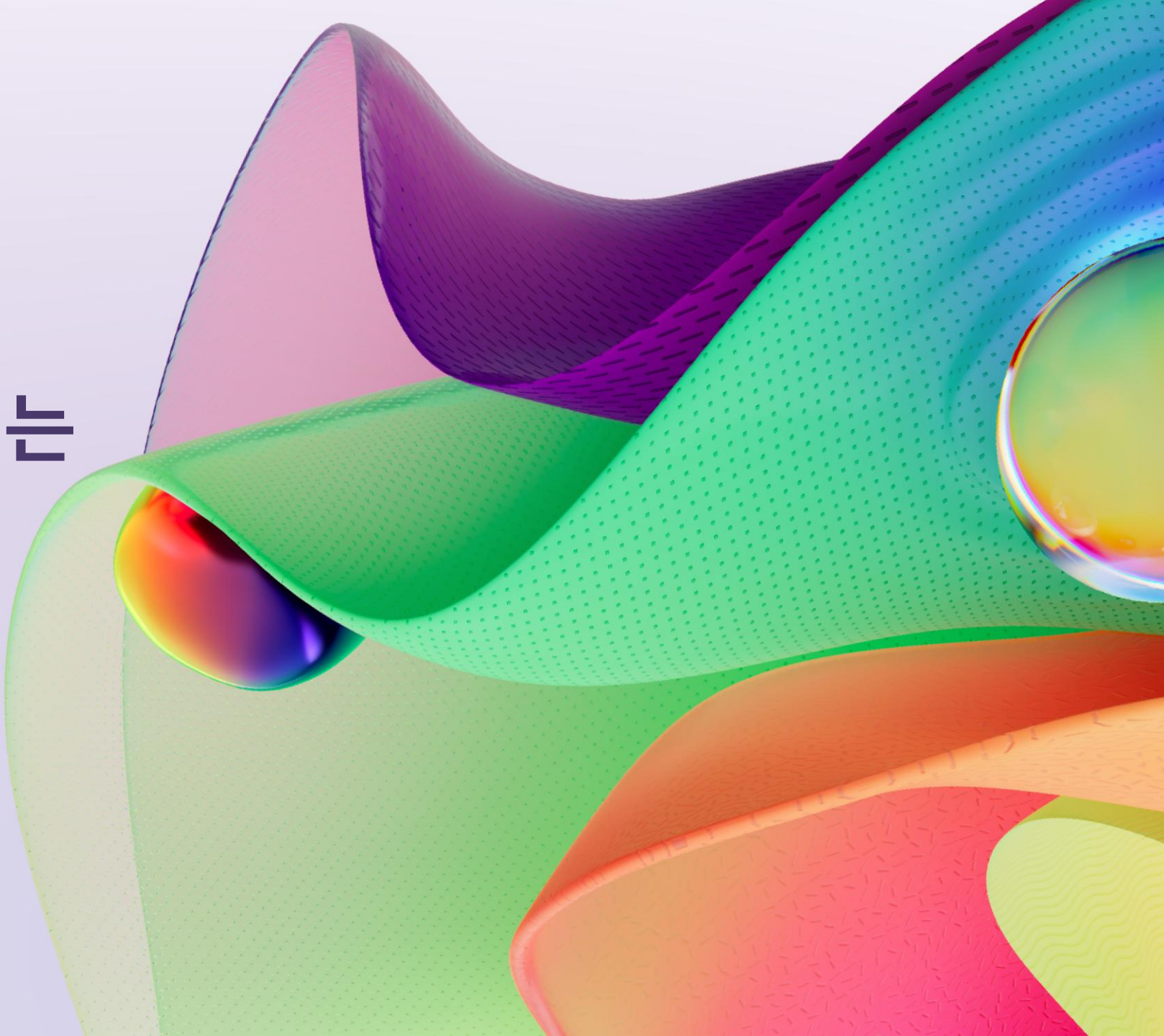
Microsoft

내 시스템에 꼭 맞는 AI 모델 선택하기

Microsoft Education Success Manager
김근희



김근희

Microsoft Education Success Manager

Microsoft Student Ambassadors Senior

@g1nya2

@geunhee-kim1227


Microsoft

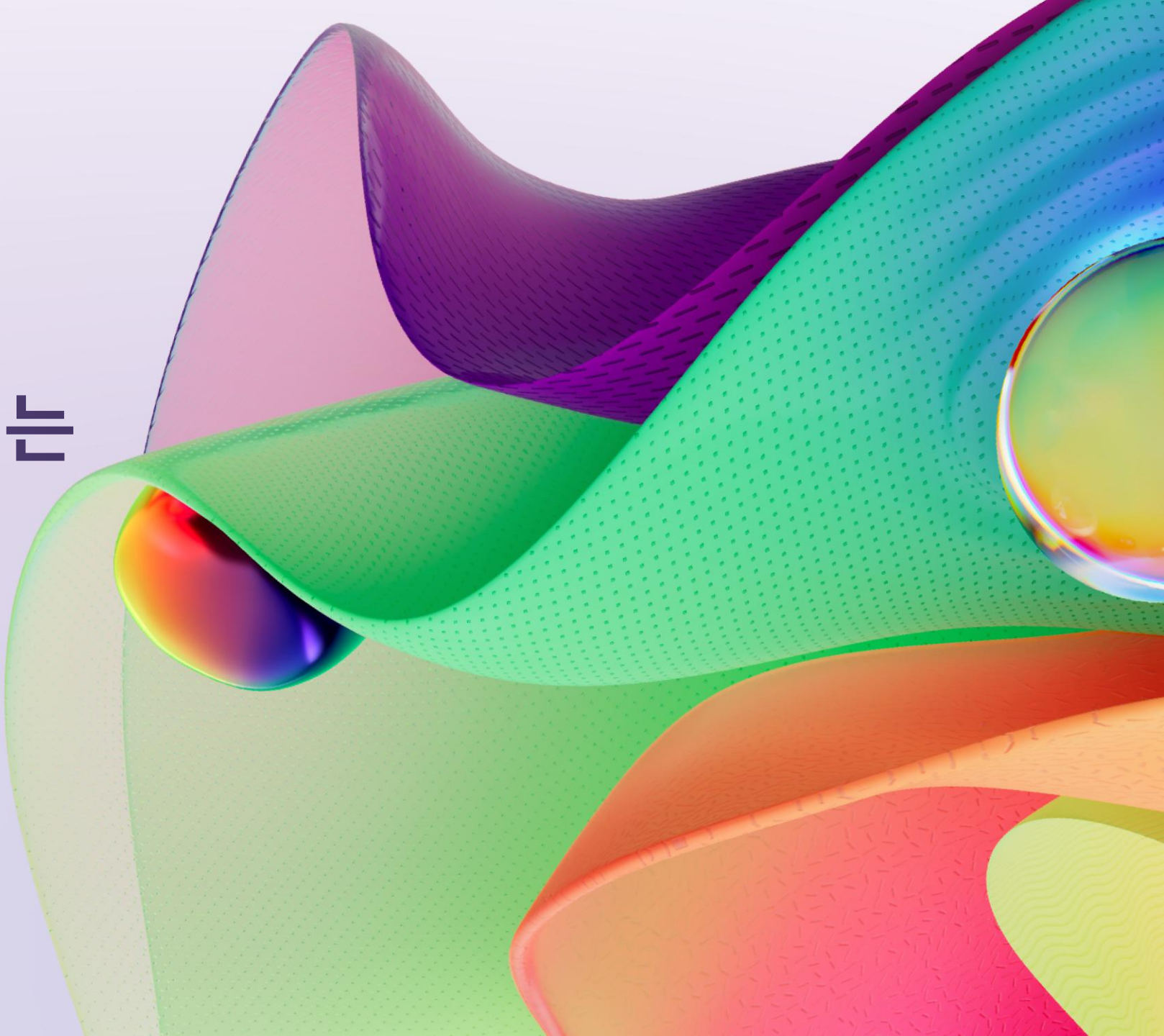

OceanLightAI


AIFactory

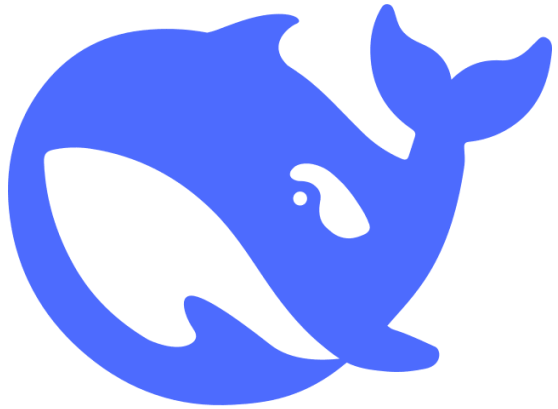
Microsoft

내 시스템에 꼭 맞는 AI 모델 선택하기

Microsoft Education Success Manager
김근희





어떤 모델 하나가 좋냐?

이 복잡한 작업들을 어떻게 나누고,
각 작업에 어떤 모델과 도구를 배치할 것인가?

어떤 모델을 사용할까?

좋은 모델을 계속 찾고 바꿀 수 있는 시스템을 만들자.


