MSAP.ai

msap.ai

MSAP.ai AAS
업무 전문가를 통한 AI Agent 개발

업무 전문가가 말로 정의하고 작은 로컬 모델이 실행하는
AI Agent 의 설계·운영 기준



대상

IT 의사결정권자·IT 담당자

일자

2026.11

CONFIDENTIAL

CONTENTS

AI Agent 스킬 설계 다섯 주제

01 **업무 정의서와 그 산출물**
업무 정의가 먼저인 이유, 업무 정의서 양식, 스킬 산출물 4종

02 **개발·운영 환경 분리와 MCP 제작 절차**
폐쇄망 반입, LiteLLM 게이트웨이, 운영 모델 기준 시험

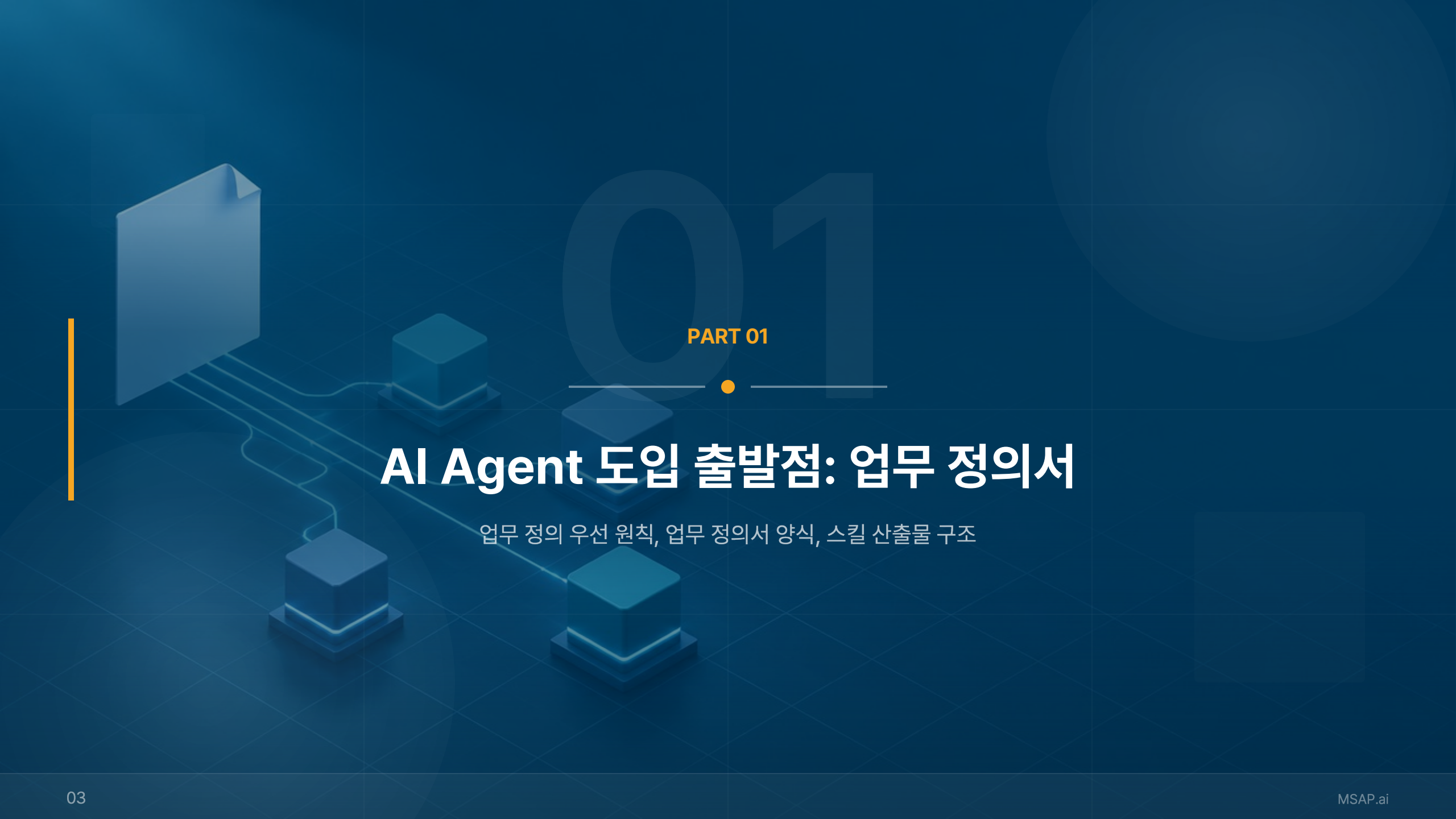
03 **작은 모델용 스킬 작성 지시와 GPU 설계**
지시문 해석·결정적 코드·질문 예시 세 지시, MoE 모델, 훅

04 **스킬 작성 사례와 합격 기준**
사례 네 건, 합격선, 결과 검증 순환, 회귀 시험 기반 변경 관리

05 **남은 한계와 파일럿 도입 순서**
작은 모델이 약한 업무, 스킬 자가 수정 위험, 파일럿 선정 점검표

KEY FOCUS 업무 정의서 한 장에서 스킬 제작·시험·반입·합격 판정까지 이어지는 설계 기준

CONFIDENTIAL



PART 01

AI Agent 도입 출발점: 업무 정의서

업무 정의 우선 원칙, 업무 정의서 양식, 스킬 산출물 구조

AI Agent 도입은 업무 정의가 먼저

KEY INSIGHT 모델부터 고르면 시범 업무 탐색·프롬프트 보강·결과 재확인·원인 추적 비용이 뒤따름 — GPU 규모는 업무 정의서로 시험한 뒤 정해도 늦지 않음

항목	모델·GPU 먼저	업무 정의 먼저
첫 의사결정	모델과 GPU 수량	맡길 반복 업무와 정답 기준
모델 크기 근거	공개 벤치마크·다른 조직 구성	정의서의 처리 규칙과 질문 예시
첫 성과 측정	장비 설치 뒤 시범 업무를 찾은 다음	스킬 제작 직후 운영 모델 시험
실패 원인 추적	모델 탓인지 정의 탓인지 구분 어려움	실패한 질문 예시를 회귀 시험으로 재실행

모델부터 고를 때 생기는 비용 5단계

1. GPU 서버 구매·설치
2. 시범 업무 탐색
3. 프롬프트 보강 (입력 토큰 증가)
4. 결과 재확인
5. 원인 추적

2~5단계는 업무 정의를 앞에 두면 줄어드는 비용

근거: 디코드 단계 지연은 메모리 대역폭이 좌우 — 장비를 늘려도 모델 호출 횟수는 줄지 않음

Source: NVIDIA, Mastering LLM Techniques: Inference Optimization (2023)

워크플로와 에이전트의 차이

KEY INSIGHT 절차와 정답 기준이 정해진 반복 업무는 워크플로 쪽 설계가 적합함 — 같은 요청에 같은 결과가 나와야 업무 시스템에 오류가 남지 않음

구분	워크플로	에이전트
다음 단계 결정 주체	개발자가 작성한 코드	언어 모델
실행 경로	매번 같음	실행마다 달라질 수 있음
맞는 업무	정해진 반복 업무	탐색형 업무
모델 호출 횟수	미리 셀 수 있음	업무 난이도에 따라 늘어남

업무 성격별 설계

정해진 반복 업무

견적·영업기회 등록·정기 점검

설계

결정적 코드 중심

절차는 고정, 요청 표현은 다양

자연어 CRM 조회·메일 기반 등록

설계

지시문 해석 + 결정적 코드

탐색형 업무

장애 원인 조사

설계

에이전트 + 사람 검토

Anthropic 권고: 가장 단순한 해법부터 시작하고, 복잡도는 결과가 좋아질 때만 더함

Source: Anthropic, Building effective agents (2024)

작은 모델 성능은 업무 결과로 측정

KEY INSIGHT 우리 질문 예시를 정답으로 처리하는지가 운영 투입 기준 — 이 확인은 GPU 추가 구매 전에 끝낼 수 있음

주장과 근거 수준

에이전트 안 모델 호출은 소수 전문 작업의 반복
NVIDIA 입장 논문 — 주장

작은 모델이 에이전트 호출 대부분에 충분
NVIDIA 입장 논문 — 실험 증명 아님

스킬 안의 스크립트는 결정적 신뢰성을 줌
Anthropic Agent Skills 설명 — 벤더 공식 설명

판단을 코드로 옮긴 스킬을 작은 모델이 정답 실행
AAS 구성 운영 관찰 — 제품 동작

두 성능 기준의 차이

구분	벤치마크 점수 기준	업무 결과 기준
측정 대상	범용 문제 풀이 능력	우리 업무 질문 예시의 정답 일치
측정 문항	공개 시험 문항	업무 전문가가 정한 질문 예시
반복 실행	평가하지 않음	같은 질문에 같은 결과인지 확인
비용과 속도	별도로 측정	도구 호출 횟수와 처리 시간을 함께 기록
판단 결과	모델 간 순위	**운영 투입 여부**

질문 예시는 스킬을 만드는 재료이면서 운영 모델의 성능을 재는 시험 문항

Source: Belcak 외, NVIDIA Research, Small Language Models are the Future of Agentic AI (arXiv 2506.02153, 2025); Anthropic Agent Skills (2025)

