



건설 데이터를 이해하는 AI : 온톨로지의 역할

이정훈, Team Leader, Solutions Engineers, Cloud & AI Platform, Microsoft
junghunlee@microsoft.com

Agenda

01 최근 AI 트렌드

모델 진화 속도 · 에이전트 확산 · 코딩 에이전트

02 AI 트랜스포메이션을 위한 엔터프라이즈 전략

Agent 전사 확산 · IQ 지능 레이어 · 보안 거버넌스

03 Ontology: The Context Layer for AI

RAG의 한계 · GraphRAG · Fabric IQ Ontology

04 건설사에 활용 가능한 Lessons Learned 공유

도면 분석 사례 · 7가지 교훈 · Ontology Agent 데모

01

최근 AI 트렌드

모델 진화 속도 · 에이전트 확산 · 코딩 에이전트

모델 출시 주기와 성능 향상 속도

2026년 출시 주기 9배 단축

① 출시 주기 9배 단축

438일

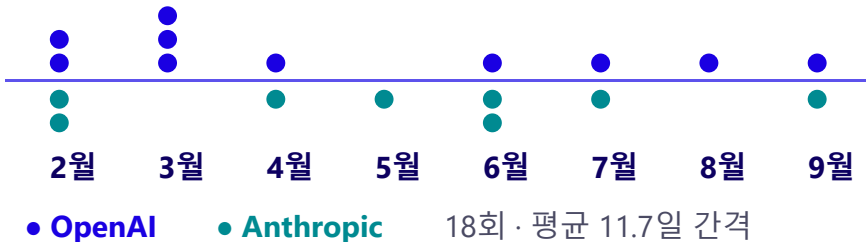
2023~2025
GPT-4 → GPT-4o → GPT-5



9배 단축

49일

2026년
GPT-5.1 이후



② 한 세대 내 성능 2배 세대 간격 56일

2.3배

AutomationBench
18.1 → 41.4

+20.6%p

Terminal-Bench 4.0
37.3 → 57.9

+15.8%p

ScreenSpot-Pro
76.9 → 92.7

③ 벤치마크 포화와 버전 교체

100% AIME 2025 · 97% SimpleQA · 96% GPQA Diamond

추적 12개 중 7개가 85% 초과

Terminal-Bench 2.0 → 2.1 → 4.0 · 1년 내 3회 버전 교체

디지털 활동 주체의 전환

트래픽 · 콘텐츠 생성 · 토큰 소비에서 확인되는 역전

① 웹 트래픽

AI · 봇 57.5%

사람 42.6%

57.5%

출처: Cloudflare 2026-06 · HTML 요청 기준

② 웹 콘텐츠 생성

AI 35%

사람 65%

35%

출처: Pew 2026-08 · 일자 표기 페이지 기준

③ 토큰 소비

사람 1배

AI 에이전트 5배

5배

출처: a16z 2026-08 · OpenRouter 기준

에이전트 토큰 사용량 2026년 2월 대비 14배 증가

디지털의 주 사용자가 사람에서 AI로 이동

코딩 에이전트 보편화와 교육 과정의 대응

현장 사용률 90% · 대학은 기초와 검증 역량으로 재편

개발자의 AI 활용 현황

90%

주 1회 이상

68%

매일

"Typing code is absolutely over"

David Fowler · Microsoft

위임

→ 실행

→ 피드백

→ 책임

스탠포드 CS 교육의 대응

비중 축소

구현 수준 코딩
보일러플레이트 · 단위 테스트
문서화

비중 강화

시스템 설계 · 추상화
AI 산출물 검증
비판적 평가

대면 평가 도입

조교 1:1 이해도 확인

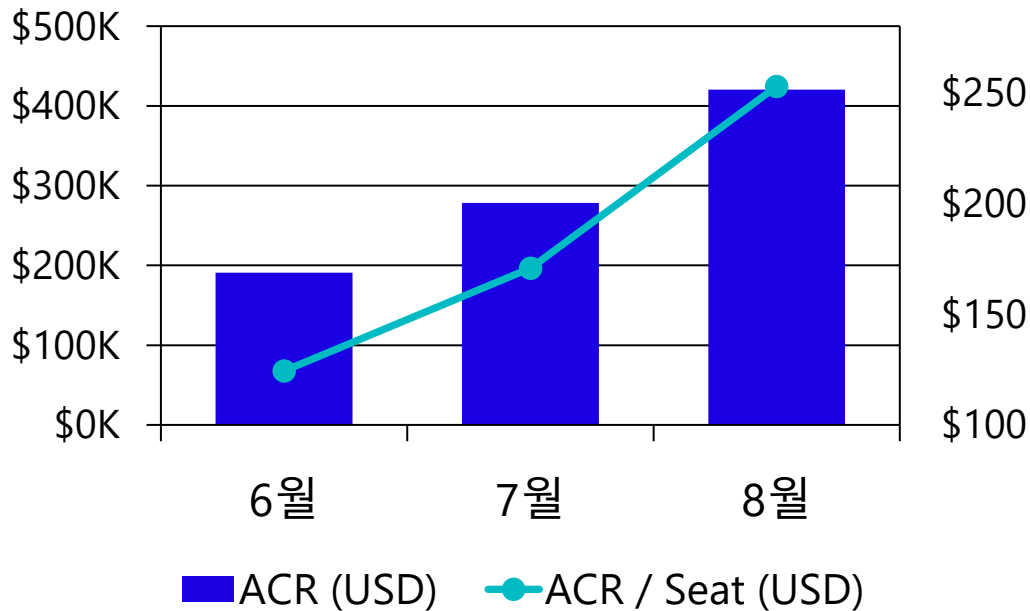
현장 시험 비중 확대

테이크홈 과제 축소

"AI를 매일 사용하는 개발자도 기본기를 갖춰야 한다" —
Chris Gregg, 스탠포드 CS

GitHub Copilot 사용량 증가 추이

국내 대형 고객 · 좌석당 사용액 2개월 2배



국내 고객 A · 약 1,600명

+51.0%

ACR · 7월 → 8월

+48.1%

ACR/Seat · 7월 → 8월

+103.3%

ACR/Seat · 6월 → 8월

+2.0%

Seats · 7월 → 8월

02

AI 트랜스포메이션을 위한 엔터프라이즈 전략

Agent 전사 확산 · IQ 지능 레이어 · 보안 거버넌스

전사 AI 전환을 위한 필수 요구 사항

01

Agent 전사 Scale

모든 직원의
AI 에이전트 활용

접근 · 개발 · 활용

전직원이 에이전트를 직접 만들고
AI의 팀장 역할 수행

02

AI를 위한 데이터

지능 레이어
(IQ Layer)