Microsoft Foundry

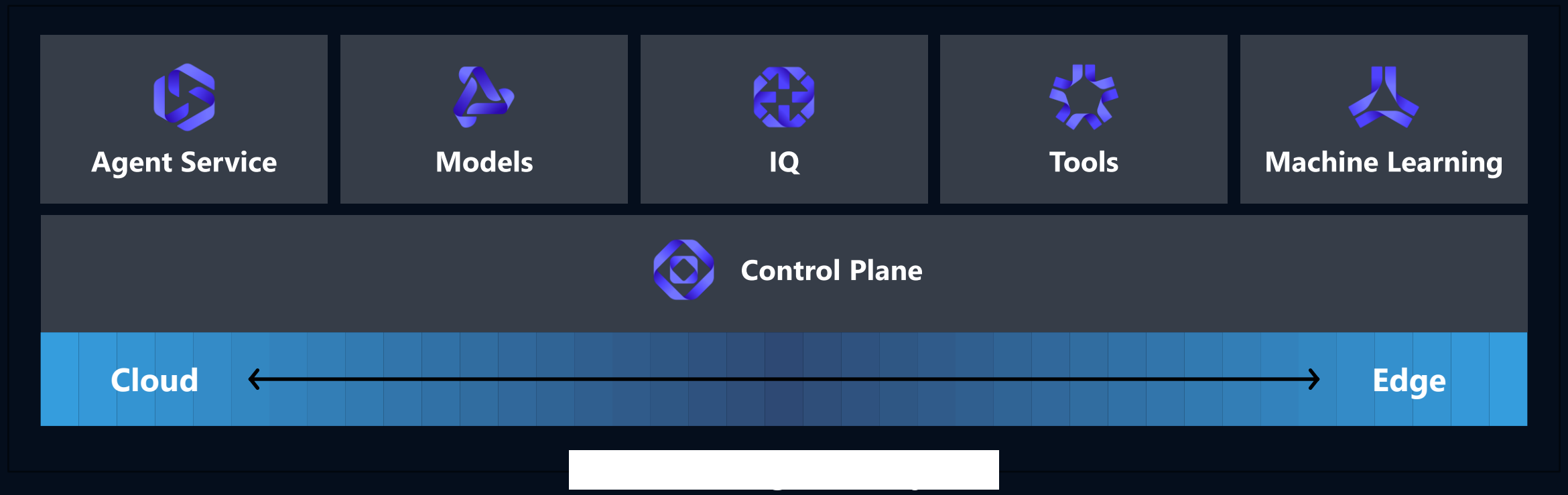
Microsoft Foundry란?

Foundry = 모델을 고르고, 에이전트를 만들고, 평가하고, 운영하는 “AI 시스템 작업대”

다양한 모델 탐색	에이전트 구성	품질/안전성 평가	배포와 버전관리

Microsoft Foundry

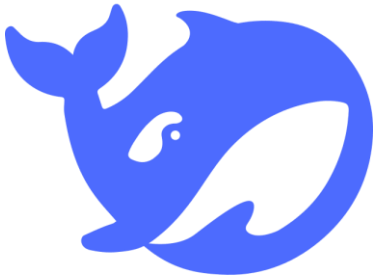
The AI app and agent factory



AI 앱 개발 방식이 바뀌었습니다

좋은 모델 선택
↓
앱에 연결
↓
배포





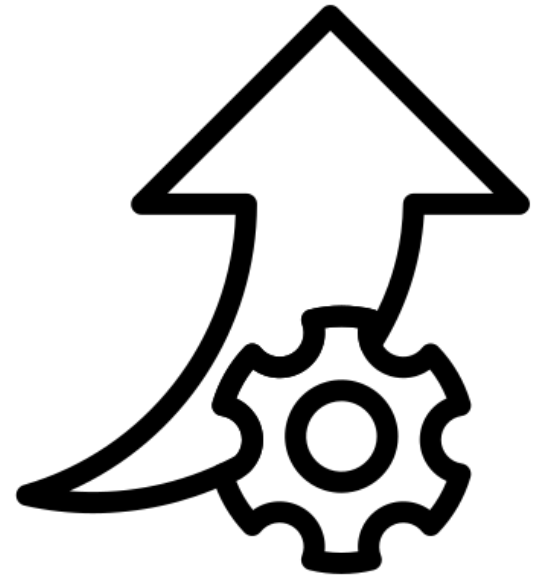

왜 더 어려워졌나?