스킬은 프론티어 AI가 만들고 로컬 모델이 실행

KEY INSIGHT 높은 추론 성능은 스킬을 만들 때 한 번만 필요함 —
매일 반복되는 실행은 판단이 코드에 있어 작은 모델로 충분함

개발 환경 (인터넷)

- 업무 전문가
- 프론티어 AI (Claude)
- MCP
- MSAP.ai AAS MCP 서버

스킬 업로드
운영 쪽 단방향

운영 환경 (폐쇄망)

- 업무 담당자
- AAS 콘솔 · Hermes 프로파일
- LiteLLM 게이트웨이
- vLLM · 활성 3B 급 MoE 모델

프론티어 AI
스킬 작성 · 만들거나 고칠 때만

작은 모델
명령 한 줄 실행 · 매 요청

MSAP.ai AAS
MCP 도구·시험 채팅·업로드 · 상시

AAS 콘솔이 정의하는 세 대상 에이전트·프로파일 = 누가 · 스킬 = 어떻게 · 예약 작업 = 언제

Source: MCP Specification 2025-06-18; Hermes Agent 문서; LiteLLM 문서

AAS 도입 검토 판단 질문

KEY INSIGHT

네 판단 모두 업무 정의서 한 장에서 출발함 —
GPU 충분 여부는 질문 예시 시험 결과로, 폐쇄망 유지 여부는 정의서의 입력값·연결 시스템 목록으로 정해짐

의사결정권자의 판단 질문	판단에 필요한 근거
1 보급형 GPU로 업무 결과를 낼 수 있는가	판단을 코드로 옮긴 설계의 처리 시간·정답 여부
2 프론티어 AI를 쓰면서 폐쇄망을 유지할 수 있는가	개발·운영 환경 분리, 스킬 업로드 경로
3 코드를 쓰지 않는 업무 전문가가 스킬 작성을 맡을 수 있는가	업무 정의서 양식, MCP 스킬 제작 절차
4 도입 뒤 남는 위험과 관리 주체는 누구인가	작은 모델이 약한 업무 유형, 운영 중 스킬 변경 통제

Source: AAS 업무 전문가 스킬 설계 백서 1장

업무 전문가와 프론티어 AI의 역할 경계

KEY INSIGHT 코딩 경험은 업무 전문가 선정 기준에서 빠짐 —
기준은 그 업무의 예외와 정답을 가장 많이 아는 사람인가임

업무 전문가

문장으로 정하는 것

- 판단 규칙
- 예외 처리
- 사람 확인 단계
- 정답 판정

산출물
업무 정의서와 질문 예시

프론티어 AI

정의서를 코드로 옮겨 작성하는 것

- 규칙 파일
- 결정적 스크립트
- 회귀 시험
- SKILL.md

산출물
스킬 디렉터리

IT 담당자
모델·접근 권한·업무 시스템 연결 같은 실행 기반을 준비

업무 정의서 8항목

1. 업무 목적
2. 요청 형태
3. 입력값
4. 처리 규칙
5. 예외
6. 확인 단계
7. 완료 상태
8. 질문 예시

Source: Anthropic, Building effective agents (2024); AAS 업무 전문가 스킬 설계 백서 2장

코드로 옮길 수 있는 업무 문장의 조건

KEY INSIGHT 세 요소 중 하나라도 빠지면 작은 모델이 매번 추측함 — 추측한 기준은 다음 요청에서 바뀌어 결과가 달라짐

입력값	무엇을 받아 처리하는가
처리 규칙	어떤 기준으로 고르고 정렬하는가
완료 상태	무엇이 되면 끝인가

업무	흐린 요청 문장	코드로 옮길 수 있는 문장
CRM 조회	"요즘 진행 중인 큰 건 좀 알려 줘"	영업 단계가 제안 또는 협상인 영업기회를 예상 금액 큰 순서로 10건, 고객사 이름이 있으면 그 고객사로 좁힘 추측으로 남던 것: 진행 중 범위 · 큰 건 기준 · 건수
영업기회 등록	"이 메일 보고 영업기회 하나 만들어 줘"	첫 줄이 영업기회면 수요기관명·사업명·마감일을 읽고, 없으면 묻고, 드라이런 확인 뒤 등록
보안 공지 반영	"보안 공지 보고 알아서 조치해 줘"	서버 목록의 제품·버전에 해당하는 항목만 조치 계획 초안, 담당자 확인 뒤 반영

문장마다 함께 적는 예외 점검 7항목

- 값이 없을 때
- 후보가 여럿일 때
- 표기가 다를 때
- 이미 처리됐을 때
- 업무 시스템이 응답하지 않을 때
- 쓰기 전 사람 확인
- 완료 상태는 업무 시스템의 상태 값으로 기재

Source: AAS 업무 전문가 스킬 설계 백서 2장

업무 정의서 8항목과 스킬 산출물 대응

KEY INSIGHT 8항목 중 6항목이 결정적 코드로 바뀜 —
스크립트로 옮긴 판단은 운영 모델이 바뀌어도 같은 결과를 냄

업무 정의서 항목	바뀌는 산출물	이어지는 지시	산출물 파일
요청 형태	지시문 해석 규칙표	① 지시문 해석	규칙 파일
입력값	값 추출 스크립트	② 결정적 코드	결정적 스크립트
처리 규칙	진행 스크립트	② 결정적 코드	결정적 스크립트
예외	정해진 질문 문구, 후보 선택 규칙	① ②	규칙 파일·스크립트
확인 단계	확인 게이트 코드	② 결정적 코드	결정적 스크립트
완료 상태	재조회 검증 코드	② 결정적 코드	결정적 스크립트
질문 예시	회귀 시험 묶음 (정의서에서 가장 긴 항목)	③ 질문 예시	회귀 시험
금지 사항	SKILL.md 금지 항목	② 결정적 코드	SKILL.md

규칙표·스크립트로 바뀌는 6항목

업무 목적은 SKILL.md 설명 문장으로 들어감

Source: Anthropic Agent Skills (2025); AAS 업무 전문가 스킬 설계 백서 2장

CRM 조회 업무 정의서 작성 예

KEY INSIGHT 업무 전문가가 정한 판단은 첫 줄 규칙·후보 되묻기 문구·브라우저 금지 세 가지 — 프론티어 AI가 이를 규칙표·이름 색인·조회 스크립트·회귀 시험으로 옮김

업무 정의서 — CRM 영업 정보 조회

업무 목적	고객사·영업기회·견적·계약 현황 질문에 같은 답
요청 형태	첫 줄이 "조회"면 이 업무로 처리
입력값	고객사 이름(정식·약칭·영문) · 영업 단계 · 기간
처리 규칙	조회 의도 7종 분류 → 이름 색인 → 의도별 항목 표
예외	후보 둘 이상이면 번호 선택, 못 찾으면 정해진 질문 후 정지
확인 단계	없음 (조회만 하고 쓰지 않음)
완료 상태	답에 나온 건수 = CRM 조회 결과 건수
금지 사항	CRM 쓰기 금지, 브라우저로 CRM 화면 열기 금지
질문 예시	(1) "조회 / A사 진행 중인 영업기회" → 영업기회 목록 (2) "A사 견적 몇 건 나갔어?" → 견적 현황 (3) "에이사 담당자 누구야" → 고객사 정보 (약칭 일치) (4) "올해 수주한 거 담당자별로" → 담당자별 현황 (5) "조회 / B" → 후보 둘 이상, 번호 선택 질문 업무 전문가가 수십 건까지 늘림

업무 전문가가 정한 판단 → 옮겨진 곳

① 첫 줄 규칙

첫 줄이 "조회"면 이 업무

→ **지시문 해석 규칙표**

운영 모델이 스킬을 고르는 부담을 덜어 줌

② 후보 선택·되묻기 문구

번호 목록, 정해진 질문 후 정지

→ **이름 색인 · 코드가 문구 그대로 출력**

모델이 질문 문구를 지어내지 않음

③ 브라우저 금지

브라우저로 CRM 화면을 열지 않음

→ **SKILL.md 금지 항목**

답을 못 찾을 때 브라우저 도구를 반복 호출하던 행동 차단

Source: AAS 업무 전문가 스킬 설계 백서 2장 (고객사 이름·금액은 일반 표현으로 대체)

스킬 산출물 4종과 점진적 공개 3단계

KEY INSIGHT SKILL.md 본문은 명령 한 줄과 금지 사항만 담아 60줄 이내로 유지함 — 작은 모델은 긴 지시를 끝까지 일관되게 따르기 어려움

규칙 파일

첫 줄 판별·조회 의도·
질문 문구·별칭

스크립트가 읽음

결정적 스크립트

값 추출부터
재조회 검증까지

명령 한 줄로 실행

회귀 시험

질문 예시별 기대 결과

변경마다 재실행

SKILL.md

이름·설명·명령 한 줄·
금지 사항, 60줄 이내

운영 모델이 읽음

점진적 공개 (Progressive Disclosure)