일 잘하는 AI 토대

비즈니스 맥락을 이해하는
숙련된 직원처럼
업무 역할을 자율 수행

03

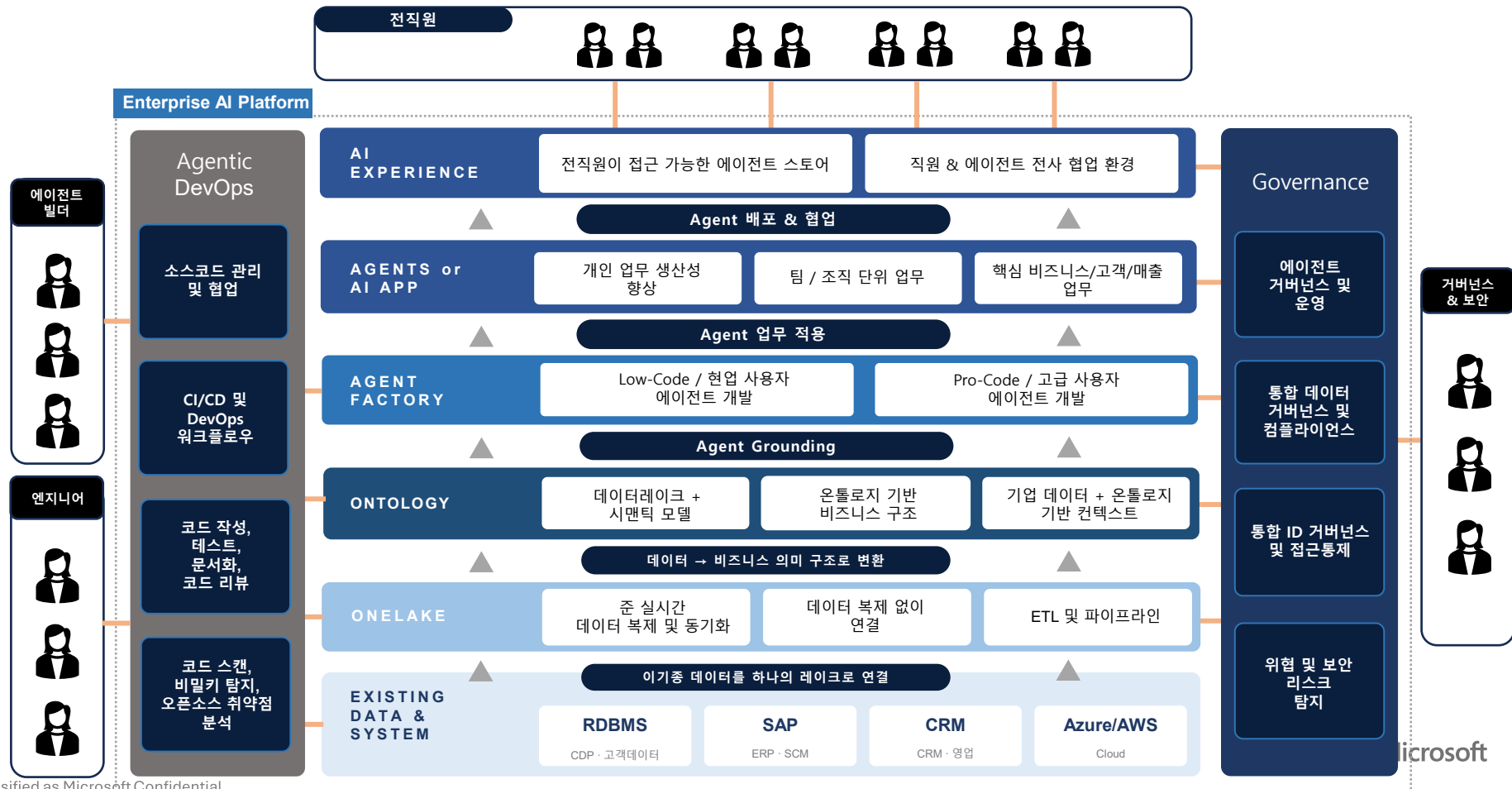
AI 보안 · 거버넌스

보안/거버넌스
Agent의 HR시스템

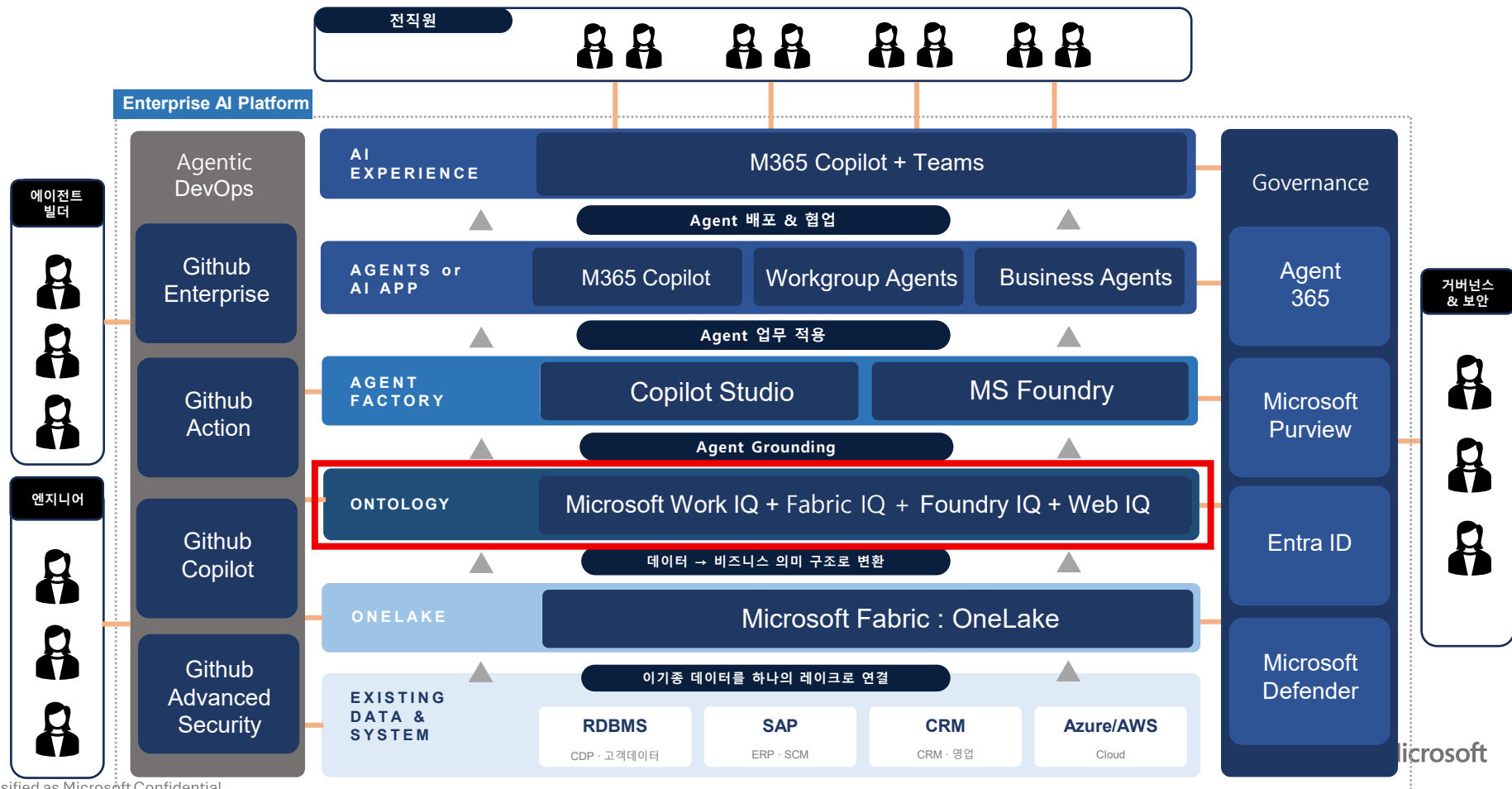
안전한 전사 확산

Agent가 제대로 일하고 있는지
직원 처럼 관리

Unified Enterprise AI Platform



Unified Enterprise AI Platform



Frontier Tuning의 필요성과 구조

왜 필요한가

01 모델의 상품화

동일 프런티어 모델을 모든 기업이 같은 가격에 구매

차별화는 어디에 어떻게 배포하는가로 이동

02 성공적인 AI 활용

플랫폼 자체가 아니라 고객별 커스터마이징

현장 워크플로에 맞춘 구성이 가치를 만듦



Frontier Tuning 구성 3요소 · Microsoft Build 2026

1

강화학습 환경 (RLE)

실제 워크플로·도구 사용·평가 신호로 학습, 운영 시스템과 분리

2

자사 데이터·도메인 지식·워크플로

콘텐츠·절차·용어·규정을 RLE에 투입, 데이터 사이언스 전문성 불필요

3

튜닝된 모델·스킬·하니스

컴플라이언스 경계 내 실행, 원본 데이터 접근 권한 상속

고객별 커스터마이징을 사람의 작업이 아닌 학습 루프로 전환 |
작업 성공률 13% → 87% (Microsoft HR 워크플로 사례)

03

Ontology: The Context Layer for AI

RAG의 한계 · GraphRAG · Fabric IQ Ontology

Intelligent의 한계

Intelligent란?

지식을 상황에 맞게 활용하고 추론·판단을 통해 새로운 문제에 적용하고 대응할 수 있는 능력

지식이 우선 과제

LLM 학습 모델의 한계

- 최근 정보가 없다
- 나의 회사 데이터가 없다

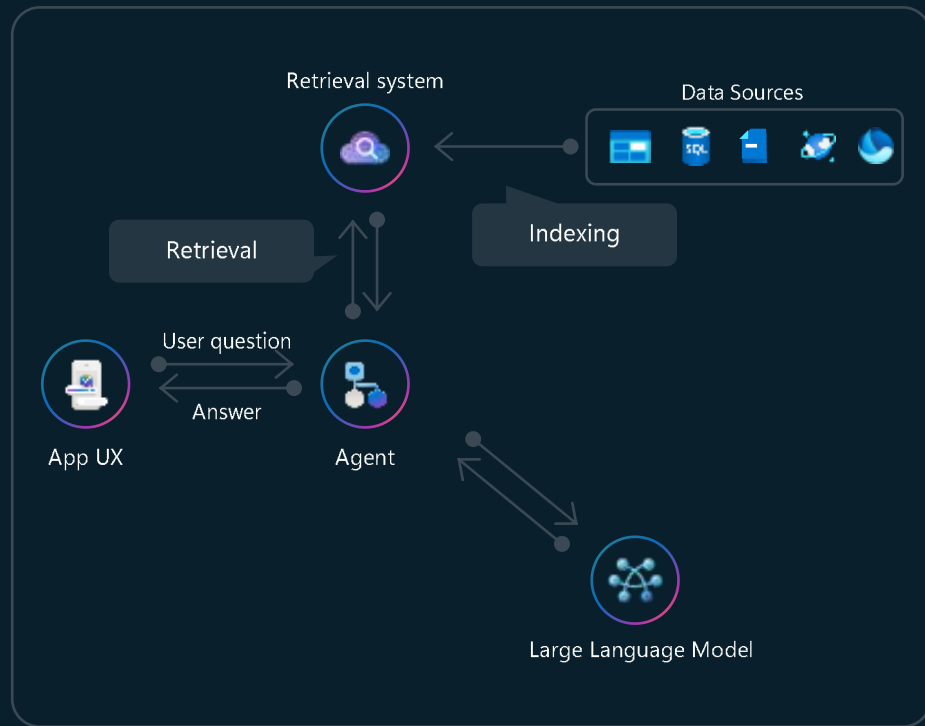
Hallucination의 위험

이걸 극복하기 위해서 **Prompt**에 **Context**[최신정보|회사데이터]를 넣자

- 어떤 데이터를 Prompt에 넣을 것인가?



Foundational RAG



왜 기존 RAG만으로는 부족할까?

기존 Vector RAG가 부딪히는 세 가지 구조적 한계

01

Semantic $\neq$ Causal

벡터 임베딩의 맹점

문서를 잘게 나눈 텍스트 조각들을 벡터로 표현해 개별적으로 검색합니다. 이 때문에 원래 문서에 존재하던 인과 관계 흐름 등이 모두 사라지고, 여러 조각 사이의 연결관계를 인식하지 못합니다.

02

Single Passage Trap

다단계 추론의 어려움

어떤 질문의 답이 A→B를 거쳐 C에 도달해야 하는 상황에서 전통 RAG는 질문과 직접적으로 연관된 조각만 찾기에 개념 B가 질문 표현에 드러나 있지 않으면 해당 조각을 건너뛰어 논리적 고리를 놓칠 수 있습니다

03

No Holistic View

전체 맥락 파악 한계

여러 문서 전반의 거시적 주제를 묻는 질문에도 취약합니다. 질문에 나온 키워드와 유사한 단어가 포함된 개별 문단들만 주로 반환될 뿐, 수많은 문서에서 공통으로 드러나는 큰 그림을 종합해 답하지는 못합니다.



Project GraphRAG

LLM-Derived Knowledge Graphs

Question: Which public figures are repeatedly mentioned across various entertainment articles?

Naive RAG response

Public figures who are repeatedly mentioned across various entertainment articles include <Public figures 1, 2, 3, and 4>. These individuals have been highlighted for various reasons, ranging from their professional achievements to their personal lives.

<Public figure 1> [...]

<Public figure 2> [...]

<Public figure 3> [...]

<Public figure 4> [...]

These figures are frequently covered due to their high-profile status and the public's interest in their careers and personal lives. Their activities, whether in music, sports, or personal relationships, have significant cultural and economic impacts, as evidenced by the media coverage and public reactions.

