3단계

내 컴퓨터에서 동작하는 생성형 AI 설치하기

찾아가는 AI 멘토링 | 생성형 AI 심화 · 소주제 16

활용 도구: Ollama

경희대학교 DX추진단 · 정보처

목차

① Ollama 소개

AI가 내 컴퓨터 안에서 직접 동작 — 인터넷 없이 사용, 입력 내용이 외부로 전혀 나가지 않음

② 맞춤 실습

기밀 문서 처리 · 비용 절감 · 오프라인 활용 등 로컬 AI의 장점을 직접 체험

① 로컬 AI — 왜 필요한가?

Ollama를 사용하면 AI가 내 컴퓨터 안에서 직접 동작하므로, 인터넷 없이도 사용할 수 있고 데이터가 외부로 전혀 나가지 않습니다.

😩 클라우드 AI 서비스의 한계

- 입력한 내용이 외부 서버로 전송됨
- 기밀 문서·개인정보 입력 시 유출 우려
- 유료 구독 비용이 월 2~3만 원 이상 발생
- 인터넷이 없으면 사용 불가
- → 보안·비용·환경 제약이 존재

✅ 로컬 AI(Ollama)로 해결

- 모든 처리가 내 컴퓨터 안에서만 이루어짐
- 기밀 문서·개인정보가 외부로 전혀 나가지 않음
- 완전 무료 — 유료 구독 없이 무제한 사용
- 인터넷 없이도 동작 — 오프라인 환경 지원
- → 보안 · 비용 · 오프라인 문제를 동시 해결

💡 핵심: 데이터가 내 컴퓨터 밖으로 절대 나가지 않는다 — 기밀 문서 처리에 가장 안전한 방법

① Ollama 소개 — 내 컴퓨터에서 AI 실행하기

핵심 특징

- 오픈소스 AI 모델을 내 컴퓨터에 설치·실행하는 도구
- Windows · Mac · Linux 모두 지원
- 설치 후 명령어 한 줄이면 AI 모델 다운로드·실행
- 인터넷 없이 완전 오프라인 동작 가능
- 완전 무료 — 모델 다운로드·사용 모두 무료
- 다양한 오픈소스 모델 중 원하는 것을 선택 설치

설치 가능 AI 모델 ('26.4 기준)

Llama 3.1 (Meta)	범용 대화·분석, 고성능 오픈소스 대표 모델
Gemma 4 (Google)	가벼운 크기, 빠른 속도, 효율적 처리
Mistral / Mixtral	유럽 오픈소스, 코딩·분석에 강점
Phi-4 (Microsoft)	초경량 모델, 저사양 PC에서도 실행
Qwen 3.5 (Alibaba)	다국어 지원, 한국어 성능 우수

💡 Tip: 일반 노트북(RAM 16GB)이면 Llama 3.1 8B, Gemma 4 9B 등을 충분히 실행할 수 있습니다.

① Ollama 작동 원리 — 설치부터 실행까지

3단계로 내 컴퓨터에 AI를 설치하고 바로 사용할 수 있습니다.

Step 1
Ollama 설치

ollama.com에서
운영체제(Windows/Mac)
에 맞는 설치 파일 다운로드
→ 더블 클릭으로 설치 완료

→

Step 2
모델 다운로드

터미널(명령 프롬프트)에서
ollama pull llama3.1
명령어 한 줄 입력
→ AI 모델 자동 다운로드

→

Step 3
로컬에서 실행

ollama run llama3.1
명령어 입력하면
AI 대화 즉시 시작
→ 인터넷 없이도 동작

핵심 포인트: 설치 후에는 인터넷을 끊어도 AI가 동작합니다.
내가 입력한 모든 내용은 내 컴퓨터의 CPU/GPU에서만 처리되고, 외부 서버로 전송되지 않습니다.

① 활용 시나리오 — 기밀 문서 처리

왜 기밀 문서에 로컬 AI인가?

- 인사 평가·급여 자료를 외부 유출 걱정 없이 분석
- 감사 자료·재무 보고서를 AI로 요약·검토
- 개인정보가 포함된 문서를 안전하게 처리
- 연구 데이터·특허 자료를 보안 환경에서 분석
- 입력 내용이 100% 내 컴퓨터 안에서만 처리됨

부서별 활용 예시

- **인사처**
  → '이 인사평가 자료를 부서별로 요약하고 핵심 성과 지표를 표로 정리해줘'
- **감사실**
  → '이 감사 보고서에서 지적 사항을 유형별로 분류하고 중요도를 매겨줘'
- **재무처**
  → '이 재무제표를 분석해서 전년 대비 변동이 큰 항목을 10개 추출해줘'
- **연구처**
  → '이 연구 데이터를 요약하고 통계적 이상치가 있는지 확인해줘'

💡 Tip: 클라우드 AI에 절대 입력하면 안 되는 자료도, Ollama라면 안심하고 AI 분석이 가능합니다.