1단계 · 모든 스킬 이름·설명
대화 시작 시 늘, 약 3천 토큰

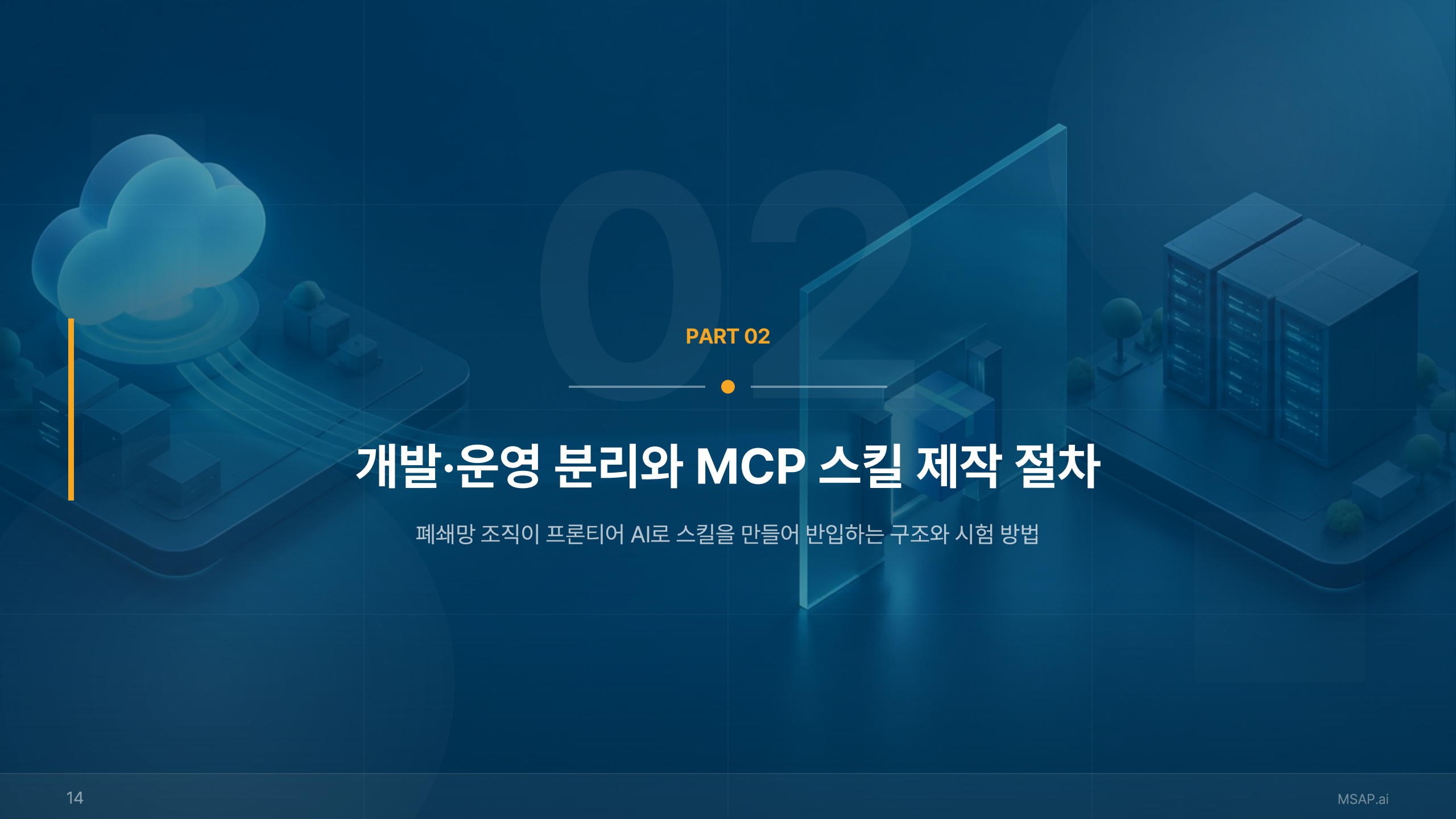
2단계 · 고른 스킬의 SKILL.md 본문
스킬을 고른 뒤에만

3단계 · 참고 파일·규칙 파일
필요할 때만 열거나 스크립트가 읽음

1단계 설명이 스킬 선택을 좌우

운영 모델은 이름과 설명만 보고 스킬을 고름 —
첫 줄 판별 기준을 설명에 직접 기재

Source: Anthropic Agent Skills (2025); Hermes Agent 문서 — Skills



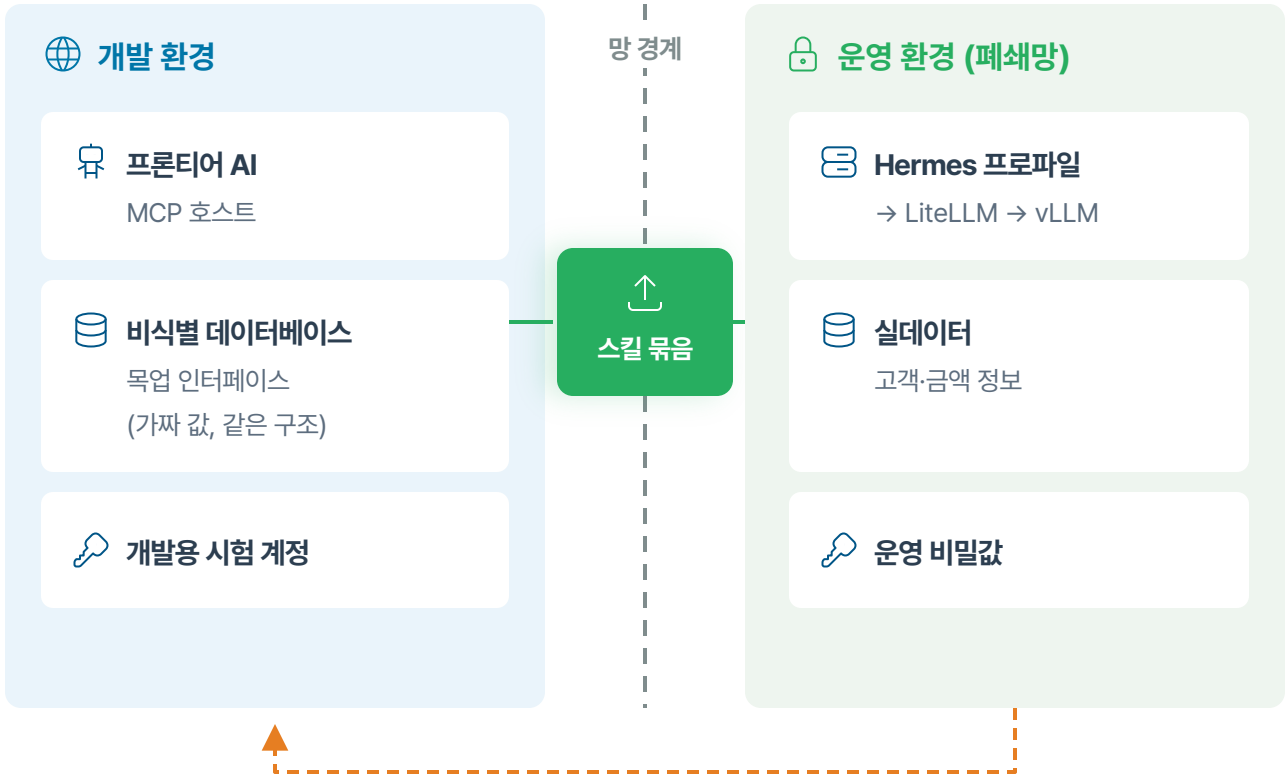
PART 02

개발·운영 분리와 MCP 스킬 제작 절차

폐쇄망 조직이 프론티어 AI로 스킬을 만들어 반입하는 구조와 시험 방법

개발 환경과 폐쇄망 운영 환경의 경계

KEY INSIGHT 프론티어 AI와 실데이터가 같은 영역에 놓이지 않음 —
기존 자료 반입 절차를 그대로 쓰므로 망 분리 정책 변경이 필요 없음



경계를 넘는 것 — 스킬 파일 묶음

- SKILL.md
- 규칙 파일
- 결정적 스크립트·플러그인 훅
- 회귀 시험 파일

경계를 넘지 않는 것

- 개발 환경의 비식별 시험 자료
- 운영 비밀값
- 실데이터와 모델 호출 기록

Source: MCP Specification 2025-06-18; LiteLLM 문서; AAS 업무 전문가 스킬 설계 백서 3장

디지털 트윈 개발 환경 구성

KEY INSIGHT 예외 행과 오류 응답까지 운영과 같게 맞춰야 개발 환경 합격이 운영에서 유지됨 — 날짜·금액 표기, 한글 처리, 건수 상한은 오류 메시지 없이 틀림

비식별 데이터베이스

구조는 복제, 값은 임의 생성
예외 행은 의도적으로 추가

실데이터 없음

목업 인터페이스

호출 규약·정상 응답·오류 응답을
운영 명세대로 재현

실데이터 없음

개발용 AAS·Hermes Agent

운영 환경과 같은 판을
컨테이너로 실행

실데이터 없음

개발용 실행 모델

운영과 같은 범주의
활성 3B 급 MoE 모델

실데이터 없음

준비 항목	운영과 맞출 내용	어긋나면 운영 환경에서 나타나는 증상
필드 이름	대소문자·밑줄·한글 필드 이름까지 동일	스크립트가 값을 찾지 못해 빈 결과 반환
응답 형식	목록 감싸는 구조, 빈 값 표기 방식	결과가 있는데도 "없음"으로 판정
오류 응답	권한 없음·대상 없음·중복의 응답 코드와 메시지	오류를 정상으로 보고 다음 단계로 진행
날짜·금액 표기	날짜 형식, 천 단위 구분, 부가세 포함 여부	금액 계산과 비교 결과가 틀림
한글 처리	한글 이름 검색의 공백·괄호·약칭 처리	같은 고객을 다른 고객으로 인식
건수 상한	페이지 나눔 방식과 한 번에 돌려주는 최대 건수	일부 결과만 보고 "대상 없음"으로 판단
인증 방식	값은 다르되 인증 절차와 만료 응답은 동일	반입 직후 모든 호출이 인증 오류