GraphRAG response

The entertainment industry is vast and diverse, encompassing film, television, music, sports, and digital media. Certain public figures stand out due to their significant contributions and influence across these sectors. The following summary highlights key individuals who are repeatedly mentioned in various entertainment articles, reflecting their impact and presence within the industry.

Actors and Directors [... 20 figures ...]

Public Figures in Controversy [... 3 figures ...]

Musicians and Executives [... 5 figures ...]

Athletes and Coaches [... 7 figures ...]

Influencers and Entrepreneurs [... 3 figures ...]

The repeated mention of these figures in entertainment articles signifies their ongoing relevance and the public's interest in their work. Their influence spans across various aspects of entertainment, from shaping cultural narratives in film and television to driving trends in music and digital media. These individuals not only contribute to their respective fields but also influence the broader cultural landscape, often becoming central figures in social discussions and public discourse.

LLM evaluation

Comprehensiveness: Winner = GraphRAG

Diversity: Winner = GraphRAG

Empowerment: Winner = GraphRAG

GraphRAG에 적합한 시나리오

엔티티 간의 관계를 이해해야 답이 나오는 업무 — 벡터 검색만으로는 한계가 있는 영역



다중 홉 관계 추론

여러 엔티티·문서를 거쳐야 답이 나오는 질문

예) "A 프로젝트 지연이 B 고객 계약에 미친 영향"



엔터프라이즈 지식 그래프

조직·제품·공급망처럼 계층·의존 구조가 뚜렷한 도메인

예) 조직도 Q&A, 부품 BOM, 시스템 의존성 매핑



문서 간 교차 요약

수백~수천 문서에서 공통 테마·패턴을 종합해야 하는 질의

예) "최근 1년 CS 티켓에서 반복되는 품질 이슈 Top 5"



규제·컴플라이언스 분석

조항·판례·정책이 서로를 참조하는 복합 규칙 탐색

예) GDPR·국내법 상호 참조, 감사 증적 추적

핵심: "관계 질문이 많다 = GraphRAG 후보" — 단일 문서 Q&A에만 머문다면 기존 RAG로 충분

Graph RAG의 3가지 대표 모델

원본 텍스트와 지식 그래프를 어떻게 결합할 것인가

①

인덱스 기반

Index-based

원본 텍스트를 그대로 유지하면서, 텍스트 청크들을 연결하는 그래프를 색인 도구로 활용한다.

대표 예

MS GraphRAG

②

지식 기반

Knowledge-based

텍스트를 구조화된 사실(Fact)로 변환하여 지식 그래프 자체를 지식 저장소로 사용한다.

대표 예

RoG, ToG

③

하이브리드

Hybrid

원본 텍스트의 상세한 맥락과 지식 그래프의 논리적 구조를 함께 결합해 활용한다.

대표 예

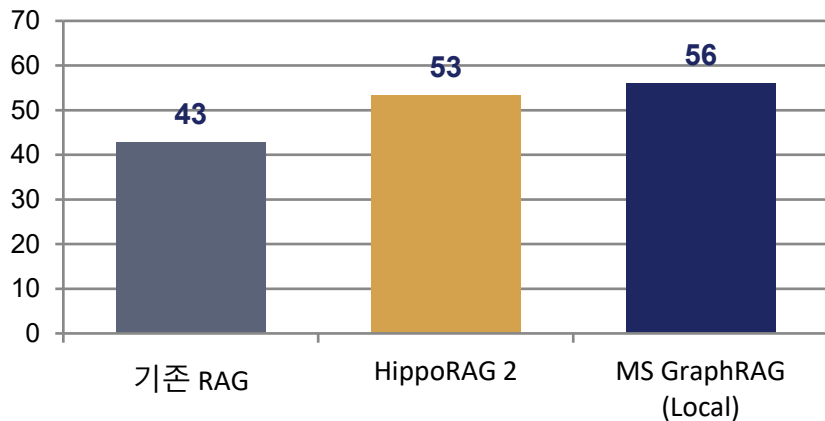
LightRAG

성능은 오르고, 비용은?

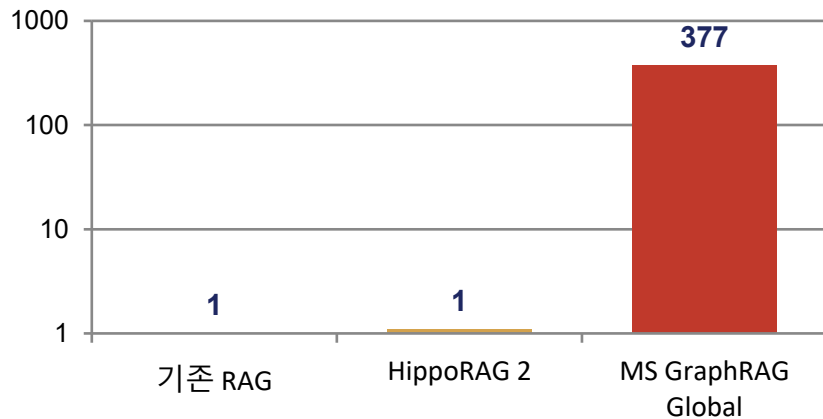
ICLR 2026 기준

성능 향상과 토큰 비용의 균형점

복합 추론 성능 (Complex Reasoning)



토큰 비용 (Token Cost, log scale)



"납기 지연 3번 이상 난 협력사 발주 건 중에, 지금 재고가 안전 재고 밑으로 떨어진 품목 있어?"

"지난 6개월간 설비 고장 직전에 반복된 알람 패턴이 지금 돌아가는 라인에도 있어?"

"작년에 놓친 딜들, 그 시점에 우리 팀 엔지니어 투입 시간이랑 상관관계 있어?"

"해지한 고객들 해지 3개월 전에 뭐가 달라졌어? 지금 그 패턴 보이는 고객 있어?"

"이 환자랑 나이·진단·수술 이력 비슷한 환자들 어떻게 치료했고 결과 어땠어?"

RAG, GraphRAG로 AI가 우리가 원하는 답을 줄 수 있는가?

RAG 중심의 AI 활용은 비즈니스 프로세스 핵심에 도달하지 못한 채, 주변 비정형 문서의 검색 및 취합 수준에만 제한되어 있음

RAG가 닿는 영역

비정형 데이터

- PDF 보고서 및 매뉴얼
- 워드/한글 업무 문서
- 이메일 및 메신저 대화록

RAG가 닿지 못하는 영역

정형 & 핵심 기간계 데이터

ERP 전사 자원 관리

CRM 고객 관계 관리

MES 생산 관리 시스템

DW 데이터 웨어하우스

수천 개의 관계형 테이블

정교하게 연결된 스키마

기업 실적, 매출, 재고 지표

실시간으로 변하는 핵심 데이터

수백 쪽짜리 매뉴얼을 요약/번역하고 키워드를 추출해 주는 AI보다,

"이번 달 공장 가동률과 재고 추이를 감안한 대체 부품 소요량을 ERP에서 추론해줘"

이 질문에 답하는 AI가 기업 AI의 **궁극적 지향점**입니다.

기존 데이터레이크 vs. AI-ready Data (AI Agent 가 이해하는 데이터)

요건	의미	예시 (건설사 데이터)	Medallion	AI-ready Data
Meaningful 의미 중심	비즈니스 언어로 엔티티와 속성이 정의된 상태	시스템 코드가 아닌 "현장", "공중", "자재"라는 비즈니스 개념으로 정의됨	✓ 커버가능	✓
Unified 통합된 의미	원천별 사일로 없이 단일 도메인 개념으로 통합	ERP·PMIS·구매 시스템에서 다르게 부르는 동일 현장·협력사를 하나의 개념으로 통합	✓ 커버가능	✓
Trusted 데이터 신뢰	데이터 출처와 품질이 보장된 상태	공사일보·자재 입고 기록의 출처 추적, 중복 제거, 이상값 제약 적용	✓ 커버가능	✓
Relationship 관계 인지	엔티티 간 연결의 의미가 선언된 상태	"프로젝트 → 공중 → 공정(작업)"으로 이어지는 시공 흐름이 데이터 차원에 선언됨	✗ 부족	✓ 핵심
Contextual 비즈니스 맥락	값이 발생한 상황과 의도가 함께 명시된 상태	같은 공기 지연이라도 설계 변경 때문인지 우천·자재 수급 문제 때문인지 그 맥락이 관계에 포함됨	✗ 부족	✓ 핵심

Graph와 Semantic Layer

기존 메달리온 인프라 위에 시맨틱 모델과 온톨로지만 있어 완성하는 차세대 AI-Ready 아키텍처

온톨로지 그라운드링 도입 효과 Microsoft 내부 벤치마크

RAG Triad 기반 A/B 테스트 · 400개 질문 / 4,000개 응답 · 온톨로지 그라운드링 vs. 미적용

Operational Intelligence

Ontologies (자율 행동 및 다대다 인과 추론 지형 선언)

 Fabric IQ

Business Intelligence

Semantic Models (Synonym 정의어 및 공식 계산 메트릭 바인딩)

 Power BI

+

Unified Data

Structured, Unstructured, Real-Time, Graph (기존 메달리온 원천 보존)

 OneLake

2.2배

Excellent 답변 비율
최고 등급 응답 2.2배 증가

+4.5p

컨텍스트 관련성
관련성 점수 4.5포인트 향상

4.5배

1:1 비교 승률
나란히 비교 시 4.5배 우위

30%

도구 호출 절감
에이전트 도구 호출 30% 감소

온톨로지 미적용

잘못된 경로를 탐색하다 정답에 도달하지 못함

온톨로지 적용

엔티티·관계라는 ‘지도’에서 출발해 정답까지 직행

기준	Ontology (OWL)	KG (RDF)	Neo4j (LPG)	Palantir	Fabric IQ Ontology
분류/유형	Class	rdf:type	Label	Object Type	Entity Type
개체 (실물)	Individual	Individual (ABox)	Node	Object Instance	Entity Instance
관계 (개체↔개체)	Object Property	IRI 속성 (Triple 술어)	Edge / Relationship	Link Type	Relationship (typed, directional)
속성 (개체→값)	Datatype Property	Literal	Property	Property	Property
식별자	IRI	IRI	업무식별자	내부 ID	Entity Type Key (업무 식별자 기반)
쿼리 언어	SPARQL (W3C 표준)	SPARQL (W3C 표준)	Cypher (→ ISO GQL)	OSDK (TypeScript/Python)	GQL (ISO 표준) + KQL · NL2Ontology

