

5회차 · 120분

AI 기술 기초

왜 그렇게 동작하는지 알면, 무엇을 조심할지가 예측됩니다

언어 모델이 하는 일은 정말 하나뿐입니다

지금까지의 글을 보고, 다음에 올 조각 하나를 고른다. 그리고 그걸 반복한다.

그런데 왜 요약이 되나

요약을 시키는 게 아니라, 요약이 나올 수밖에 없는 상황을 만드는 것

(긴 글 1,000줄)

위 글의 요약:

→ 다음에 올 자연스러운 말 = 요약문

한국어: 오늘 회의는 취소되었습니다.

영어:

→ 다음에 올 자연스러운 말 = 번역문

따라 나오는 것

무엇을 잘하고 무엇을 못 하나

잘하는 것

왜

사람 글에 흔한 패턴

예 1

요약 · 번역 · 형식 변환

예 2

초안 작성 · 분류

예 3

문장 다듬기

못 하는 것

왜

✓ 본 적 없거나 계산이 필요

예 1

✓ 정확한 계산

예 2

✓ 최신 사실 · 내부 정보

예 3

✓ 매번 동일한 출력

오른쪽 네 가지는 전부 “도구”로 해결합니다. 모델을 키워서 해결되는 게 아닙니다.

환각

“모르겠습니다”는 자연스러운 이어짐이 아닙니다

질문

우리 회사 2023년 매출이 얼마였지?

답변

2023년 매출은 약 47억원입니다.

질문 뒤에 오는 자연스러운 말은 답입니다. 아는 게 없어도 **형태는 만들어집니다.**

환각 줄이기

기억에서 만들게 하지 말고, 자료를 주고 그 안에서 답하게

위험

질문 방식

"근로기준법 연차 규정 알려줘"

무엇을 하나

기억에서 생성

출처

없음 (지어낼 수 있음)

안전

질문 방식

✓ "아래 조문에서 정리해줘" + 조문

무엇을 하나

✓ 주어진 자료에서 추출

출처

✓ 내가 준 자료

검색·DB 도구 연결도 같은 원리입니다 — 지어낼 필요를 없애는 것.

실습 · 30분

자료를 붙여보기 — 네 번 물어봅니다

1

자료 없이

우리 규정을 그냥 묻기. 숫자를 말하면
지어낸 것

2

붙이고 다시

"이 문서만 근거로. 쪽수를 붙여줘"

3

함정 질문

문서에 답이 없는 것을 묻기

4

한 문단만

잘라 붙이면 답이 흔들립니다

2단계에서 붙은 쪽수를 실제로 열어보세요. 출처가 붙었다는 것과 출처가 맞다는 것은 다릅니다.

“우리 데이터로 학습시키면?”

자료에 넣을까, 가중치에 넣을까

자료 붙이기(RAG)

무엇을 바꾸나
모델이 아는 것

언제 반영
문서 고치면 다음 질문부터

착수 비용
며칠~몇 주

정책이 바뀌면
한 줄 고치면 끝

파인튜닝

무엇을 바꾸나
✓ 말투 · 형식 · 판단 습관

언제 반영
✓ 재학습해야 (수 시간~수 일)

착수 비용
✓ 몇 주 + 데이터 만드는 사람 시간

정책이 바뀌면
✓ 그때까지 옛 정책을 확신에 차서 말함

파인튜닝은 사실을 집어넣는 도구가 아닙니다. 지시 → 자료 붙이기 → 도구 → 검색 구축 → 파인튜닝 순서로만 내려가세요.

학습의 종류

네 가지 다 학습입니다 — 어디에 남느냐가 다릅니다

01

이번 대화 안에만

지시 — 이번 요청의 행동. 인컨텍스트 러닝이라 부르고, 대화가 끝나면 사라집니다

02

자료에

자료 붙이기(RAG) — 아는 것. 비모수적 기억이라 모델이 아니라 저장소가 배웁니다

03

도구와 절차에

도구 · 스킬 · 평가 · 가드레일 — 할 수 있는 일. 모델을 갈아타도 남습니다

04

모델 가중치에

파인튜닝 — 말투 · 형식. 사전학습 · 정렬 다음의 맨 아래 한 칸

아래로 갈수록 오래 남고, 고치기 어렵고, 비쌉니다. 되돌리기가 가장 비싼 칸을 먼저 잡는 것이 가장 흔한 실수입니다.

도구 사용

모델에 손을貸아주기



4단계가 핵심입니다 — 실행은 시스템이 하므로 그 사이에 사람 승인을 끼울 수 있습니다.

도구를 한 번 쓰는 것과 스스로 반복하는 것

도구 사용

반복

1회

사람 개입

매 턴

실패 양상

한 번 틀리면 한 번

사람의 일

질문하고 답 보기

에이전트

반복

✓ 목표 달성까지

사람 개입

✓ 시작과 끝

실패 양상

✓ 틀린 전제 위에 계속 쌓음

사람의 일

✓ 목표를 정확히 주고 멈출 지점 정하기

에이전트는 틀린 방향으로도 열심히 갑니다. 그래서 사람의 일이 앞으로 이동합니다.