Source: AAS 업무 전문가 스킬 설계 백서 3장; 주황 행은 운영 응답 표본과 글자 단위 대조 권장

완성 스킬의 폐쇄망 반입 6단계

KEY INSIGHT 반입 승인서 해시와 운영 스킬 해시를 대조하면 승인 없는 변경을 찾을 수 있음 — 파일럿은 스킬 업로드, 두 번째 부서부터 스킬 허브가 적합함

1	2	3	4	5	6
개발 환경 시험	**묶음 확정**	**반입 승인**	**업로드·게시**	**설치 보안 검사**	**운영 재시험**
업무 전문가	IT 담당자	보안 담당자	IT 담당자	시스템 (스킬 허브)	업무 전문가
시험 결과	**판 번호·파일 해시**	**반입 승인서**	업로드 기록	검사 결과	같은 회귀 시험 결과

반입 승인서의 해시 ↔ 운영 환경 스킬 파일의 해시 대조 — 어긋나면 승인 없는 변경

반입 대상 유형별 확인

유형	반입 때 확인할 것
스킬을 부르는 스킬	불리는 스킬이 같은 에이전트에 함께 켜져 있는가
큰 스킬	참고 파일 경로가 스킬 디렉터리 안의 상대 위치인가
한글 이름 스킬	이름 글자 깨짐 여부와 해시 일치 확인
첫 줄 판별 규칙	첫 줄 낱말 목록이 규칙 파일에 들어 있는가

스킬 업로드 — 파일럿

콘솔에서 묶음을 올려 특정 에이전트에 바로 켬, 한 부서 한 에이전트에 빠름

스킬 허브 — 두 번째 부서부터 사용

설치마다 보안 검사, 어느 에이전트가 어느 판을 설치했는지 저장소에 남음

Source: Hermes Agent 문서 — Skills Hub; AAS 업무 전문가 스킬 설계 백서 3장

LiteLLM 호출 경로와 보안 점검 9항목

KEY INSIGHT 게이트웨이 모델 목록에 외부 제공자가 없다는 설정 자체가 증빙이 됨 — 보안 담당자는 아홉 항목으로 파일럿 승인 여부를 판단함

1 요청 접수
AAS 콘솔·사내 채널
업무 담당자의 채팅 요청

→

2 스킬 실행
Hermes 프로파일
스킬 선택·스크립트 실행

→

3 호출 통제
LiteLLM 게이트웨이
가상 키·예산·속도 상한·호출 기록·장애 우회

→

4 생성
vLLM 로컬 모델
요청당 도구 호출 0~1회

→

5 읽기·쓰기
업무 시스템
쓰기는 사람 확인 뒤에만

→

6 응답
AAS 콘솔·채널
스크립트 결과 원문 전달

보안 담당자의 파일럿 승인 전 점검 9항목

1 개발 환경 실데이터 없음
비식별 DB·목업 값 표본 검토

2 반입 묶음 해시 일치
운영 스킬 해시 ↔ 반입 승인서 해시

3 설치 보안 검사
스킬 허브 설치 검사 결과 확인

4 게이트웨이 단일 경로
LiteLLM 모델 목록에 외부 제공자 없음

5 에이전트별 가상 키·예산
키·예산·속도 상한 설정 확인

6 호출 기록 보존
관측 도구에 기록이 쌓이는지 확인

7 쓰기 전 사람 확인
확인 게이트·도구 호출 동의 설정

8 운영 중 스킬 수정 차단
skills.write_approval·차단 훅 확인

9 비밀값은 스킬 파일 밖
묶음 안 계정·키 문자열 검사

Source: LiteLLM 문서; vLLM 문서; MCP Specification 2025-06-18; Hermes Agent 문서

MSAP.ai

MCP 연결 구조와 AAS 도구 5묶음

KEY INSIGHT 업무 전문가는 대화창 하나에서 제작부터 시험까지 진행함 —
MCP 표준이라 프론티어 AI 애플리케이션을 바꿔도 AAS 쪽 설정은 그대로임

호스트
프론티어 AI 애플리케이션
도구 호출 전 사용자 동의

→

클라이언트
AAS 연결 설정
연결 대상·인증

→

서버
MSAP.ai AAS MCP 서버
노출 범위·권한 검사

통신 규약: JSON-RPC 2.0

AAS 도구 묶음	업무 전문가가 하는 일	통제
에이전트 목록	시험 대상 에이전트 고르기	조회 전용
스킬 관리	생성·폴더 업로드·적용·켜기	스크립트 포함 시 명시 확인
채팅	세션 시작·보내기·기다리기·승인 응답	승인 응답은 한 번·세션·항상·거부 중 선택
예약 작업	같은 질문 예시를 주기적으로 다시 실행	프로파일 단위 cron 일정
비밀값 관리	계정·키 등록	값은 스킬 파일 밖에 보관

런타임 설정과 외부 시스템 연결은 이 도구로 바뀌지 않음 — IT 담당자 몫

Source: MCP Specification 2025-06-18; AAS 업무 전문가 스킬 설계 백서 4장

스킬 제작 4단위와 등록 4단계

KEY INSIGHT 단위마다 다음으로 넘어가는 조건을 정해 둠 —
SKILL.md를 먼저 길게 쓰거나 다른 프로파일에 올리는 실수를 막음

프론티어 AI와의 대화 4단위 — 단위별 넘어가는 조건

① 업무 정의서 제시
AI가 판단 목록을 정리

넘어가는 조건
판단 목록 승인

② 규칙·스크립트·시험 생성
판단을 코드로 먼저 옮김

넘어가는 조건
회귀 시험 전부 통과

③ SKILL.md 확정
코드가 생긴 뒤 마지막에 작성

넘어가는 조건
명령·금지 사항만 남음

④ 등록·활성화
MCP 도구로 AAS에 등록

넘어가는 조건
시험 프로파일에서 켜짐

④ 단위 안의 등록 4단계

1. **업로드** — 검사 통과·식별값 받음
2. **적용** — 스크립트 실행 허용 확인
3. **활성화** — 옛 동명 스킬 충돌 확인
4. **반영 확인** — 등록 파일과 원본 일치

조건 없이 진행할 때 생기는 실수
SKILL.md를 코드보다 먼저 길게 작성 · 시험 대상이 아닌 다른 프로파일에 등록 · 옛 동명 스킬과 충돌

Source: AAS 업무 전문가 스킬 설계 백서 4장

합격 판정은 운영 모델 기준

KEY INSIGHT 프론티어 AI 실행 결과는 정답표로만 씀 —
판단 이관 전 운영 모델은 같은 스킬에서 턴당 도구 11~15회, 191~207초가 걸림

구분	프론티어 AI 직접 실행	운영 모델 (AAS 채팅)
판정 역할	정답표	**합격 판정 대상**
긴 SKILL.md	맥락으로 빈 곳을 메움	단계 누락·순서 뒤바뀜
확인 게이트	스스로 지킴	코드가 쥐지 않으면 건너뜀
출력 전달	원문 유지	자기 말로 고쳐 씀

시험 기록 항목

질문 원문 · 기대 결과 · 운영 모델 결과 · 상태 이력 · 실패 유형

도구 호출 수 3회 이하 · 소요 시간 몇 분 안

개선 전 운영 모델: 턴당 도구 11~15회, 191~207초

실패 유형별 수정

지시문 해석 실패

엉뚱한 스킬을 고르거나 추측으로 진행

→ 첫 줄 규칙표·질문 문구 보강

결정적 코드 부족

모델이 단계·확인 게이트를 판단

→ 그 판단을 스크립트로 이관

질문 예시 부족

새 표현에서 결과가 달라짐

→ 예시 추가·회귀 시험으로 고정

수정 뒤 회귀 시험 전체 재실행

Source: AAS 구성 운영 관찰 (한 업무 에이전트, 외부 벤치마크 아님); AAS 업무 전문가 스킬 설계 백서 4장



PART 03

작은 모델용 스킬 작성 지시와 GPU 설계

지시문 해석·결정적 코드·질문 예시 세 지시와 MoE 모델·훅 구성

지시 ① 지시문 해석

KEY INSIGHT 명령 선택과 되묻기를 코드 규칙표로 둠 — 모델이 추측으로 엉뚱한 업무를 시작하는 실패가 사라짐

요청 첫 줄로 명령 고르기

요청 예시 (첫 줄 / 본문 일부)	해석 근거	선택 명령
"견적 / ○○공사 모니터링 구축 건, 라이선스 20식"	첫 줄 명령 낱말	**견적 작성**
"영업기회 / 첨부 공고 기준 등록, 지난 견적 번호 참고"	첫 줄 명령 낱말 (본문의 "견적"은 판별에서 제외)	**영업기회 등록**
"서버 1에 OS 취약점 조치됐어?"	신호어 "취약점" 단독 일치	**패치 적용 조회**
"지난달 공지된 CVE는 몇 대에 적용됐어?"	신호어 "CVE" 단독 일치	**패치 적용 조회**
"이거 처리해 줘"	일치하는 명령 없음	**되묻기 후 정지**