* RDF: Resource Description Framework · OWL: Web Ontology Language · LPG: Labeled Property Graph · IRI: Internationalized Resource Identifier
* SPARQL: W3C 표준 RDF 쿼리 언어 · Cypher: Neo4j 고유 언어 · GQL: ISO 39075 표준 그래프 쿼리 언어 (2024)
* Palantir OSDK: 전용 SDK 방식 (별도 그래프 쿼리 언어 없음, API 기반 접근) · Fabric IQ: GQL for Graph, KQL for Eventhouse, NL2Ontology 자동 라우팅
* Fabric IQ Ontology 출처: Microsoft Learn 공식 문서 (2025.11 Public Preview 기준)

왜 Microsoft는 'LPG + 온톨로지'로 지식 그래프를 만들었나

그래프도 목적에 따라 방식이 다르다 — 의미는 온톨로지로 정의하고, 실행은 LPG로 빠르게

Fabric의 선택

온톨로지 (Ontology) — 의미를 정의

- 현장·공종·협력사·자재 등 업무 용어를 한 곳에 정의
- 엔터티·관계·규칙을 선언
- 사람과 AI 에이전트가 같은 개념으로 일함

+

LPG (Labeled Property Graph) — 관계를 빠르게 실행

- 노드와 관계에 속성을 직접 부착 (예: 투입기간·인원)
- 여러 단계를 잇는 다단계(Multi-hop) 탐색에 최적
- 대규모 분석·운영 환경의 성능

=

지식 그래프

Knowledge Graph

업무 의미 + 실행 성능

→ 기업과 AI 에이전트가 실제로 쓰는 지식 그래프

다른 방식과 비교

RDF / 트리플 스토어

방식	주어-술어-목적어 트리플로 저장
강점	형식 의미론, OWL 기반 자동 추론
한계	구조가 복잡하고 대규모 운영 시 성능 부담
적합	규제·법령 검증, 공공 Linked Data

관계형 테이블 (SQL JOIN)

방식	외래키와 JOIN으로 관계 표현
강점	익숙한 SQL, 별도 저장소 불필요
한계	관계가 깊어질수록 JOIN 폭증·성능 저하
적합	단순 조회·집계

건설사 예시

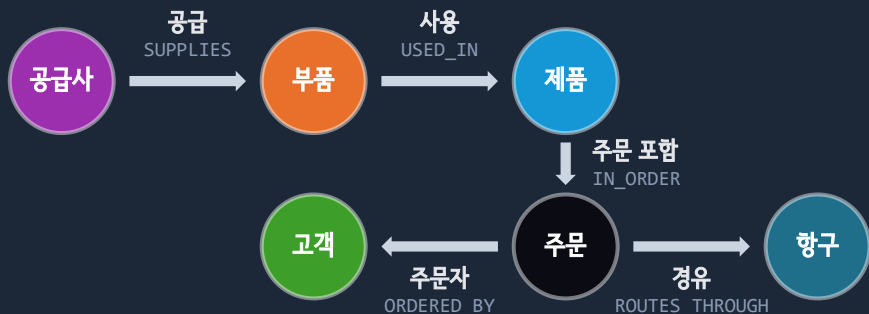
“레미콘 협력사 공급이 멈추면 어느 현장·공정·발주처까지 영향을 받나?”

→ 온톨로지가 경로의 '의미'를 알려주고, LPG가 그 경로를 빠르게 따라간다

공급망에서의 Graph

비즈니스 질문

“공급사나 항구 배송이 지연되면, 어떤 주문과 고객이 영향을 받나?”



같은 질문을 SQL로 쓰면

```
SELECT DISTINCT o.order_id, c.name
FROM Suppliers s
JOIN Parts p ON s.supplier_id = p.supplier_id
JOIN Products pr ON p.part_id = pr.part_id
JOIN OrderItems oi ON pr.product_id = oi.product_id
JOIN Orders o ON oi.order_id = o.order_id
JOIN Customers c ON o.customer_id = c.customer_id
WHERE s.supplier_id = 'S-123';
```

문제점:

- JOIN 폭증 — 단계가 깊어질수록 JOIN이 계속 늘어남
- 연결용 중간 테이블(OrderItems 등)이 곳곳에 필요
- 대용량 데이터에서 성능 저하
- 단계 수가 정해지지 않은 N단계 탐색은 일반화 불가

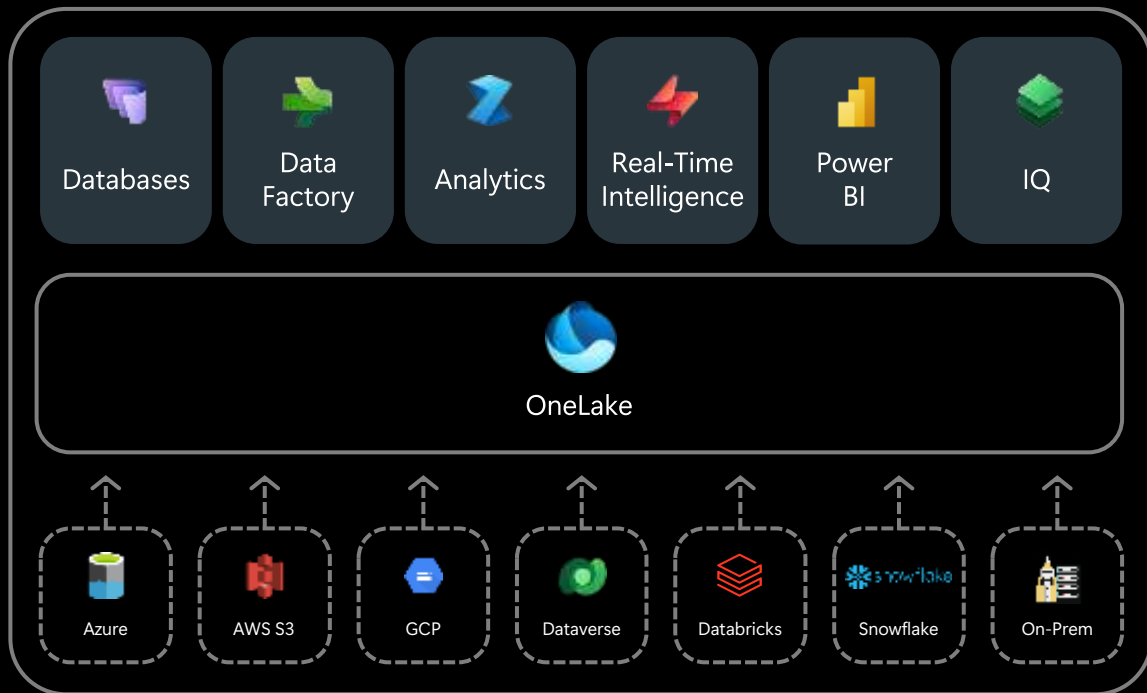
GQL 쿼리

```
MATCH (s:Supplier {id: 'S-123'})
-[:SUPPLIES]->(:Part)-[:USED_IN]->(:Product)
-[:IN_ORDER]->(o:Order)-[:ORDERED_BY]->(c:Customer)
RETURN o.id, c.name
```

관계 경로를 그대로 한 줄 패턴으로 표현

OneLake로 데이터 통합

클라우드 간 Shortcut과 Mirroring



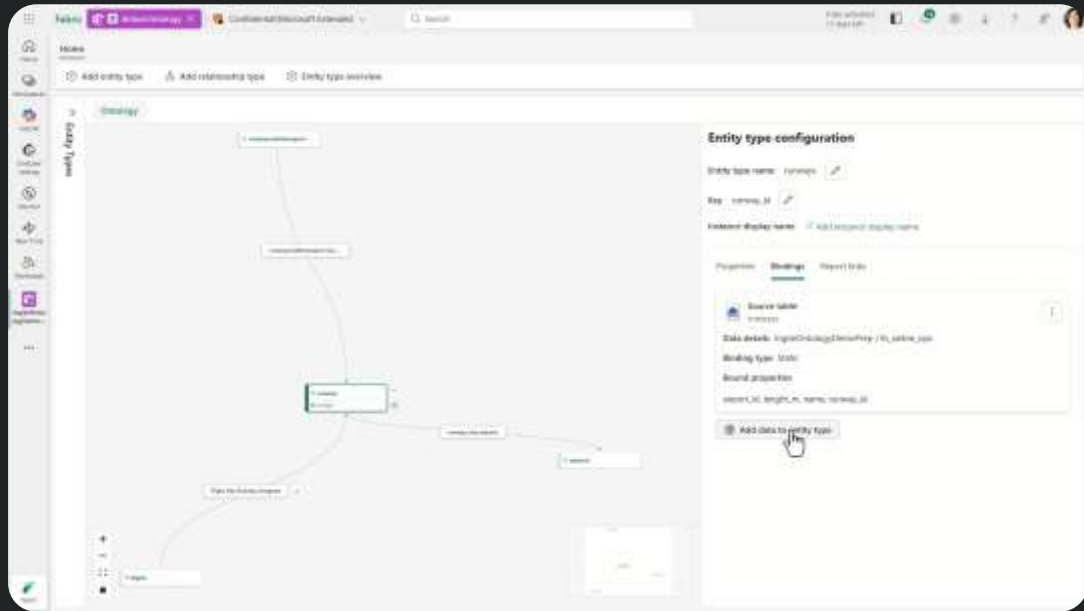
- 최초의 멀티 클라우드 SaaS 데이터 레이크
- Shortcut으로 여러 클라우드와 온프레미스의 데이터를 제자리에서 연결
- Shortcut은 즉시 연결 — 데이터 복제나 이동 없음
- Mirroring으로 데이터베이스 전체를 OneLake에 그대로 반영