선택지가 너무 많아짐

공개 벤치마크만으로는 부족

사용량·토큰·처리량에 따라 변동

품질·안전·성능을 운영에서 관리



User Request → Task Decomposition →

Model Selection → Evaluation →

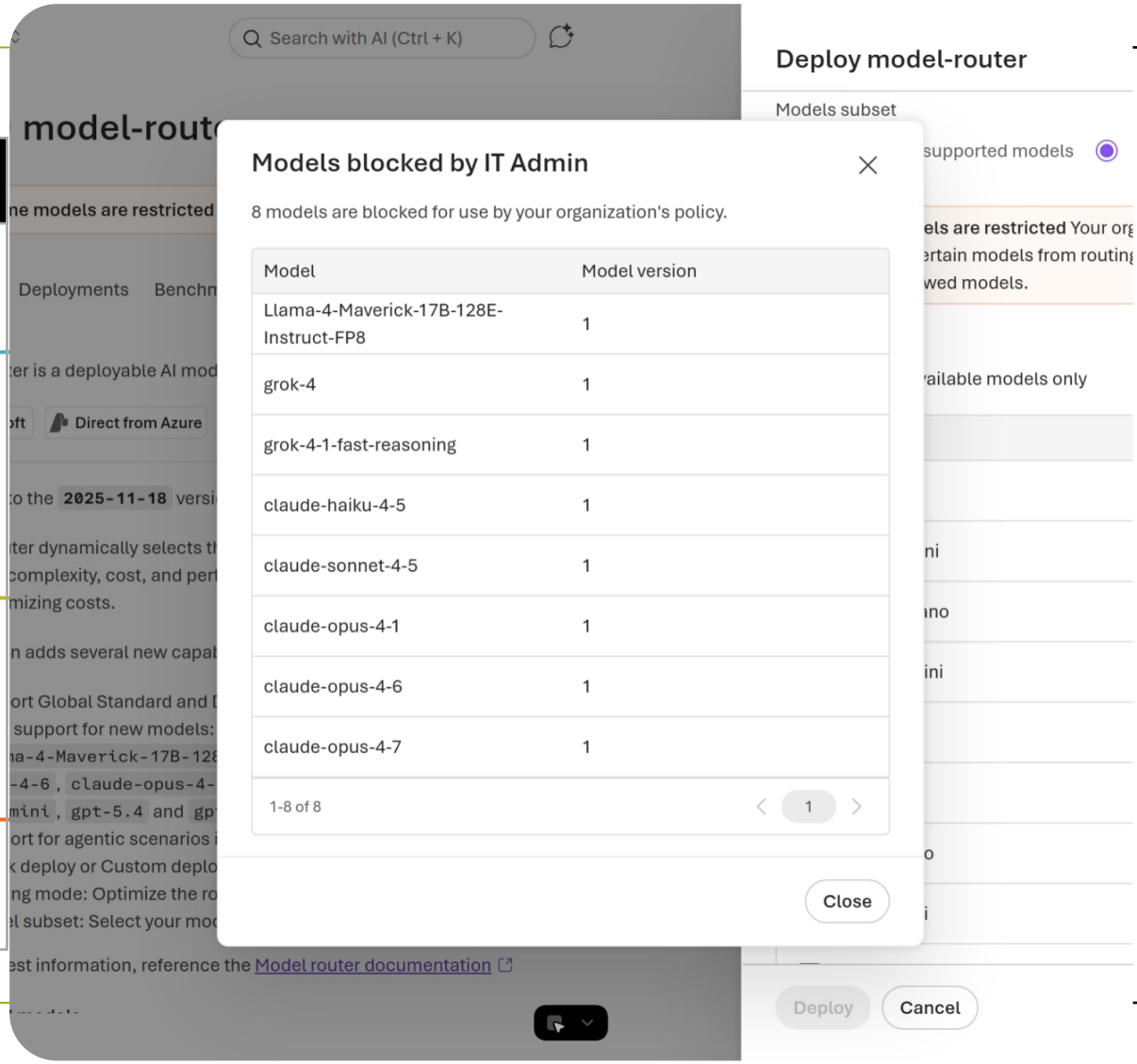
Optimization → Operation → 다시 개선

다양한 모델 할당하는 방법

소비자 요청사항

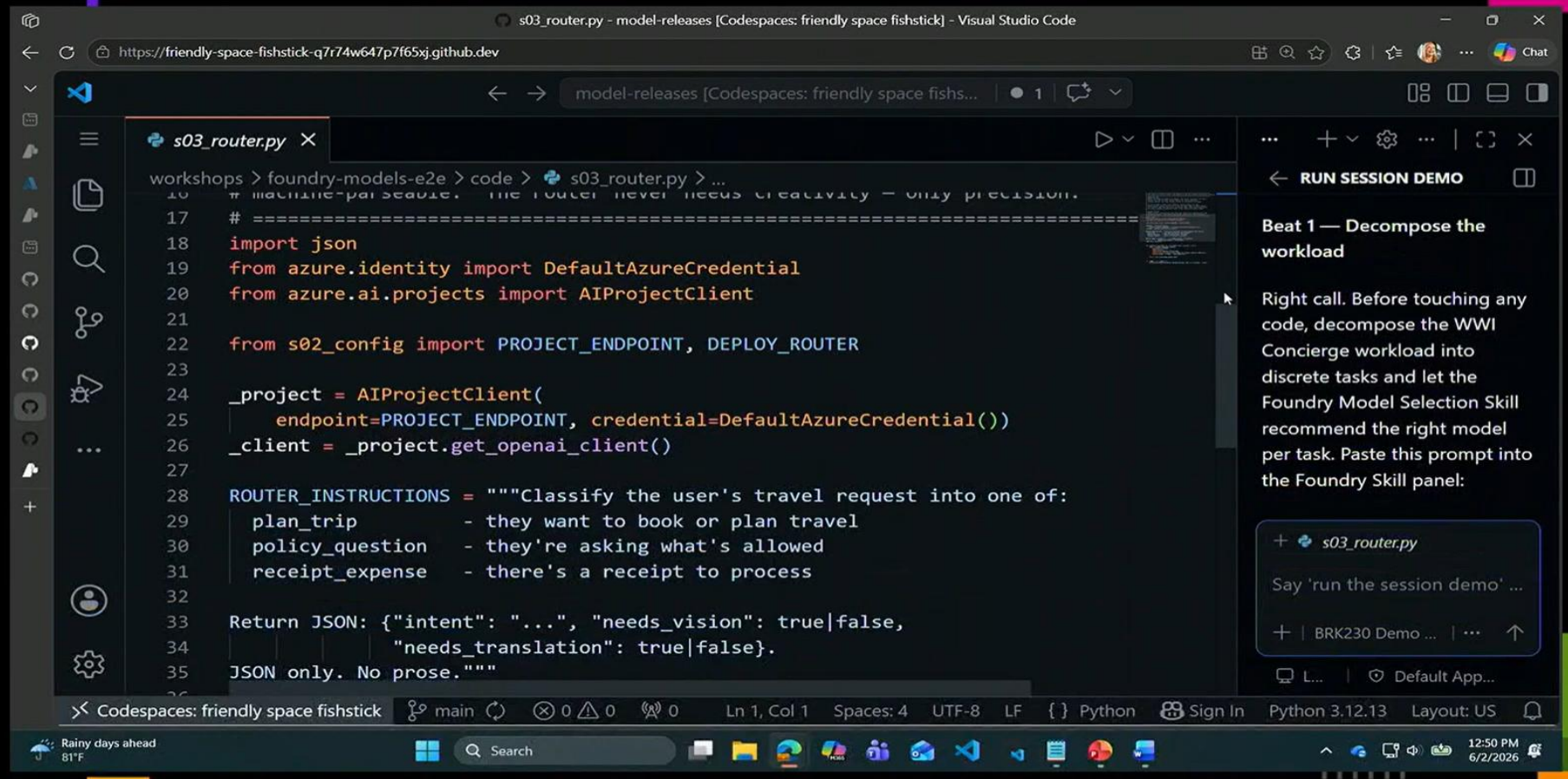
해야하는 일들

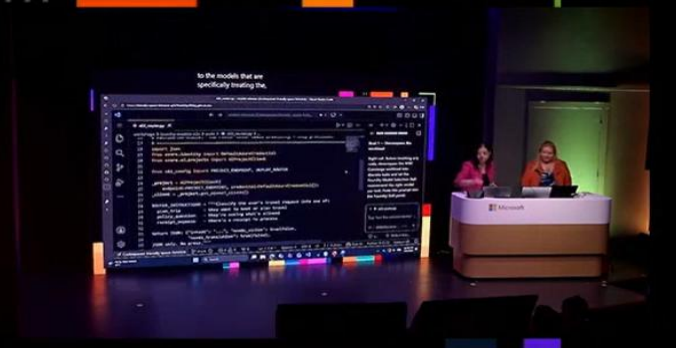
- 여행 경로 탐색하기
- 영수증 읽기
- 계획하고 툴을 불러오기



평가하기

- 품질
- 비용
- 시간


s03_router.py - model-releases [Codespaces: friendly space fishstick] - Visual Studio Code
https://friendly-space-fishstick-q7r74w647p7f65xj.github.dev
Chat
model-releases [Codespaces: friendly space fishs...
s03_router.py
workshops > foundry-models-e2e > code > s03_router.py > ...
17 # ==========
18 import json
19 from azure.identity import DefaultAzureCredential