코드가 그대로 내는 되묻기 문구

상황	출력할 질문	답 형식
명령 판별 불가	"어떤 업무인가요? 1) 견적 2) 영업기회 등록 3) 패치 조회"	번호
고객사 후보 여러 곳	"고객사가 두 곳과 일치합니다. 1) ○○공사 2) ○○공단"	번호
필수 값 누락	"사업명이 없습니다. 사업명을 한 줄로 적어 주세요."	한 줄 문장
대상 서버 누락	"대상 서버 이름을 적어 주세요. 예: web-01"	서버 이름

멈춘 뒤 이어서 실행

사용자 답(번호·문자열)을 스크립트가 그대로 받아 다음 단계 실행, 상태 파일에 단계 기록

Source: AAS 업무 전문가 스킬 설계 백서 5장

지시 ② 결정적 코드

KEY INSIGHT 여섯 판단을 진행기 스크립트가 맡음 —
모델 몫은 명령 한 줄 실행과 출력 원문 전달로 줄어 같은 요청에 같은 결과가 나옴

스크립트로 옮기는 여섯 판단

값 추출
라벨·정규식·이름 색인으로
요청에서 값을 읽음

후보 선정
대조 점수로 고르고
동점이면 되묻기

순서
모델이 매 턴 고르지 않는
고정 단계

확인 게이트
드라이런 결과를 보여 주고
승인 문구가 있어야 진행

멱등 쓰기
같은 키면 새로 만들지 않음
— 두 번 받아도 한 건

재조회 검증
쓴 뒤 다시 읽어 대조,
마지막 단계에 반드시 둠

진행기 도입 전후

구분	도입 전	도입 후
명령 선택	모델이 SKILL.md를 읽고 판단	진행기 해석 단계
단계 순서	모델이 매 턴 고름	고정 순서
실행	여러 도구 차례 호출	진행기 명령 한 줄
사용자 응답	요약·재작성	**스크립트 출력 원문**

가장 먼저 옮길 것: 확인 게이트와 쓰기

Source: Anthropic Agent Skills (2025); AAS 업무 전문가 스킬 설계 백서 5장

지시 ③ 질문 예시와 회귀 시험

KEY INSIGHT 모델 교체·설정 변경·규칙 변경·파일 변경 때마다 같은 묶음을 다시 돌림 — 결과가 달라진 곳을 기계 대조로 바로 찾음

질문 예시에 담는 여섯 가지 꼴

1. 표준 꼴
2. 줄임·구어 꼴
3. 다른 업무 낱말이 섞인 꼴
4. 값이 빠진 꼴
5. 후보가 여럿인 꼴
6. 범위 밖 요청

한 의도, 여러 표현 → 모두 "패치 적용 조회"

"서버 1에 OS 취약점 조치됐어?"

"web-01 이번 달 보안 패치 적용 여부"

"웹 1호기 패치 했어?"

회귀 시험 한 건의 필드

입력	기대 명령	기대 값
되묻기 여부	쓰기 실행 여부	기대 출력

전체 묶음을 다시 돌리는 네 시점

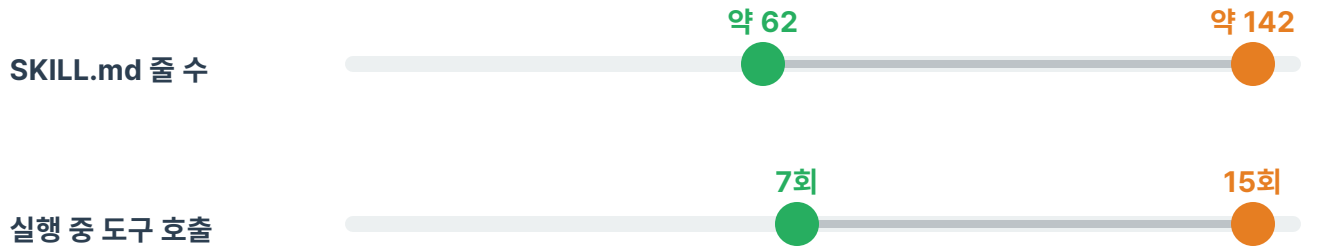
- 운영 모델 교체
- 서빙·추론 설정 변경
- 업무 규칙 변경
- 스킬 파일 변경 감지

Source: Anthropic, Building effective agents (2024); AAS 업무 전문가 스킬 설계 백서 5장

결정적 코드 한 줄 지시의 한계

KEY INSIGHT 줄 수와 도구 호출은 절반으로 줄었으나 확인 게이트와 후보 고르기는 모델에 남음 — 이관할 판단을 이름으로 지정해야 승인 없는 쓰기를 막을 수 있음

"결정적 코드 중심으로 다시 작성" 한 줄만 준 재작성 결과



확인 게이트	모델 → 모델 (변화 없음)
후보 고르기	모델 → 모델 (변화 없음)

재작성한 작은 모델은 조회를 묶어 줄였지만 결과를 바꾸는 판단은 못 찾음,
프론티어 모델도 한 줄 지시만으로는 같은 선택을 할 수 있음

→ 여섯 판단을 이름으로 적어 이관 범위를 명시

스킬 받기 전 검토 두 가지

판단 목록의 코드 이관 여부

SKILL.md의 조건 문장과
진행기 단계를 하나씩 짝지음

합격: 짝 없는 판단 0건

업무 전문가

확인 게이트의 코드 소유

승인 문구 없이 쓰기 명령을 부르면
진행기가 거부하는지 시험

합격: 승인 없는 쓰기 거부

IT 담당자

Source: AAS 구성 운영 관찰 (한 업무 에이전트, 근사치); Anthropic, Building effective agents (2024)

토큰 생성 속도는 메모리 대역폭이 결정

KEY INSIGHT 활성 파라미터가 작은 MoE 모델은 토큰마다 읽는 가중치가 적어 같은 장비에서 더 빠름 — 고가 장비 대신 모델 구조로 속도를 확보함

추론의 두 단계

프리필 (Prefill)

입력 전체를 한 번에 처리

좌우하는 것: GPU 연산 성능·입력 토큰 수

디코드 (Decode)

토큰을 하나씩 생성

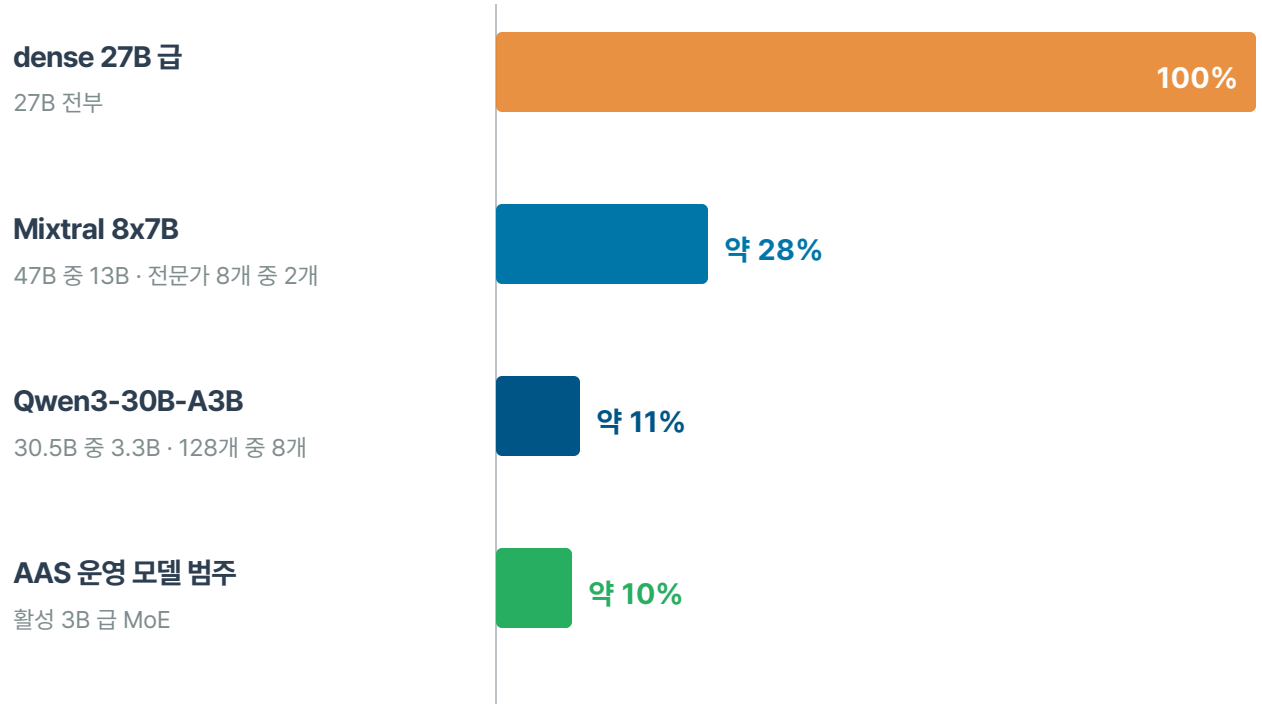
좌우하는 것: 메모리 대역폭

도구 호출마다 두 단계 반복

도구 결과가 다음 입력에 쌓여 프리필이 길어지고
디코드도 한 차례 더 돎

→ 호출 횟수를 줄이는 설계가 먼저

토큰마다 읽는 가중치 비율 (활성 파라미터 / 전체)



