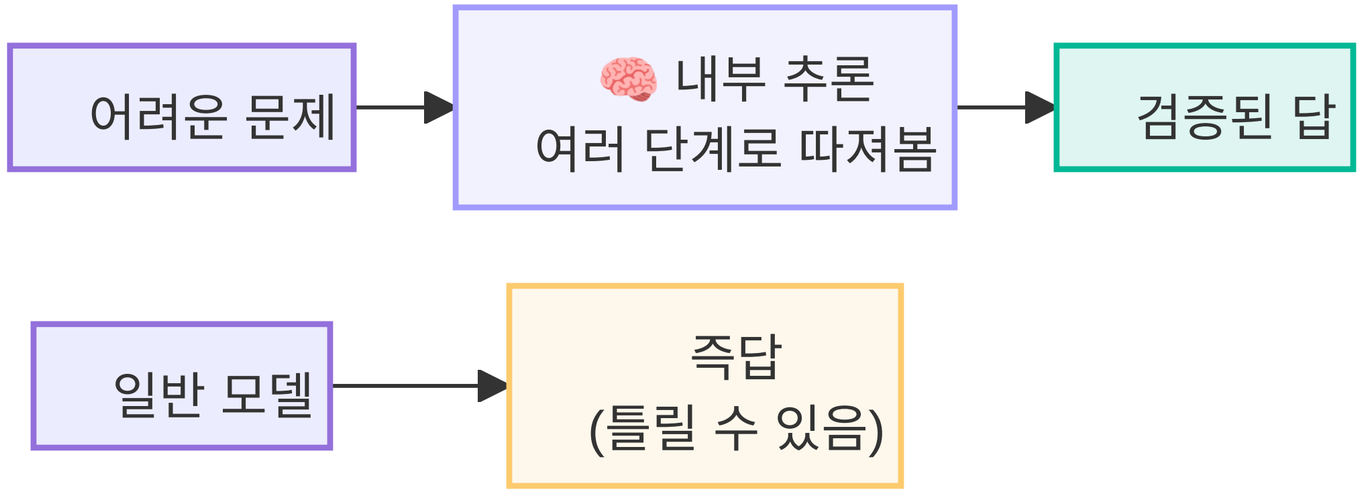
왜 좋은가

- 전체 파라미터는 **크게**(지식 ↑)
- 매 호출 연산은 **작게**(속도·비용 ↓)

> 💡 "큰 용량 + 빠른 추론"을 동시에. 최신 대형 모델의 흔한 비결.



추론 모델 — "답하기 전에 생각한다"



무엇이 다른가 (o 시리즈 등)

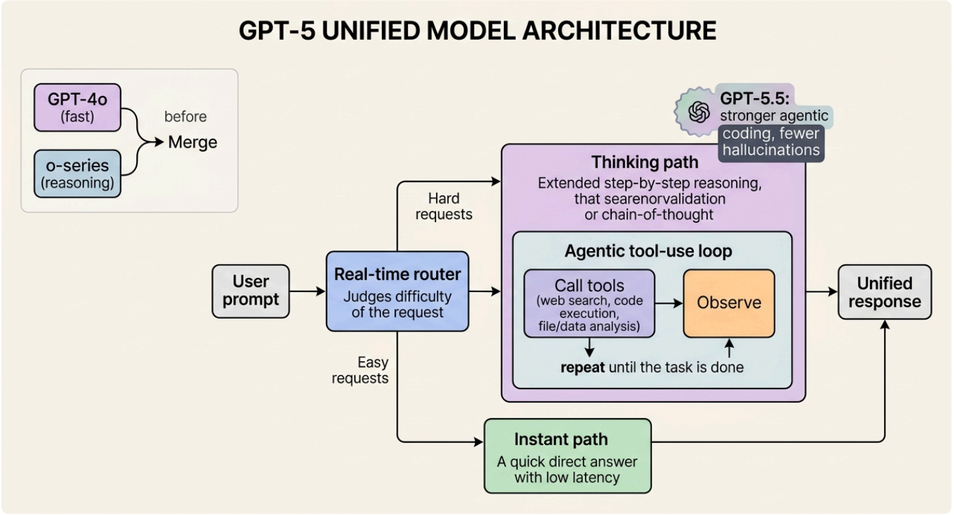
- 답을 내기 전 **긴 사고 과정**(chain-of-thought)을 거침
- **추론 시간을 더 쓰면** 정답률 ↑ (test-time compute)
- 수학·코딩·다단계 추론에 강함

⚖️ 트레이드오프

- **느리고 비쌈** (토큰 더 소모)
- 단순 작업엔 과함

> 🔑 선택 기준: **쉬운 작업 → 빠른 모델**, **어려운 추론 → 추론 모델**. Day4 Agent 에서 모델 라우팅으로 연결.

GPT-5 — "난도를 보고 경로를 스스로 고른다"



> GPT-5 통합 구조 · 실시간 라우터가 난도를 판단해 즉답/추론 경로로 자동 분기(과거 GPT-4o·o 시리즈 통합)

무엇이 합쳐졌나

- 과거: 빠른 **GPT-4o** vs 추론 **o 시리즈**를 사용자가 직접 선택
- GPT-5: **실시간 라우터**가 난도 판단 → 즉답/추론 경로 자동 분기
- 추론 경로엔 도구 호출·검증을 반복하는 **에이전트 루프**

개발자 관점

- 모델명·버전은 자주 바뀜 → **상수/환경변수**로 한곳에서 관리
- 선택 기준은 그대로: **추론 난도 · 응답 속도 · 비용**

> 🔑 라우터가 골라줘도 **"왜 이 경로인가"** 를 이해해야 비용·지연을 통제한다.

출처: Day 1 GPT·ChatGPT 진화

언제 어떤 모델을 부를까

상황	권장 모델	이유
일반 챗봇·요약·RAG 답변	**gpt-4o-mini**	빠르고 저렴, 충분히 똑똑
임베딩(검색용 벡터)	**text-embedding-3-small**	RAG 표준, 매우 저렴
복잡 추론·어려운 코딩	o 시리즈 / GPT-5	정확도 우선, 비용 감수
비용 0·오프라인·민감정보	**로컬 LLM(Ollama)**	Day4 — 프라이버시
모델 교체 유연성	(LangChain 추상화)	Day2 — 한 줄로 교체

💡 핵심 원칙: **추론 난도·응답 속도·비용** 세 축으로 고른다. 비싼 모델을 모든 호출에 쓰지 않고, **모델명은 상수/환경변수**로 한곳에서 관리한다.

핵심 6가지

1. ✅ GPT 진화 3축: **규모 · 정렬 · 양식**
2. ✅ GPT-3 → ChatGPT 의 핵심은 **RLHF** — **지식 주입이 아니라 사람 선호 정렬**
3. ✅ RLHF 3단계: **사전학습 → SFT → 보상모델+강화학습**, 정렬만으론 **환각 못 막음 → RAG**
4. ✅ 2022 이후: GPT-4 → **4o(멀티모달)** → **o 시리즈(추론)** → **GPT-5 통합 라우터**
5. ✅ **MoE** = 큰데 빠름, **Thinking/GPT-5** = 난도 보고 즉답·추론 경로 자동 분기
6. ✅ 모델 선택은 **난도·속도·비용** 기준 — 메인은 `gpt-4o-mini`, 모델명은 **상수/환경변수**