Ontology in Fabric IQ

Ontology로 비즈니스의 의미와 흐름을 모델링하세요

Public Preview



비주얼 기반의 **no-code** 도구로
비즈니스 온톨로지를 정의합니다

비즈니스 개념, 프로세스, 현실 세계의
관계를 포함한 **온톨로지**를 구축합니다

OneLake의 **정형 · 실시간 · 공간**
데이터를 온톨로지와 결합합니다

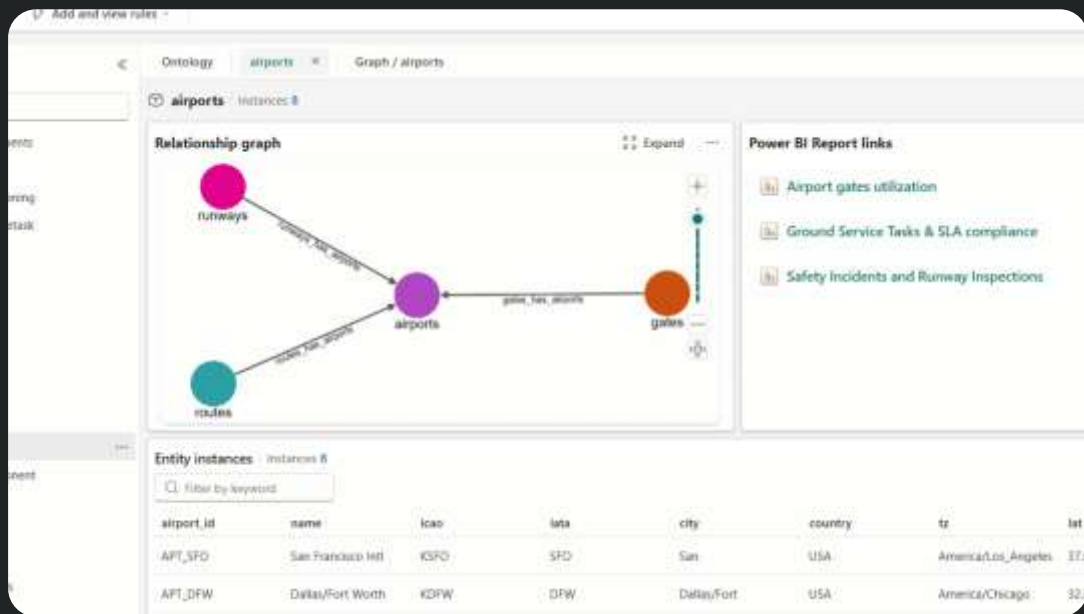
2천만 개 이상의 **semantic model** 을
기반으로 온톨로지 구축을 가속화합니다



Ontology-Powered Graph in Fabric IQ

그래프 분석을 통해 총체적인 인사이트를 확보하세요

Public Preview



프로세스, 사용자, 공급망 전반에 걸친 패턴과 연쇄 효과를 발견합니다

전통적인 BI 도구로는 포착하기 어려운 기회와 리스크를 탐지합니다

Fabric Graph 기반의 거버넌스 그래프를 손쉽게 활용할 수 있습니다

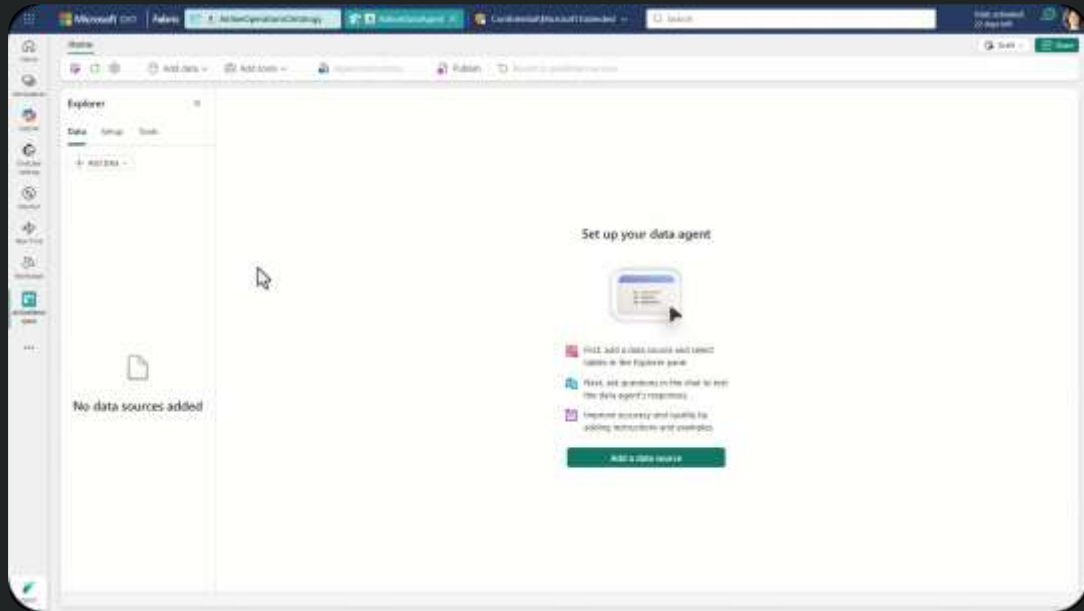
데이터 사일로를 넘어, 전체 시스템 관점에서의 의사결정이 가능합니다



Data Agent in Fabric IQ

Ontology 기반의 강화된 Data Agent를 만나보세요

Public Preview



자연어로 질문하고, 비즈니스 맥락에
적합한 답을 얻을 수 있습니다

단순 스키마가 아닌, 온톨로지 기반의
개념과 관계를 중심으로 추론합니다

변화하는 비즈니스 상황에 맞춰,
자동으로 답변을 고도화합니다.

하나의 semantic foundation 위에서
AI 에이전트가 동작함으로써, 조직 전반을
아우르는 인사이트를 도출합니다



AI가 전통적인 데이터레이크 기반으로 답변할 때

Data Agent 시나리오 - 전통적인 데이터 모델

그러면 현재 재고가 부족한 상품이 있어?

현재 재고가 부족한(재고 수량이 낮거나 0에 가까운) 상품은 조회 결과 없습니다.
즉, 모든 상품이 충분한 재고를 보유하고 있거나, 최소 임계값 이하로 부족한 제품은 없는 상태입니다.

추가로 궁금한 점이 있으면 말씀해 주세요!

1 step completed ^

🕒 Response time: 65 sec

⊗ Analyzed IQ_Model_02 SemanticModel for: "현재 재고가 부족한(재고 수량이 낮거나 0에 가까운) 상품이 있는지 알려주세요."

Semantic Model 활용 ver.

"재고 부족"은 현재 보유 수량 대비 예약된 수량으로 판단해줘. 이렇게 판단하면 재고가 부족한 상품이 있을까?

"현재 보유 수량 - 예약된 수량" 기준으로 가용 재고가 0 이하인, 즉 실제로 재고가 부족한 상품이 존재합니다.

- Product 102: 가용 재고 -15 (보유 31, 예약 46)

이 제품은 예약된 수량이 보유 수량을 초과하며 재고 부족 상태입니다. 필요한 조치(추가 발주 등)가 요구됩니다.

1 step completed ^

🕒 Response time: 46 sec

✔ Analyzed IQ_Model_02 SemanticModel for: "현재 보유 수량에서 예약된 수량을 뺀 '가용 재고'가 0 이하인 상품이 있는지 알려주세요."



Data Agent in Fabric IQ

Data Agent 시나리오 - Ontology의 맥락 기반 추론

현재 재고가 부족한 상품이 있습니다.

- 상품명: Product 102
- 브랜드: 에스쁘아
- 카테고리: makeup
- 실제 보유 수량(on_hand): 31
- 예약 수량(reserved): 46

따라서, 예약된 수량이 실제 보유하고 있는 수량보다 많아서 재고가 부족한 상황입니다.

1 step completed ▾

그러면 현재 재고가 부족한 상품이 있어?



Analyzed IQ_Ontology_02 Ontology for: "현재 재고가 부족한 상품이 있는지 확인해줘. 재고 부족의 기준은 실제 보유 수량이 예약 수량보다 적은 경우야."

</> Query code

Ontology

```
{ "entitySelector": { "query": "MATCH (node_Product:Product)-[edge_has_state:has_state]->(node_InventoryState:InventoryState)\nWHERE node_InventoryState.on_hand_qty < node_InventoryState.reserved_qty\nRETURN node_Product.product_id AS product_id, node_InventoryState.inventory_id AS inventory_id, node_InventoryState.on_hand_qty AS on_hand_qty, node_InventoryState.reserved_qty AS reserved_qty, TO_JSON_STRING(node_Product) AS product_json, TO_JSON_STRING(node_InventoryState) AS inventory_state_json", "queryType": "GQL" }
```

Query output

Ontology 활용 ver.



Data Agent in Fabric IQ

Data Agent 시나리오 - Ontology의 맥락 기반 추론

해당 제품의 재고 부족 이유가 뭐야? 판매가 많이 돼서 그런거야, 아니면 납품이 지연돼서 그런거야?

조사 결과, 해당 제품(P-102)의 재고 부족 사유는 다음과 같이 파악됩니다.

1. 판매 실적:

- 최근 판매 이벤트(2026-01-12)에 5개가 판매되었습니다.
- 전체적으로 판매 이벤트 수가 많지는 않은 것으로 보입니다.

2. 납품(입고) 정보:

- 최근 입고(납품) 예정일은 2026-01-28이었으나, 실제 입고는 2026-01-31에 383개가 들어왔습니다.
- 예정일보다 납품이 3일 지연되었습니다.

요약: 재고 부족의 주된 원인은 '납품(입고)의 지연'으로 보입니다. 판매가 급증해서 생긴 부족이라기보다는, 예약 수량이 큰 영향이 큰 것으로 판단됩니다.

2 steps completed

Analyzed IQ_Ontology_02 Ontology for: "상품 ID가 P-102인 제품의 최근 판매량, 판매 이벤트 수, 판매 이벤트 발생 시점 정보를 알려줘."

Query code

Ontology

```
{ "entitySelector":  
  { "query": "MATCH (node_SalesEvent:SalesEvent)\nWHERE LOWER(node_SalesEvent.`product_id`) = LOWER(\n`P-
```

Ontology 활용 ver.