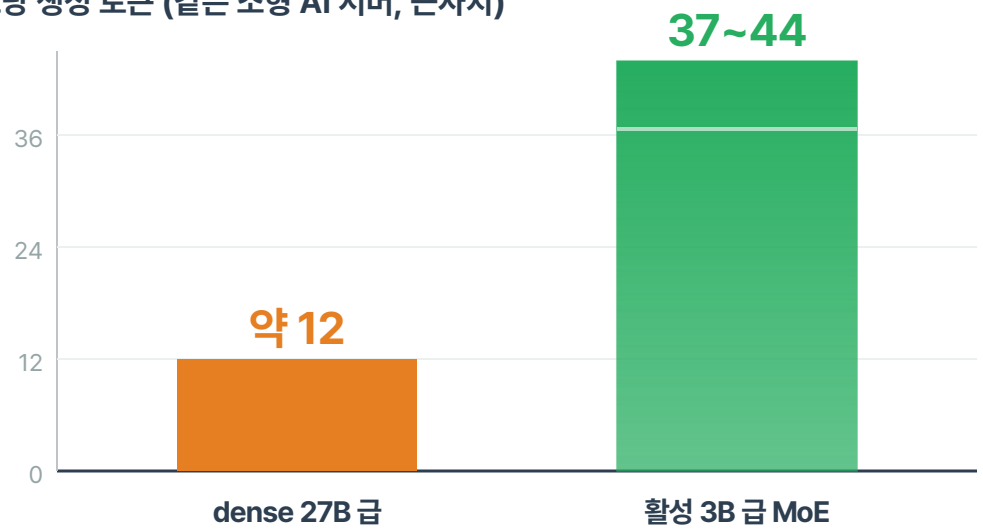
읽는 가중치가 적을수록 같은 대역폭에서 초당 토큰이 늘어남

Source: NVIDIA, Inference Optimization (2023); Jiang 외, Mixtral of Experts (arXiv 2401.04088); Qwen3-30B-A3B 모델 카드

dense 27B와 활성 3B MoE의 속도 비교

KEY INSIGHT 초당 생성 토큰이 약 12에서 37~44로 늘어남 — 단 속도만으로는 부족하고 턴당 도구 호출을 1회 안팎으로 유지해야 응답 시간이 줄어듦

초당 생성 토큰 (같은 소형 AI 서버, 근사치)



생성 분량	dense 27B 급	활성 3B 급 MoE
도구 인자 100토큰	약 8초	**2~3초**
답변 300토큰	약 25초	**7~8초**

AAS 서빙 구성 5요소

vLLM OpenAI 호환 서버
로컬 MoE 모델 서빙

병렬 도구 호출 끄기
parallel_tool_calls=false — 요청당 도구 호출 0~1회

생각 모드 끄기
판단이 코드에 있어 긴 추론 토큰이 필요 없음

LiteLLM 게이트웨이
호출 통제·기록 단일 경로

Hermes Agent 프로파일
스킬·훅을 실행하는 에이전트 단위

Source: vLLM 문서; Qwen3 모델 카드; AAS 구성 운영 관찰 (근사치, 외부 벤치마크 아님)

결정적 코드 중심의 역할 분담

KEY INSIGHT 스킬 고르기에 판단 모델을 더해도 정확도는 약 86%에서 89%로 오르고 시간은 20배 넘게 늘어남 — 판단 모델은 자유 답 해석과 결과 충족 판정에만 둠

결정적 코드		첫 줄 라우팅 · 값 추출 · 조회 · 후보 선정 · 순서 확인 게이트 · 멱등 쓰기 · 재조회 검증
로컬 LLM	활성 3B 급 MoE	요청 받기 · 진행기 명령 한 줄 · 결과 원문 전달 하지 않음: 업무 판단 · 결과 고쳐 쓰기 · 스킬 수정
판단 모델		사용자 자유 답 해석 · 결과 충족·실패 유형 판정 재시도 여부는 코드가 정함
사람		업무 규칙 정의 · 쓰기 전 최종 확인 1회 · 예외 처리

스킬 고르기에 판단 모델을 더하면

로컬 LLM 선택만
약 86% · 1초 미만

→

판단 모델 검토 추가
약 89% · 약 20초

Source: AAS 구성 운영 관찰 (한 업무 에이전트, 근사치); AAS 업무 전문가 스킬 설계 백서 6장

코드 이관 뒤 남는 공백 세 곳과 훅

KEY INSIGHT

세 훅이 남은 공백을 모델 밖에서 막음 —
훅 적용 뒤 턴당 도구 호출은 1회 이하로, 답은 진행기 원문 그대로 유지됨

요청 한 턴과 훅의 자리



① 모델 호출 전 훅 · 라우팅

공백

본문의 다른 업무 낱말 때문에
엉뚱한 스킬 선택

막는 방법

첫 줄을 등록 명령과 대조해
진행기를 바로 실행

② 출력 직후 훅 · 전달 충실도

공백

모델이 질문·수치·확인 문구를
줄이거나 바꿔 전달

막는 방법

답을 진행기 원문으로 교체

③ 도구 호출 전 훅 · 스킬 수정

공백

대화 뒤 배경 검토가
SKILL.md·스크립트를 수정

막는 방법

스킬 디렉터리 쓰기 차단, 원본
매니페스트(파일별 해시)로 복원

스킬 변경 경로는 하나: 개발 환경 → 회귀 시험 → 재업로드

Source: Hermes Agent 문서 — Plugins·Hooks; AAS 구성 운영 관찰

PART 04

AAS 스킬 작성 사례와 합격 기준

영업·IT 운영 사례 네 건과 합격선·결과 검증·변경 관리 기준

CRM 조회 응답 394초에서 14초로 단축

KEY INSIGHT 이름 색인·조회 의도 7종·회귀 문항 53건을 코드로 고정해 틀릴 수 있는 단계를 없앰 — 속도와 정확도가 함께 좋아짐

응답 시간

394초 → 14초

판단을 코드로 옮긴 뒤

도구 호출

36회 → 0회

그중 브라우저 조작 26회

입력 토큰

약 127만 → 한 턴 분량

도구 결과가 입력에 쌓이지 않음

결과

오답 → 정답

같은 활성 3B 급 MoE 모델

업무 전문가가 정한 문장	코드로 옮긴 판단	모델에게 남은 일
기관 이름이 조금 달라도 같은 곳을 찾음	이름 색인 조회, 연도·날짜·말머리를 뗀 유사도 비교	없음
무엇을 묻는지는 질문 문장으로 판단	의도 규칙 7종 판별	없음
여러 곳이 맞으면 후보를 보여 주고 물음	후보 수 판정, 정해진 질문 문장 출력	질문 그대로 전달
답에 CRM 기록과 문서 링크를 붙임	결과 양식 생성, 링크 조회	결과 그대로 전달
새 표현을 알려 주면 다음부터 알아들음	별칭 목록 추가, 회귀 문항 53건으로 고정	없음

127만 토큰의 원인: 도구 호출마다 앞선 도구 결과가 다음 모델 입력으로 다시 들어감

Source: 한 업무 에이전트의 AAS 구성 운영 관찰 (외부 벤치마크 아님)

영업기회 등록 진행기 8단계

KEY INSIGHT SKILL.md 281줄은 52줄로, 턴당 도구 호출 11~15회는 1회로 줄어듦 — 작은 모델의 결과 JSON이 프론티어 모델 실행 결과와 일치함



거둬들인 판단: 이전 건의 답으로 처리돼 생긴 중복 등록, 답을 지어 넣어 수요기관 칸에 협력사 이름이 들어간 오류

항목	문서를 읽고 엮던 구성	진행기 + 작은 모델	같은 명령 + 프론티어 모델
SKILL.md 길이	281줄	**52줄**	52줄 (같은 스킬)
턴당 도구 호출	11~15회	**1회**	1회
응답 시간	191~207초	**18초 + 33초 (두 턴)**	11초 + 21초 (두 턴)
입력 토큰	36만~48.5만	**약 8.3만**	해당 없음
결과	두 번 중 한 번 인증 경로 오류로 실패	**결과 JSON 일치**	기준