20 from azure.ai.projects import AIProjectClient
21
22 from s02_config import PROJECT_ENDPOINT, DEPLOY_ROUTER
23
24 _project = AIProjectClient(
25 endpoint=PROJECT_ENDPOINT, credential=DefaultAzureCredential())
26 _client = _project.get_openai_client()
27
28 ROUTER_INSTRUCTIONS = """Classify the user's travel request into one of:
29 plan_trip - they want to book or plan travel
30 policy_question - they're asking what's allowed
31 receipt_expense - there's a receipt to process
32
33 Return JSON: {"intent": "...", "needs_vision": true|false,
34 "needs_translation": true|false}.
35 JSON only. No prose."""
RUN SESSION DEMO
Beat 1 — Decompose the workload
Right call. Before touching any code, decompose the WWI Concierge workload into discrete tasks and let the Foundry Model Selection Skill recommend the right model per task. Paste this prompt into the Foundry Skill panel:
s03_router.py
Say 'run the session demo' ...
BRK230 Demo ...
Default App...
Codespaces: friendly space fishstick
main
Ln 1, Col 1
Spaces: 4
UTF-8
LF
Python
Sign In
Python 3.12.13
Layout: US
Rainy days ahead
81°F
Search
12:50 PM
6/2/2026



감사합니다.

Evaluate

Evaluators in Foundry

Built-in evaluators

Ready-made checks for common AI system signals

Quality

Risk & safety

Agentic

Custom evaluators

Use custom checks to encode **your** business logic

Prompt-based

Code-based

Rubric-based

demo

Custom Evaluators

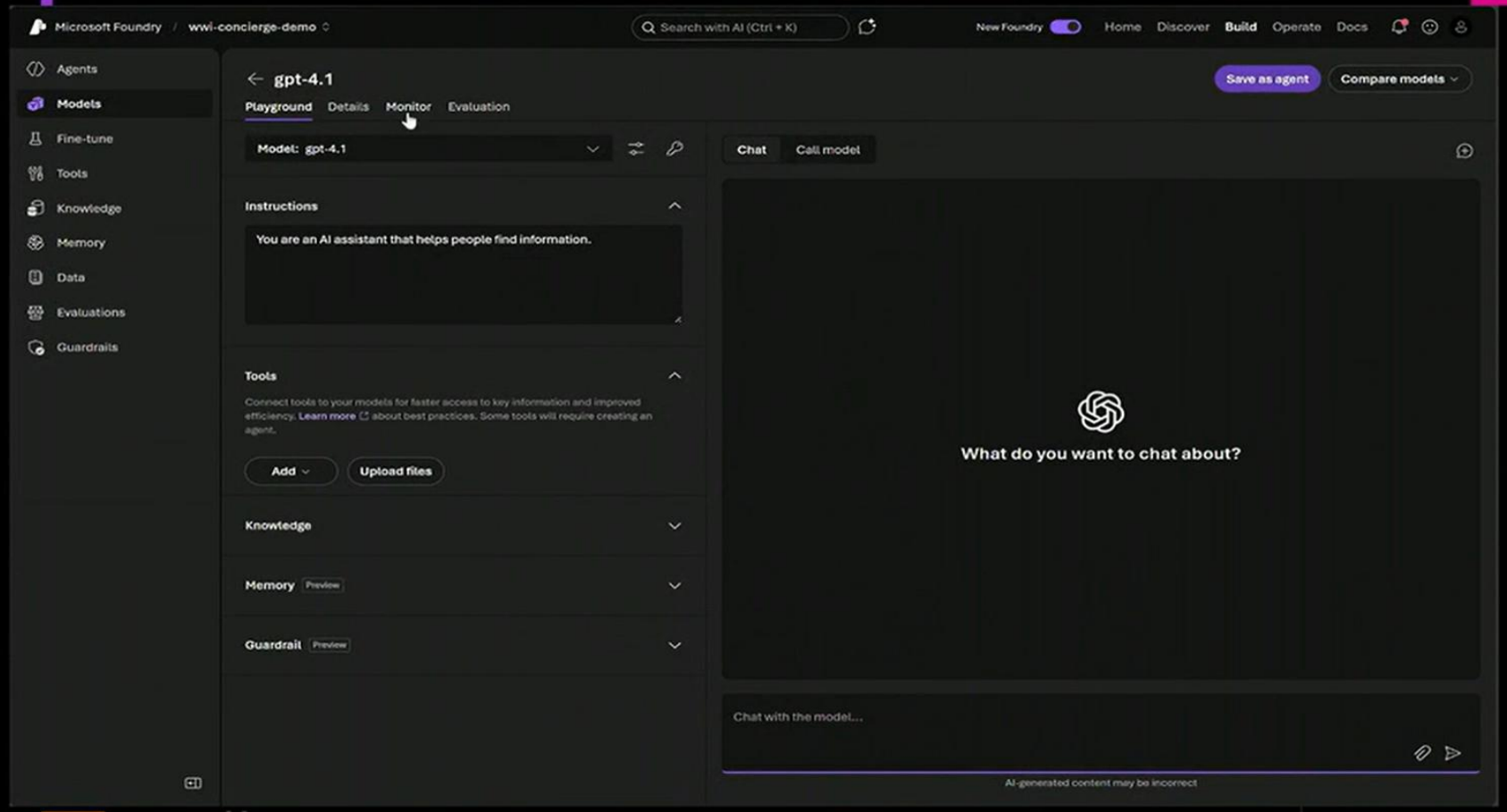
Try it at: aka.ms/build/BRK230

그걸 보여줄게.