Analyzed IQ_Ontology_02 Ontology for: "상품 ID가 P-102인 제품에 대한 최근 입고(납품) 주문 정보와 실제 입고일, 예정 입고일, 입고 수량 정보를 알려줘."

Query code

Ontology

```
{ "entitySelector": { "query": "MATCH (node_PurchaseOrder:PurchaseOrder)-  
[edge_fulfilled_by:fulfilled_by]->  
(node_Delivery:Delivery)\nWHERE LOWER(node_PurchaseOrder.`product_id`) = LOWER(`P-
```

Select data from Fabric to use in your Graph



All



Endorsed in your org



My data



Favorites

Filter by keyword

Filter

Explorer

	Name	Type	Owner	Location	Endorsement
	C360	Lakehouse	Justin Fine	My Workspace	—
	FraudDemo	Lakehouse	Justin Fine	My Workspace	—
	C360_Demo	Lakehouse	Justin Fine	My Workspace	—
	C360_Demo_Data	Lakehouse	Justin Fine	My Workspace	—
	AddressTable	Semantic model	Sreramam Narasimhan	SQL DB Native Bug ...	—
	LDBC_SF3_600mb_data	Lakehouse	Yuchen Wang	TestDataSources_DO...	—
	LDBC_SF1	Lakehouse	Shashwathi Nagaraj...	FabricGraphBugBash...	—
	Import	Lakehouse	Justin Fine	My Workspace	—
	ADW_DEMO_DATA	Lakehouse	Justin Fine	My Workspace	—
	graphinstance_3eace2764b1b47...	Lakehouse	Justin Fine	My Workspace	—
	a1	Lakehouse	Gila Cohen	SQL DB Native Bug ...	—

Add

Cancel

Ontology Preview에서 얻은 교훈

+ 핵심 변화

Ontology: Graph 적재 기반 → 복사 없는 공유 의미 계층

✓ new experience

1 생성 자동화

- AI 기반 ontology 초안 생성
- Power BI measure → metric
- Agent 제안 → 검토 후 적용

2 실행 분리

- Graph는 선택(opt-in)
- 원본 엔진 직접 조회
- Keyless entity · 직접 관계

3 재사용 확대

- 에이전트 · 대시보드 · MCP
- 원본 보안(RLS·OLS·CLS) 적용
- RDF · OWL 연동 · 버전 관리

무엇을 배웠고, 왜 바꿨나

배운 점

수작업 모델링의 도입 장벽

→ 기존 Power BI 자산에서 시작

배운 점

전체 graph 적재의 비용 · 최신성 부담

→ 의미 정의와 실행 엔진 분리

배운 점

에이전트별 정의 · 규칙 중복

→ 한 번 정의, 여러 곳에서 재사용

+ New item

New folder

Import

Migrate

Filter by keyword

Filter



Choose from predesigned task flows or add a task to build one

Select from one of Microsoft's predesigned task flows or add a task to start building one yourself.

Select a predesigned task flow

Add a task

Import a task flow

	Name	Status	Type	Task	Owner	Refreshed	Next refresh	Endorse...	Sensitivity	Included in app
	Notebook		Folder	—	—	—	—	—	—	
	da_A_tables		Data agent	—	MOD Admini...	—	—	—	—	
	da_B_prompt		Data agent	—	MOD Admini...	—	—	—	—	
	da_C_Ont		Data agent	—	MOD Admini...	—	—	—	—	
	lh_zeb		Lakehouse	—	MOD Admini...	—	—	—	—	
	lh_zeb		SQL analytics...	—	MOD Admini...	—	—	—	—	
	ont_zeb_certification		Ontology	—	MOD Admini...	—	—	—	—	

Explorer

Overview

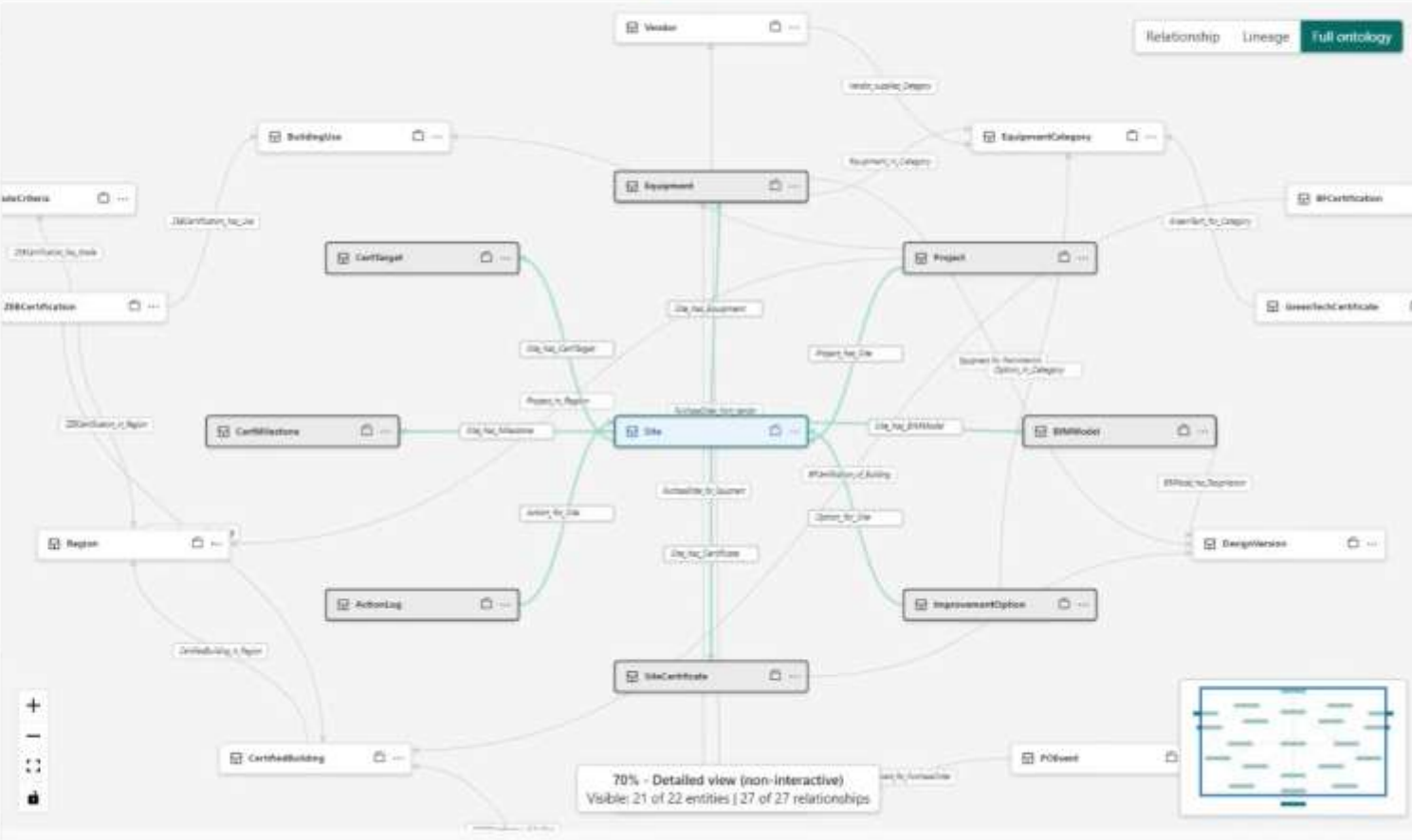
- Rules
- Shared properties

Ontology

Filter by name

Entity Types

- Region
- BuildingUse
- ZEBGradeCriteria
- EquipmentCategory
- CertifiedBuilding
- ZEBCertification
- GSEEDCertification
- BFCertification
- GreenTechCertificate
- Project
- Site**
- BIMModel
- DesignVersion
- SiteCertificate
- CertTarget