Source: 한 업무 에이전트의 AAS 구성 운영 관찰 (외부 벤치마크 아님); 주황 단계는 사람에게 돌아가는 곳

견적 초안 판단 모델 판정 결과

KEY INSIGHT 판단 모델이 시범 12건의 맞음·틀림을 모두 정확히 판정함 — 단 실제 생성 여부는 볼 수 없어 코드 재조회를 앞에 둠

견적 한 건의 검증 순서

1. **코드** 요청 원문 값·거래처 정확 일치·금액 규칙
   어긋나면 경고 또는 차단, 사람에게 질문
2. **판단 모델** 첫 초안이 요청 원문·사용자 답과 맞는가
   어긋나면 초안에 경고 한 줄
3. **사람** 초안 검토와 수정 지시
   값을 고치면 1번부터 다시
4. **판단 모델** 최종 확인 직전 초안이 요청과 맞는가
   확인 메시지에 경고 한 줄
5. **사람** 생성 승인 1회
   승인 전에는 쓰지 않음
6. **코드** 견적번호 재조회·PDF 저장·링크 기록
   재시도 1회 또는 사람에게 넘김

시범 문항 판정 결과

운영 판단 모델

12/12	7/7	약 3초
맞음·틀림 정확 판정	어긋난 곳 찾음	판정 시간 중앙값 (첫 기동 약 9초)

충족 점수: 어긋난 초안 0.14 이하 · 맞는 초안 0.83 이상
맞는 초안과 어긋난 초안의 점수 간격이 넓음

비교: 4B 급 소형 판단 모델

7/12	1/7	약 0.9초	빠르지만 판정이 부정확

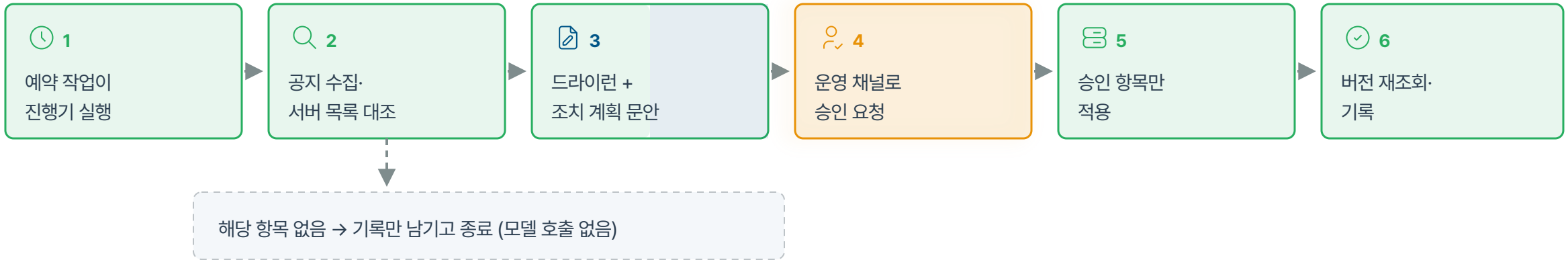
판단 모델은 결과 글만 봄 —
실제 생성 여부는 6단계 코드 재조회가 확인

Source: AAS 견적 업무 시범 문항 (12건, 범위 한정 · 외부 벤치마크 아님)

보안 공지 반영 운영 에이전트 설계안

KEY INSIGHT 수집·대조·적용·재조회는 코드, 조치 문안은 모델, 적용 결정은 사람이 맡음 — 해당 항목이 없는 회차에는 모델을 부르지 않아 비용이 들지 않음

코드 모델 사람 설계안 · 실측 없음



공지 값 추출

CVE 번호·제품명·영향 버전을 코드가 읽음

못 읽은 공지는 추측하지 않고 사람에게 넘김

폐쇄망 운영

공지는 망 연계로 반입한 파일을 읽음

외부 인터넷 접속 없이 같은 순서로 동작

파일럿 전 할 일

설계안 단계라 실측 수치 없음

파일럿 전에 합격선(도구 호출·응답 시간·정답)부터 정함

Source: AAS 업무 전문가 스킬 설계 백서 7장 (설계안, 실측 수치 없음)

실제 요청으로 재는 스킬 합격선

KEY INSIGHT

세 지표 모두 AAS 채팅 화면에서 읽을 수 있음 —
업무 전문가가 스크립트를 열지 않고도 합격 여부를 판정함

도구 호출
3회 이하
판단이 코드로 옮겨졌는가

응답 시간
수 분 이내
채팅 화면의 요청~응답 소요 시간

결과
정답
같은 요청 두세 번 반복해도 일치

측정 5단계

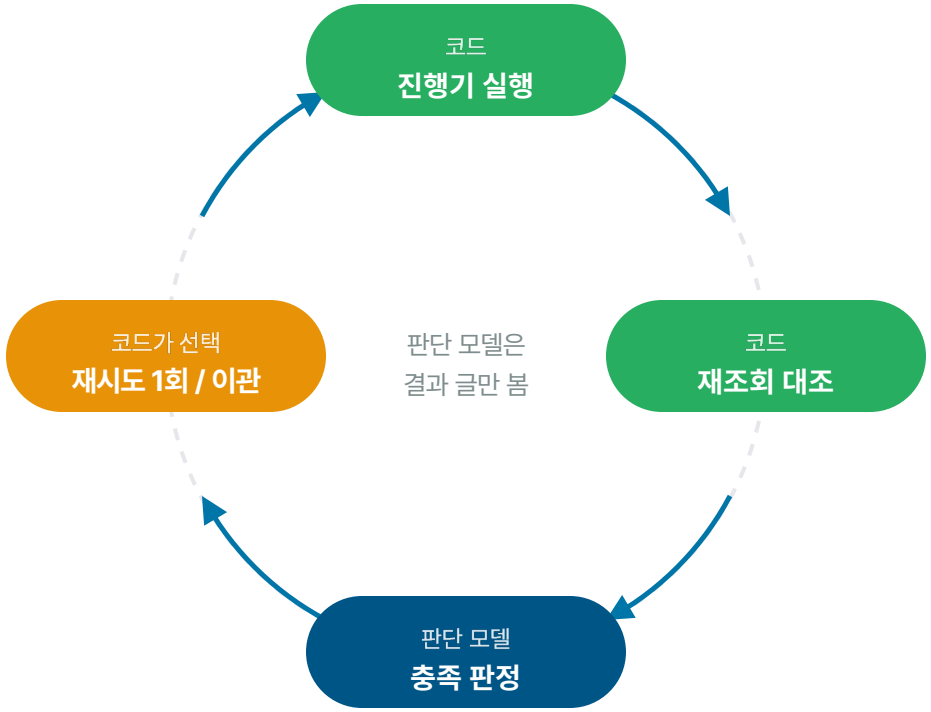
1 준비 | 2 입력 | 3 관찰 | 4 대조 | 5 반복

사례	도구 호출	응답 시간	결과	판정
CRM 조회 · 개선 전	36회	394초	오답	불합격
CRM 조회 · 개선 후	0회	14초	정답	합격
영업기회 등록 · 개선 전	11~15회	191~207초	두 번 중 한 번 실패	불합격
영업기회 등록 · 개선 후	1회	18초 + 33초	결과 JSON 일치	합격
한 줄 지시 재작성 대조	15회 → 7회	—	확인 게이트가 모델에 남음	불합격

Source: AAS 구성 운영 관찰 (한 업무 에이전트, 외부 벤치마크 아님)

결과 검증 순환과 재시도 1회 상한

KEY INSIGHT 쓰기 후 재조회가 어긋나면 재시도 없이 바로 사람에게 넘김 — 다시 실행하면 중복 등록이나 값 덮어쓰기가 생길 수 있음



실제 생성 여부는 코드 재조회가 판정보다 먼저 확인

실패 유형	처리	이관 대상
값 누락·요청 모호	재시도 없이 정해진 질문 출력 후 정지	요청자
업무 시스템 일시 오류 시간 초과·연결 실패	재시도 1회, 다시 실패하면 이관	IT 담당자
결과가 요청과 다름	판정 사유를 붙여 재시도 1회, 다시 실패하면 이관	업무 담당자
쓰기 후 재조회 불일치	**재시도 없이 즉시 이관**	업무 담당자· IT 담당자
확인 게이트 대상 금액·외부 발송·삭제	판정과 무관하게 사람 확인 후 실행	업무 담당자

재시도는 언제나 1회까지, 재시도 여부는 코드가 결정
— 같은 실패를 반복하며 비용을 쓰지 않음

Source: AAS 업무 전문가 스킬 설계 백서 8장

회귀 시험 기반의 스킬 변경 관리

KEY INSIGHT 전체 예시 합격선 충족과 기존 예시 무결 통과 두 조건으로 변경을 승인함 — 승인 기준이 숫자로 남아 담당자가 바뀌어도 판단이 같음

스킬 변경 승인 6단계

1. **업무 정의서 개정·변경 사유 기록**
   업무 전문가 · 바뀐 규칙이 문장으로 적혀 있음
2. **질문 예시 추가, 기존 예시 유지**
   업무 전문가 · 새 예시마다 기대 결과
3. **프론티어 모델이 MCP로 스크립트·시험·SKILL.md 수정**
   스크립트 수준 시험 전부 통과
4. **운영 모델로 전체 예시 묶음 재실행**
   승인 조건: 전체 합격선 충족 + 기존 예시 무결 통과
5. **스킬 업로드 반입 승인**
   IT 담당자 · 4단계 기록 첨부
6. **운영 환경에서 대표 예시 재실행**
   업무 전문가·IT 담당자 · 개발 환경 결과와 일치