Optimize

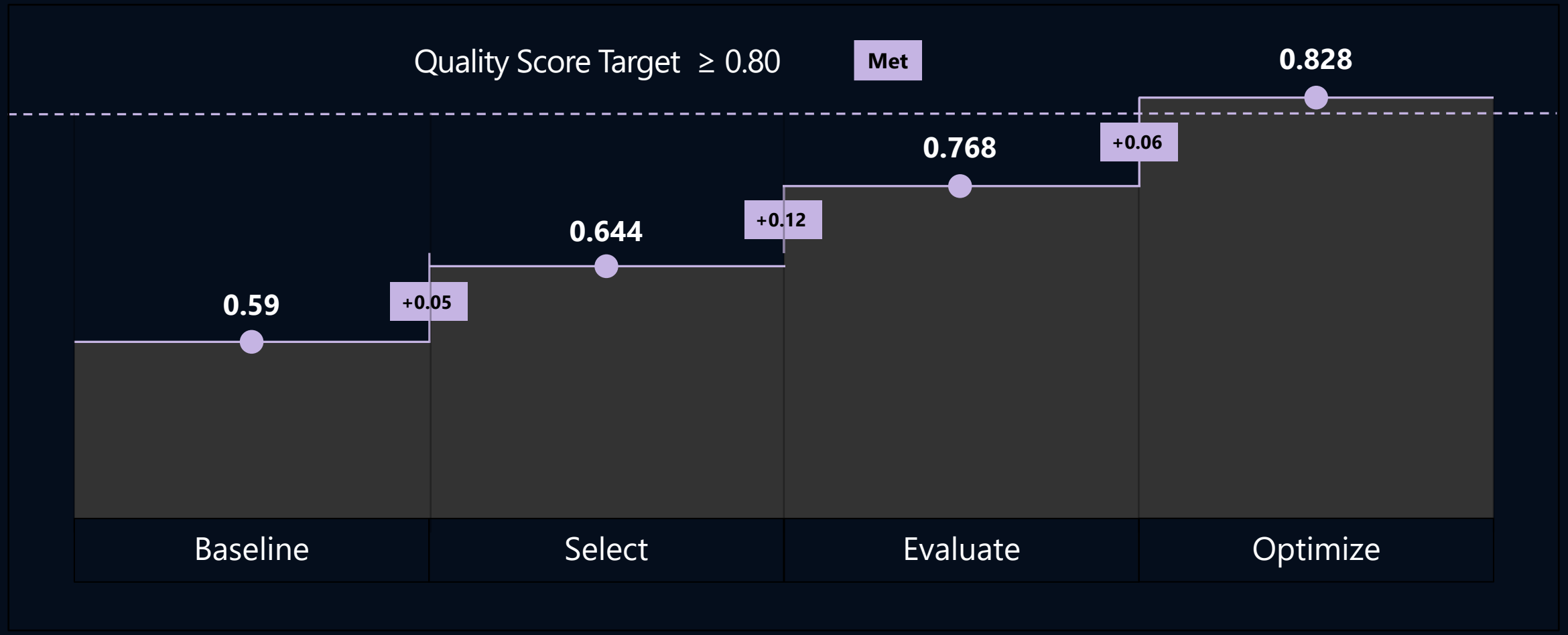
Operate

Foundry 모니터링



자세한 내용은 다루지 않겠지만, 파운드리 포털에서 이 모습을 빠르게 보여드리자면,

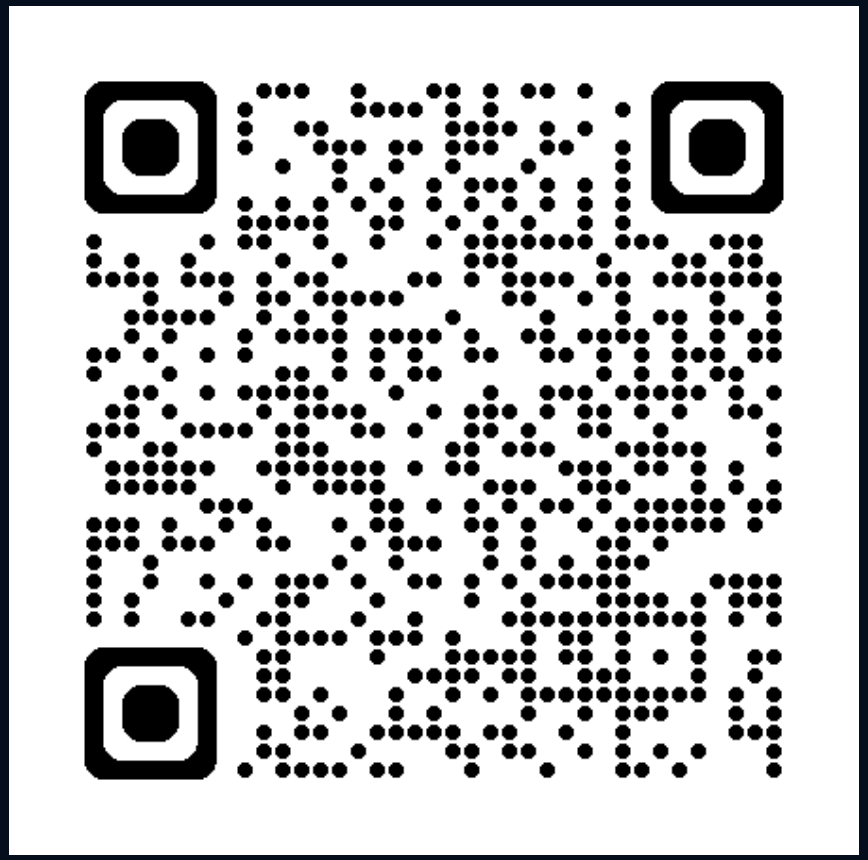
Hill Climbing



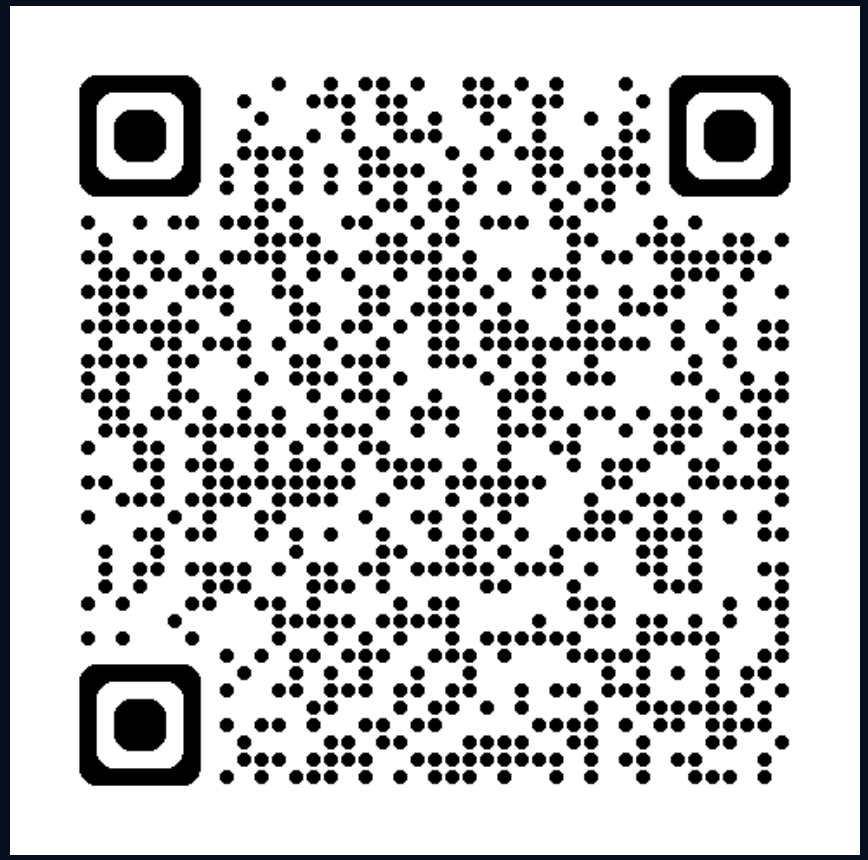
AI 서비스는 모델 호출이 아니라 시스템 설계다

Build the System,
Not Just the Prompt

1. 모델 선택은 일회성 결정이 아니다
2. 작업별로 모델을 고르고 평가해야 한다
3. 비용·속도·품질은 함께 최적화해야 한다
4. 운영하면서 계속 개선해야 한다



Build 영상



GitHub 자료

Microsoft

내 시스템에 꼭 맞는 AI 모델 선택하기

Microsoft Education Success Manager
김근희