맥락

의사결정

실행



Intelligence layer



Agents

Microsoft IQ 플랫폼

AI를 위한 통합 인텔리전스



Work IQ

우리 조직이
일하는 방식

사람, 협업, 업무 흐름에 대한
맥락



Fabric IQ

우리 비즈니스가
운영되는 방식

비즈니스 엔터티, 기간제
시스템(ERP·PMIS 등),
실행(액션)에 대한 맥락



Foundry IQ

에이전트가 검증된
지식을 활용하는 방식

정책, 공식 문서, 지식 베이스에
대한 맥락



Web IQ

웹 인텔리전스를 활용하는
방식

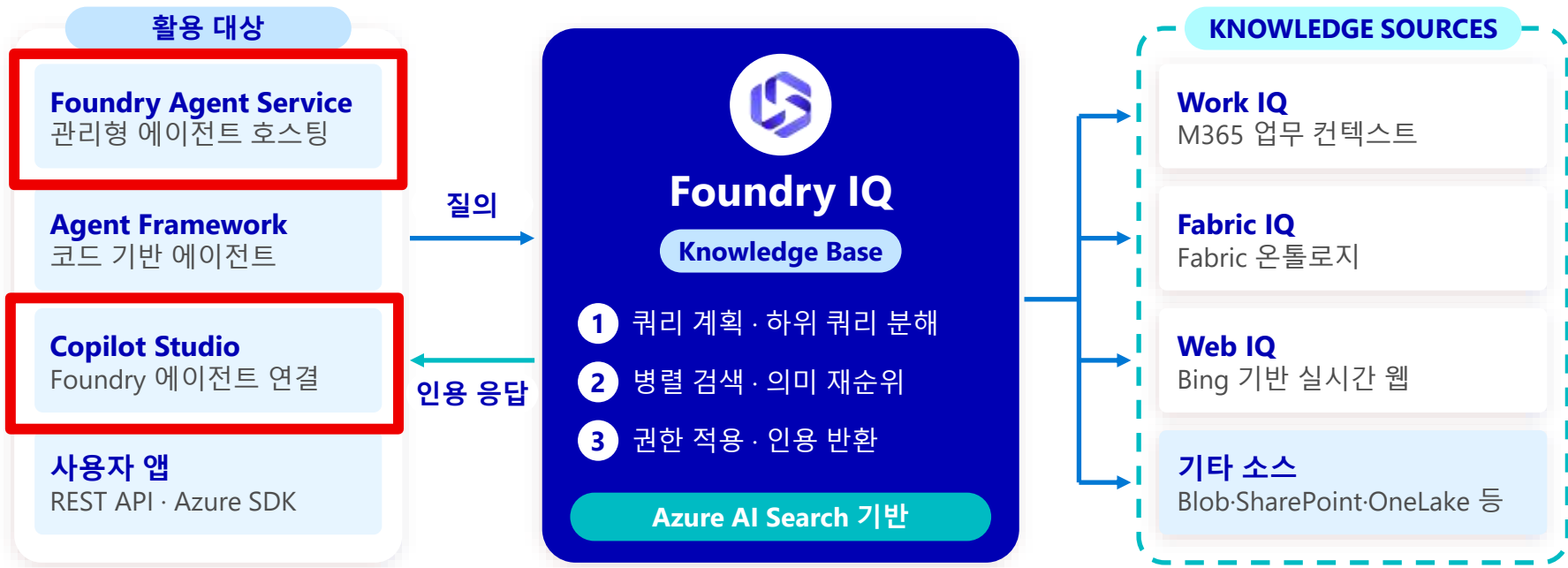
외부 리서치, 규제, 공개 웹
지식에 대한 맥락

aka.ms/MicrosoftIQ

Foundry IQ 구조

+ Foundry IQ

에이전트 질의 → Knowledge Base가 각 Knowledge Source 병렬 검색 → 인용 응답



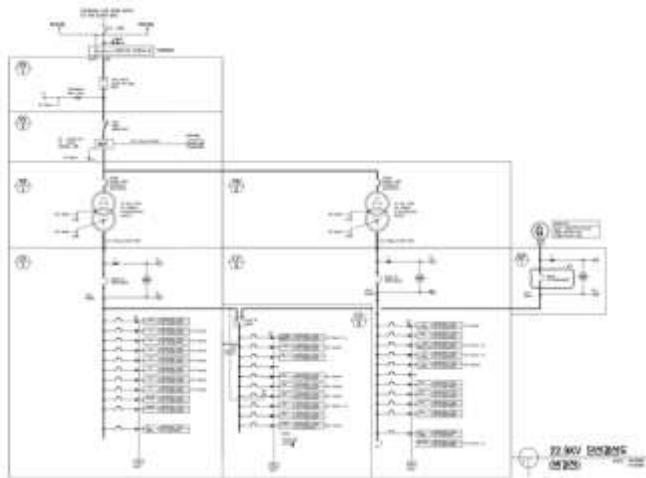
04

건설사에 활용 가능한 Lessons Learned 공유

도면 분석 사례의 교훈 · Ontology 사례 ·
건설사 데이터로 본 Ontology 효과

도면이미지에서, 패널별 자재 목록으로

기기의 연결과 사양을 읽고, 같은 패널에 속하는 자재를 하나의 BOM으로 묶습니다.



사용자 제공 대체 샘플 ① · 실제 PoC 평가 입력이 아님

입력

데이터센터 수변전설비 PDF

CAD PXF 등 객체 구조정보 없이
그림과 활용 가능한 선·글자에서 출발

추출 단위

그룹 → 패널 → 자재

각 패널에 어떤 자재가
어떤 사양·수량으로 있는가

산출물

패널별 BOM

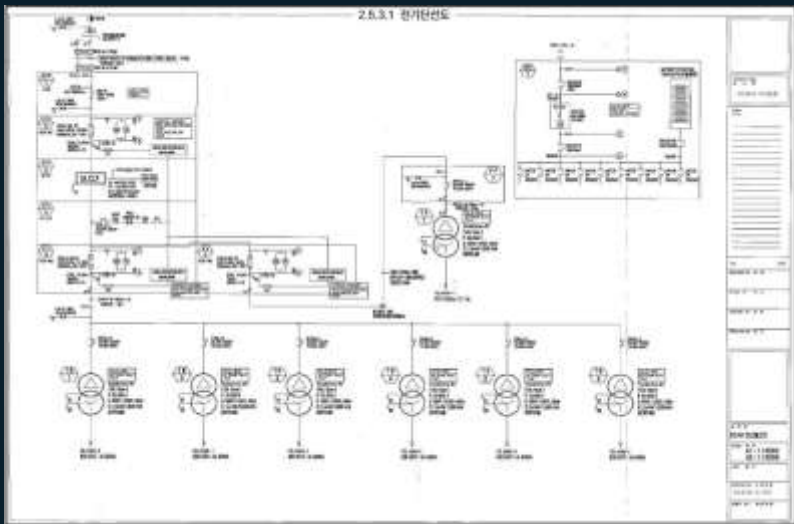
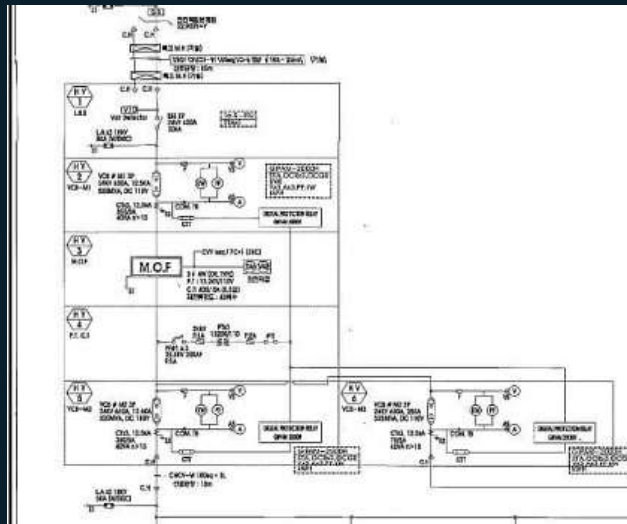
자재명 · 사양 · 수량을
정답지와 대조할 수 있는 목록

패널 단위로 읽고, 전체 도면과 다시 대조합니다

확대해서 읽되, 패널 이름과 위치를 함께 유지해야 결과를 올바르게 모을 수 있습니다.

전체 보기 · 위치와 계통의 맥락

부분 확대 · 글자와 기호 확인



분할 → 추출 → 수합 이후에도 누락·중복과 패널 귀속을 다시 확인합니다.

정확도는 높아지고, Reasoning 모드 차이는 줄었습니다

Task: 힌트 없이 박스 좌표를 JSON으로 반환 · 단순 흰 배경 · 모델별 5회 평균 IoU

5.2: high – none = +0.038

Reasoning none · GT와 예측 비교

5.4: -0.003 · 5.6/6 계열: 약 0



Reasoning high · GT와 예측 비교



모델	IoU · high	IoU · none	Δ (high–none)	best (high)
gpt-5.2	0.732	0.694	+0.038	0.813
gpt-5.4	0.823	0.826	-0.003	0.825
gpt-5.6-luna	0.954	0.954	±0	0.956
gpt-5.6-terra	0.956	0.956	-0.001	0.956
gpt-5.6-sol	0.954	0.954	±0	0.954
gpt-6-astra	1.000	1.000	±0	1.000

모델별 IoU · 리즈닝 high



이 단순 박스 과제의 관측입니다. 실제 도면에서도 같은 효과가 있는지는 별도 평가해야 합니다.

공통 지침에 도면 규칙과 이미지 예시를 더합니다

이 실험에서는 L3(지침 + 도면 노트 + 참조 이미지)를 채택했습니다.

L1-L4 · 비교한 프롬프트 구성

레벨	프롬프트 구성	쉽게 말하면
L1	일반 지침만	"육각형 안 판넬 이름 다 찾아줘" — 도메인 힌트 없이 맨음
L2	L1 + 도면 노트	이 도면들의 명명 규칙을 알려줌 (예: LV CRA-8A 형식, 하이픈 규칙, J/B · CON · SF- 코드 등)
L3	L2 + 참조 예시 이미지	노트에 더해 "판넬 이름 박스가 이렇게 생겼다"는 주석 예시 그림 1장 첨부 (최종 채택)
L4	L1 + 예시 이미지 (노트 제외)	노트를 빼고 예시 그림만 — "노트 vs 예시 중 뭐가 중요한지" 가르는 대조군

L1: 일반 지침 → L2: 도면 노트 추가

L3: 참조 이미지 추가 · L4: 노트 제외

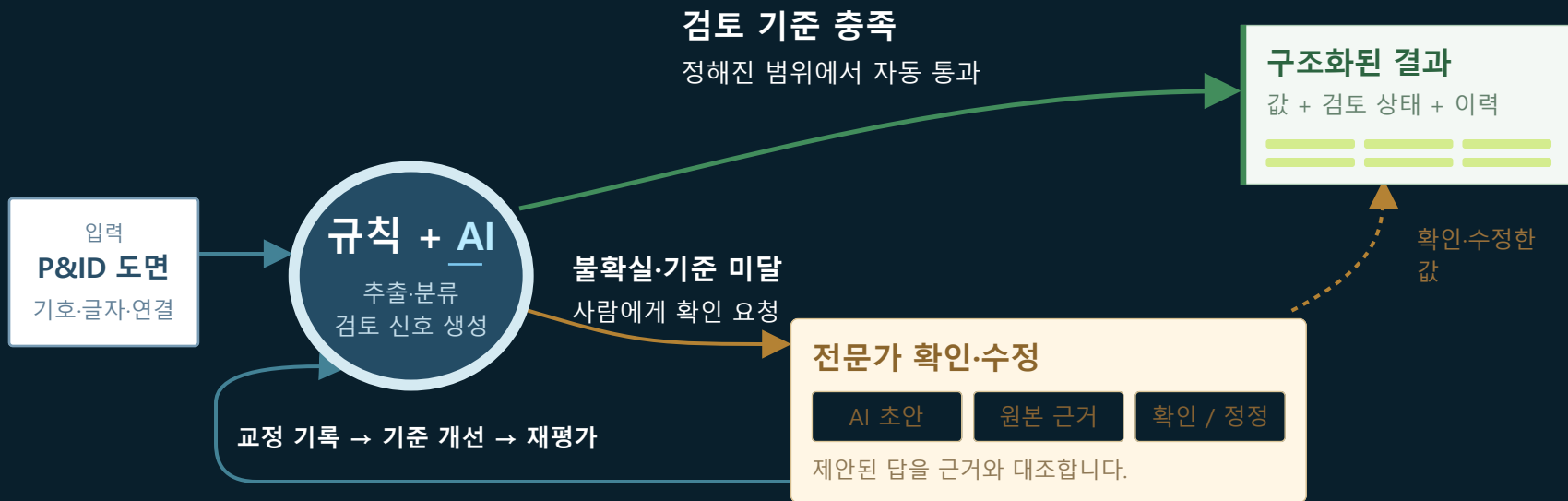
페이지별 검출 결과 · 원문 표

페이지	L1	L2	L3	L4
p1	4/20	10/20	18/20	6/20
p2	12/12(+오탐18)	10/12	12/12	12/12
p3	12/16	12/16	14/16	4/16
p4	0/17	11/17	17/17	0/17
p5	0/26	22/26	24/26	—

예시 이미지는 규칙 설명을 대체하기보다 보완합니다. 새로운 도면에는 다시 평가가 필요합니다.