운영 모델 교체 때 점검 6항목

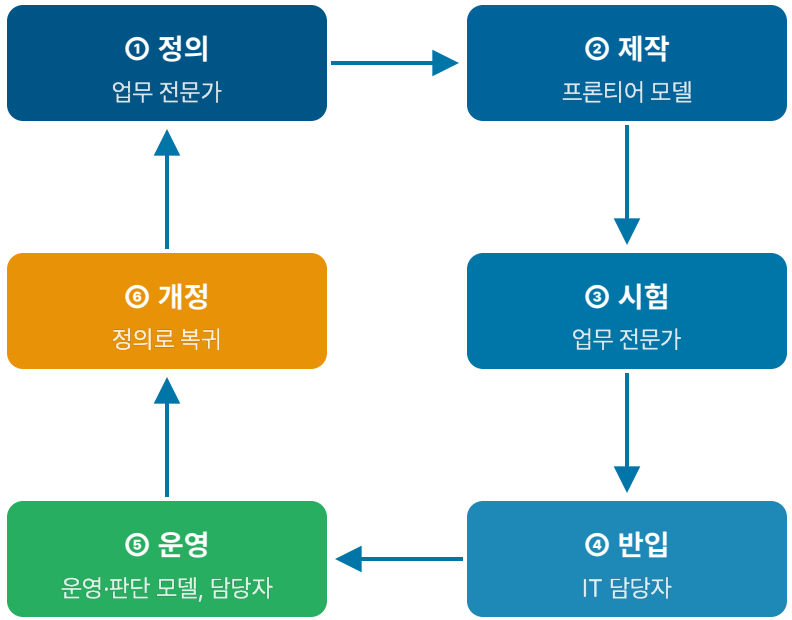
- 업무 결과 일치 (결과 JSON 대조)
- 첫 줄 라우팅이 같은 스킬을 고르는가
- 전달 충실도 (진행기 원문 유지)
- 도구 호출 형식 오류 0건
- 생각 모드·양자화 설정 기록
- 합격선 세 지표 충족

LiteLLM 게이트웨이 뒤에서 두 모델을 시험용 프로파일로 나란히 비교

Source: AAS 업무 전문가 스킬 설계 백서 8장; LiteLLM 문서

스킬 생애 주기와 책임 분담

KEY INSIGHT 개정은 언제나 정의 단계로 돌아가 제작·시험·반입을 다시 거침 — 운영 기록은 비식별 질문 예시로 바뀌어 다음 판의 재료가 됨



되돌아가는 경로

- 시험 불합격 → ② 제작
- 운영 중 이관된 실패 → ⑥ 개정
- 파일 변조 → 원본 목록으로 복원

대상	업무 전문가	IT 담당자	의사결정권자
업무 정의서·질문 예시	**소유·작성**	보관 위치 관리	대상 업무 선정
합격 판정	**정답 판정**	측정 환경 준비	기준값 승인
운영 반입	**반입 요청**	업로드 승인·원본 목록	반입 조건 승인
개정·모델 교체	**사유·예시 작성**	비교 실행	반영 승인

Source: AAS 업무 전문가 스킬 설계 백서 8장

PART 05

AAS 도입의 남은 한계와 파일럿 순서

작은 모델의 한계, 스킬 자가 수정 위험, 파일럿 대상 선정과 단계별 결정

작은 모델이 약한 업무 유형

KEY INSIGHT 매번 판단이 달라지는 업무와 범용 상담은 코드로 옮길 판단이 거의 없음 — 이런 업무는 반복되는 하위 단계만 스킬로 떼어 냄

업무 유형	작은 모델에서 생기는 현상	권하는 처리
매번 판단이 달라지는 업무	규칙 밖 요청이 모델 추측으로 넘어감	사람이 판단, 스킬은 자료 수집·초안
라벨 없는 자유 문장 입력	값 추출 실패로 되묻기 반복	요청 양식 지정, 질문 예시로 추출 규칙 보강
긴 추론이 필요한 업무	도구 호출 증가, 결과 불일치	큰 모델 또는 사람, 후속 반복 처리만 스킬화
범용 상담	코드로 옮길 판단이 거의 없음	파일럿 대상에서 제외
결과 글에만 의존하는 검증	실제 생성 여부를 확인하지 못함	코드 재조회를 판단 모델보다 앞에 배치

근거 수준 — 운영 반영 판단에 가까워지는 순서

NVIDIA 입장 논문
방향 주장, 비교 실험 아님

작은 모델 구성이 맞을 수 있다는 가설

▶

Mixtral·Qwen3 공개 자료
모델 구조 사양

활성 파라미터와 생성 속도의 관계 설명

▶

AAS 구성 전후 관찰
특정 업무·특정 모델의 동작

설계 원칙의 예시, 다른 업무에 그대로 적용 불가

▶

자기 업무로 잰 합격선 결과
자기 업무·모델·장비

운영 반영 판단의 직접 근거

Source: Belcak 외, NVIDIA Research (arXiv 2506.02153); Mixtral of Experts (arXiv 2401.04088); Qwen3-30B-A3B 모델 카드

운영 중 스킬 자가 수정 위험과 설정

KEY INSIGHT 작은 모델이 고친 스킬에서는 확인 게이트가 사라질 수 있음 —
쓰기 승인 대기·수정 차단 훅·원본 복원으로 변경 경로를 개발 환경 하나로 고정함

운영 스킬이 바뀌는 네 경로

대화 중 skill_manage 호출

바뀌는 것: 운영 스킬 본문·스크립트
위험: 회귀 시험을 거치지 않은 변경

대화 뒤 배경 검토

바뀌는 것: 모델이 고쳐 쓴 스킬 파일
위험: 사용자가 모르는 사이의 변경

스킬 허브 설치

바뀌는 것: 새 스킬 추가
위험: 기존 업무 스킬과 라우팅 충돌

운영 환경에서 사람이 직접 수정

바뀌는 것: 개별 파일
위험: 개발 환경 정본과 어긋남

운영 설정 점검 6항목

점검 항목	설정·장치	담당
스킬 쓰기 승인 대기	skills.write_approval 켜기	IT
스킬 수정 도구 차단	도구 호출 전 훅	IT
원본 복원	반입 때 남긴 원본 매니페스트	IT
변경 경로 단일화	개발 수정 → 회귀 시험 → 업로드	업무·IT
승인 대기 건 검토	정의서와 대조해 승인·반려	업무
버전 변경 재확인	Hermes Agent 판 변경 시 훅 시험	IT

업무 = 업무 전문가, IT = IT 담당자

Source: Hermes Agent 문서 — Skills; AAS 업무 전문가 스킬 설계 백서 9장

파일럿 대상 업무 선정 점검표

KEY INSIGHT

여섯 항목을 모두 통과한 업무 하나에 프로파일 하나로 시작함 —
연 1~2회 업무는 정의서·시험 작성 노력이 회수되지 않음

선정 점검 6항목

- [ ] **반복 빈도**
  같은 형태의 요청이 정기적으로 반복되는가
- [ ] **규칙 명시 가능성**
  입력값·처리 규칙·완료 상태를 문장으로 쓸 수 있는가
- [ ] **결과 판정 가능성**
  업무 전문가가 정답 여부를 바로 판정할 수 있는가
- [ ] **데이터 반출 제한**
  자료가 조직 밖으로 나가면 안 되는가
- [ ] **쓰기 영향 범위**
  잘못 기록했을 때 되돌릴 수 있는가
- [ ] **업무 전문가 확보**
  정의서를 쓰고 정답을 판정할 담당자가 있는가

파일럿에 맞지 않는 업무

- 사내 규정 범용 상담
- 영업 전략 수립·인사 평가
- 여러 로그를 엮는 장애 원인 분석
- 자유 형식 메일에서 여러 값을 뽑아 바로 기록하는 업무
- 연 1~2회 하는 업무

시작 단위: 업무 하나 + 프로파일 하나

Source: AAS 업무 전문가 스킬 설계 백서 9장

파일럿 7단계와 단계별 결정 항목

KEY INSIGHT 첫 결정은 부서 반복 업무 하나와 업무 전문가 지정 —
GPU 규모는 운영 측정 뒤 부서 확산 단계에서 정함

의사결정권자가 정할 것	단계	IT 담당자가 준비할 것
업무 하나·업무 전문가 지정	**1 대상 업무 선정**	데이터 반출 제한 확인
작성 인력·시간 배정	**2 업무 정의서 작성**	비밀값·접근 권한 목록
프론티어 AI 사용 범위 승인	**3 개발 환경 구성**	디지털 트윈·MCP 연결
시험 요청 범위	**4 스킬 제작·드라이런**	시험 프로파일
합격선 값·결과 승인	**5 합격선 확인**	도구 호출 수·응답 시간 기록
반입 승인	**6 스킬 업로드 반입**	보안 검사·쓰기 승인 대기·훅
확산 범위·GPU 규모	**7 운영 측정·부서 확산**	가상 키·예산·모델 서빙

Source: AAS 업무 전문가 스킬 설계 백서 9장

THANK YOU

AAS 파일럿 첫 단계: 업무 정의서 한 장

부서 반복 업무 하나를 골라 지시문 해석·결정적 코드·질문 예시를 담은
업무 정의서로 파일럿 대상 확정

MSAP.ai

msap.ai

https://www.msap.ai/

02-6953-5427

hello@msap.ai

© MSAP.ai. All rights reserved.

CONFIDENTIAL