① 활용 시나리오 — 비용 절감 · 오프라인 사용

비용 절감

- 유료 구독 없이 무제한 사용 — 완전 무료
- ChatGPT Plus 월 $20, Claude Pro 월 $20 → 절약
- 부서 전체가 사용해도 추가 비용 없음
- API 호출 비용이 발생하지 않음
- 대량 문서 처리 시 비용 부담 제로

오프라인 · 특수 환경 활용

- 인터넷이 없는 환경에서도 AI 사용 가능
- 네트워크 차단 보안 구역에서 활용 가능
- 출장·이동 중 오프라인 상태에서도 작업 가능
- 서버 장애·네트워크 불안정에도 영향 없음
- 외부 서비스 의존도를 줄여 업무 연속성 확보

💡 비용 시뮬레이션: 직원 10명 × ChatGPT Plus $20/월 = 연간 약 320만 원 → Ollama라면 0원

① 클라우드 AI vs 로컬 AI(Ollama) 비교

구분	클라우드 AI	로컬 AI (Ollama)
데이터 보안	외부 서버로 전송됨	내 컴퓨터에서만 처리 ✓
비용	월 2~3만 원 구독료	완전 무료 ✓
인터넷 필요	필수 (항상 연결)	불필요 (오프라인 가능) ✓
응답 품질	★★★ (최고 수준)	★★~★★★ (모델에 따라)
응답 속도	빠름 (서버 GPU)	PC 사양에 따라 다름
추천 용도	일반 업무 · 고성능 분석	기밀 문서 · 보안 환경 · 비용 절감

💡 결론: 보안·비용이 중요하면 Ollama(로컬), 최고 성능이 필요하면 클라우드 — 상황에 맞게 병행 사용

② 맞춤 실습: 내 컴퓨터에서 AI 실행하기

실습 A
Ollama 설치 + 첫 실행

[실습 과정]
1. ollama.com 접속
2. 운영체제에 맞는 설치 파일 다운로드
3. 설치 완료 후 터미널(명령 프롬프트) 열기
4. ollama pull llama3.1 입력
   → 모델 다운로드 (약 4~5GB)
5. ollama run llama3.1 입력
   → AI 대화 시작
6. 간단한 질문으로 동작 확인

→ 내 컴퓨터에서 AI가 동작하는 것을 확인

실습 B
기밀 문서 AI 분석

[준비] 기밀 수준의 테스트 문서

[실습 과정]
1. 인터넷 연결을 끊기 (비행기 모드)
2. Ollama가 오프라인에서도 동작하는지 확인
3. 테스트 문서 내용을 AI에 입력:
   '이 내용을 3가지로 요약해줘'
4. '핵심 지적 사항을 표로 정리해줘'
5. 네트워크 모니터로 외부 전송 없음 확인

→ 오프라인에서 기밀 문서 처리를 체험

실습 C
외부 도구 연동 체험

[추가 활용]
Ollama를 다른 도구와 연동하면
더 편리하게 사용할 수 있습니다.

[연동 예시]
- Open WebUI: 웹 브라우저에서 ChatGPT처럼 대화 인터페이스 사용
- VS Code + Continue: 코드 편집기에서 AI 코딩 보조
- Dify(로컬): 로컬 RAG 챗봇 구축

→ 로컬 AI의 확장 가능성을 확인합니다

⏱️ 실습 시간: 15분 | 모델 다운로드에 시간이 걸리므로, 사전에 설치를 완료해두면 원활합니다

오늘의 핵심 정리

- Ollama를 사용하면 AI가 내 컴퓨터에서 직접 동작하여 입력 내용이 외부로 전혀 나가지 않는다
- 인사·감사·재무 등 기밀 문서를 외부 유출 걱정 없이 AI로 요약·분석할 수 있다
- 유료 구독 없이 완전 무료로 무제한 사용 — 인터넷 없는 오프라인 환경에서도 동작한다
- 보안이 중요한 업무는 Ollama(로컬), 최고 성능이 필요한 업무는 클라우드 — 상황에 맞게 병행한다