컨텍스트

모델은 아무것도 기억하지 않습니다

01

매번 다시 넣는 것

대화가 이어지는 건 지난 대화를 통째로 다시 보내기 때문

02

창에는 크기가 있음

차면 오래된 것부터 밀려납니다 — 대개 앞쪽에 중요한 게 있습니다

03

뒤로 갈수록 비쌌

20번째 요청 = 앞의 19턴 전부 + 새 질문

“모델이 멍청해진 것 같다”의 상당수가 이것입니다. 모델은 그대로이고 보고 있는 것이 달라졌습니다.

기억

창을 닫으면 남는 것이 없습니다

- 1층 — 지금 보고 있는 대화 전체. 창을 닫으면 사라집니다
- 2층 — 앞부분을 압축한 요약. 이것도 창 안에 있어 함께 사라집니다
- 3층 — 대화 밖 파일에 적어둔 것. 지우기 전까지 남습니다
- 밀려나기 전에도 나빠집니다 — 끝난 중간 단계가 쌓이면 품질이 조용히 떨어집니다

메모리에 자격증명을 적지 마세요. 한 번 적히면 이후 모든 세션의 컨텍스트로 다시 읽힙니다.

좋은 지시

네 가지 요소

01

목표

빠지면 엉뚱한 것을 만듦. “무엇이 문제다”로 적으면 명확해집니다

02

범위

빠지면 건드리면 안 될 곳을 건드림. 하지 말 것으로 적으세요

03

완료 조건

빠지면 판정이 갈림. 판정이 수동이면 자동화 이득이 깎입니다

04

맥락

빠지면 바퀴를 다시 발명. 이미 실패한 것도 알려주세요

일회성 질문에는 과합니다. 반복 작업과 에이전트에 맡기는 작업에 씁니다.

실습 · 30분

나쁜 지시 다섯 개를 직접 고쳐봅니다

나쁜 지시

보고서

지난달 매출 보고서 만들어줘

메일

거래처에 사과 메일 써줘

정리

이 데이터 정리해줘

무인 작업

매일 아침 매출 슬랙에 올려줘

빠진 요소

보고서

✓ 맥락 — B제품 단종을 안 주면 이상 징후로 해석

메일

✓ 범위 — 보상 언급 금지. 안 적으면 약속이 됩니다

정리

✓ 완료 조건 — “추측해서 채우지 말 것”

무인 작업

✓ 실패 시 동작 — 틀린 숫자보다 안 올리는 게 낫습니다

사람이 중간에 보지 않는 작업은 “잘 됐을 때”가 아니라 “안 됐을 때”를 적어야 합니다.

평가(EVALS)

개인은 눈으로 판정하지만 조직은 못 합니다

개인

누가 판정

본인

언제

즉시

기준

자기 감

하루 건수

몇 건

조직

누가 판정

✓ 매번 다른 사람

언제

✓ 나중예, 또는 안 함

기준

✓ 사람마다 다름

하루 건수

✓ 수백~수천

수백 건을 사람이 다 보면 자동화의 이득이 사라집니다. 그래서 판정 자체를 자동화합니다.

평가 시작하기

케이스 20개면 시작할 수 있습니다

- 1주차 — 실제 케이스 20개를 스프레드시트에 (입력 / 기대 결과)
- 2주차 — 돌려서 몇 개 맞는지 센다. 이게 기준선
- 3주차 — 프롬프트를 고칠 때마다 다시 돌린다
- 이후 — 틀린 케이스가 나오면 데이터셋에 추가

AI 출력은 매번 조금씩 다릅니다. "100% 통과"가 아니라 "기준선보다 나아졌는가"가 판정입니다.

모델 고르기

제일 좋은 것 하나로 다 돌리면

- 분류 · 태깅 · 대량 반복 → 경량
- 대부분의 일상 업무 → 중급
- 검수 · 최종 판정 · 복잡한 판단 → 상급
- 두 개로 시작하세요 — 일상용 하나, 검수용 하나

5배

경량으로 될 일에 내는 값

상급 · 중급 · 경량은 성능과 가격이 함께 움직입니다

실습 · 30분

컨텍스트를 일부러 넘겨보기

1

표식 심기

대화 맨 앞에 규칙 둘 — 형식(이니셜)
과 사실(코드명)

2

길게 채우기

긴 글을 5~8회 나눠 붙이며
정리시킵니다

3

되묻기

규칙을 다시 알려주지
말고 물어봅니다

4

되살리기

되묻기 · 제약 다시 적기 · 요약 후 새 대화
— 무엇이 듣는지

가장 위험한 결과는 "모르겠습니다"가 아니라 **다른 코드명을 자신 있게 말하는 것**입니다. 모델은 잊었다고 말하지 않고 채웁니다.

정리

오늘 가져가실 것

- 1 모델은 다음 말을 맞히는 기계 — 그래서 잘하는 것과 못 하는 것이 예측됩니다
- 2 환각은 구조에서 나옵니다. 자료를 주고 그 안에서 답하게 하세요
- 3 사실은 찾아 붙이고, 말투와 형식만 학습으로 고정합니다
- 4 실행은 시스템이 합니다 — 그 사이가 승인을 넣을 자리입니다
- 5 창을 닫으면 사라집니다. 남길 것은 대화 밖 파일에
- 6 조직에서는 판정도 자동화해야 합니다 (Evals)
- 7 대량은 싼 모델로, 검수만 비싼 모델로