AI의 초안을, 사람이 근거를 보고 확정합니다

초기에는 전수 검토합니다. 검증된 기준과 위험 수준에 따라 검토 범위를 조정합니다.



개념도 · 경로의 굵기는 처리 비율이 아닙니다. 초기에는 전수 검토하며, 교정 저장에 곧 모델 재학습은 아닙니다.

HITL의 목적: 이번 실행의 오류를 확인·수정하고, 확정값·근거·검토 이력을 남기는 것.

태스크마다 정확도·시간·비용의 균형이 다릅니다

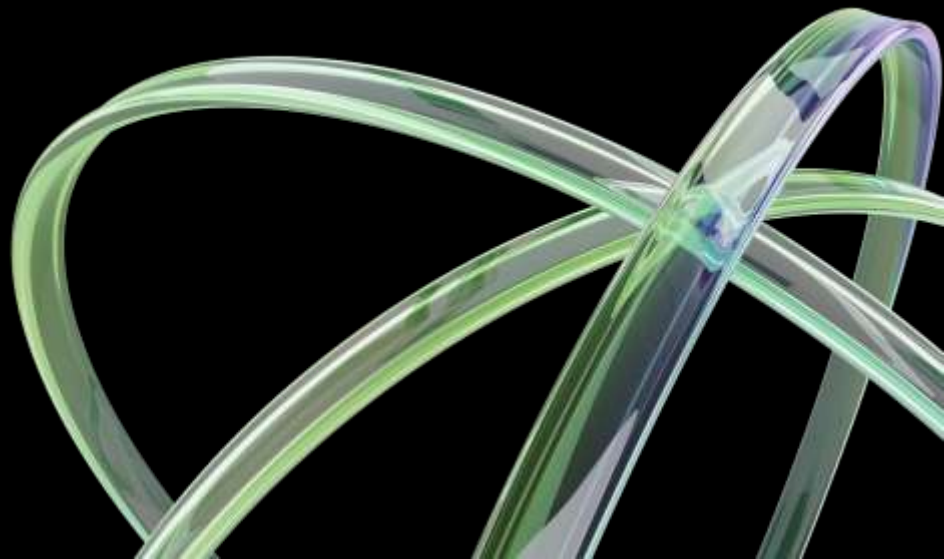
도구를 고정하기보다, 해당 도면의 조건과 측정 결과에 맞춰 조합합니다.

태스크	검토할 도구	강점	주의할 조건
텍스트·좌표	Azure DI / OCR	글자와 위치 추출	축약어·작은 글자 확인
선·경계	규칙 / SVG 기하	재현 가능한 처리	벡터 유무·도면 관례
이름·의미	VLM	문맥·시각 정보 해석	오류·비용·지연
최종 확정	검증 규칙 + HITL	정합성·전문가 판단	검토량과 미검출 오류

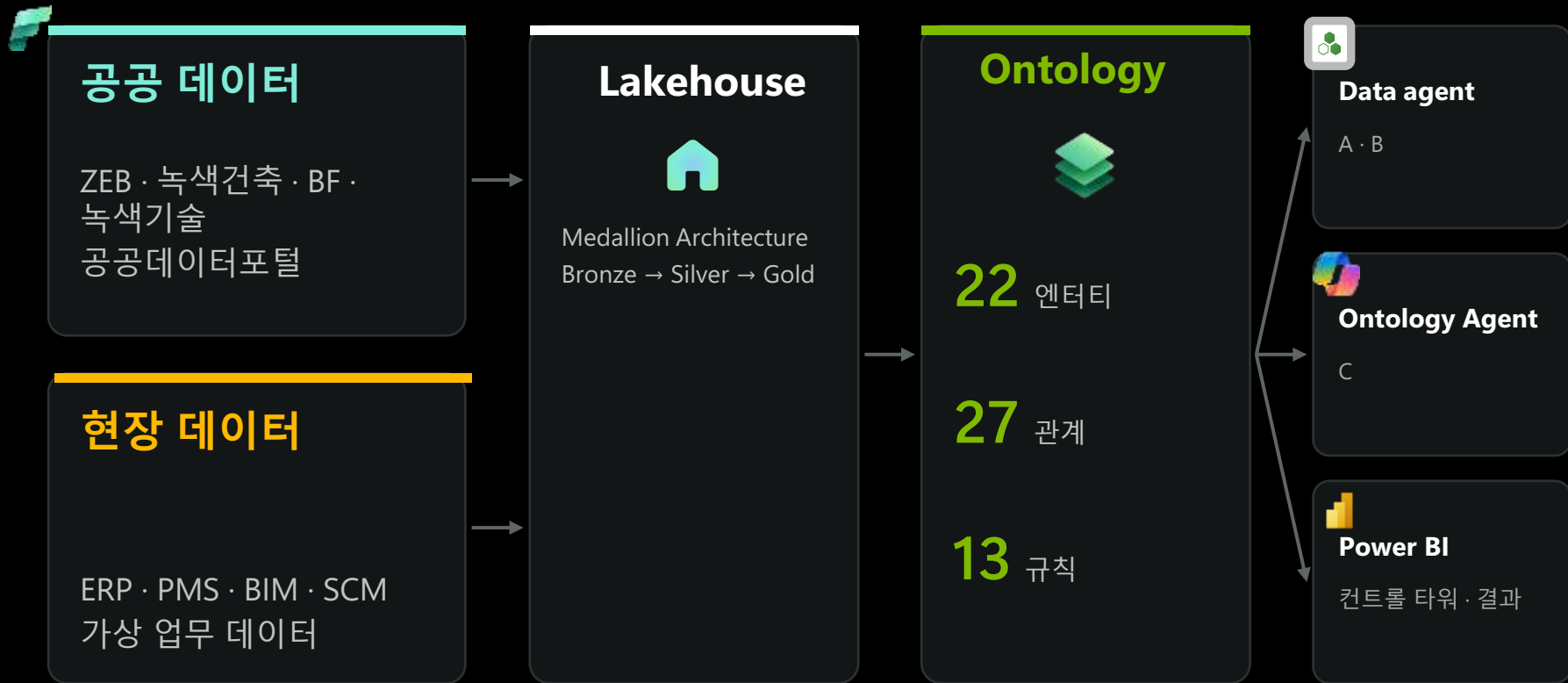
아래는 도구 선택 기준입니다. 실제 채택 여부는 해당 태스크의 품질·지연·비용으로 결정합니다.

Ontology가 AI Agent의 업무 이해, 지식 검색 및 추론에 미치는 역할과 효과

- 기존 AI 지식검색 방식 대비 Ontology 적용 시 정확도·신뢰도 및 업무 수행 수준의 변화



건설사 기반 데이터 구성



제로에너지건축물(ZEB) 인증 현황

한국에너지공단 공공데이터 9,760건(2025-12-31 기준) · 등급 = 에너지자립률 기준과 1차에너지소요량 기준 중 높은 등급(+ BEMS 필수)

9,760

누적 ZEB 인증

2,732

2025년 인증

+23.0%

전년 대비

45.1%

4등급 이상 비율

시도별 ZEB 인증

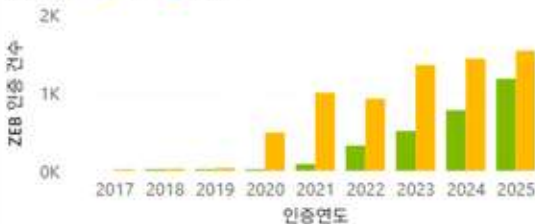
원 크기 = 인증 건수 · 경기 1,897 · 경남 870 · 경북 828



연도별 인증

예비인증(실제 단계)과 본인증(준공 후)

인증 구분 ● 본인증 ● 예비인증



등급별 인증

+ > 1 > 2 > 3 > 4 > 5 · '4등급 이상' = +1·2·3·4

인증 구분 ● 본인증 ● 예비인증



등급 기준

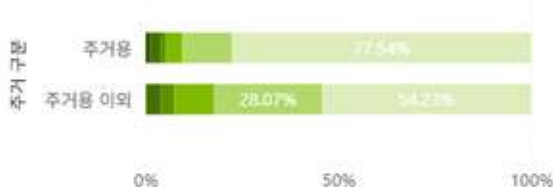
자립률 ≥ 하한 또는 1차에너지 < 기준이면 그 등급 → 높은 쪽

등급	자립률 하한(%)	1차에너지 주거	1차에너지 비주거
ZEB+	120	-10	-70
1등급	100	10	-30
2등급	80	30	10
3등급	60	50	50
4등급	40	70	90
5등급	20	90	130

주거용 vs 주거용 이외 등급 분포

주거용은 1차에너지 기준값이 다름

등급 ● ZEB+ ● 1등급 ● 2등급 ● 3등급 ● 4등급 ● 5등급



비교 구성

A 테이블만

Data agent

Gold 22개 테이블

업무 설명 없음

B 지시문·설명

Data agent

Gold 22개 테이블
에이전트 지시문
데이터 원본 지시문
테이블·열 설명

업무 규칙 13개

C 온톨로지

Ontology Agent

Fabric IQ 온톨로지
엔터티 22 · 관계 27
규칙 13

업무 규칙 13개

19문항 × 3회

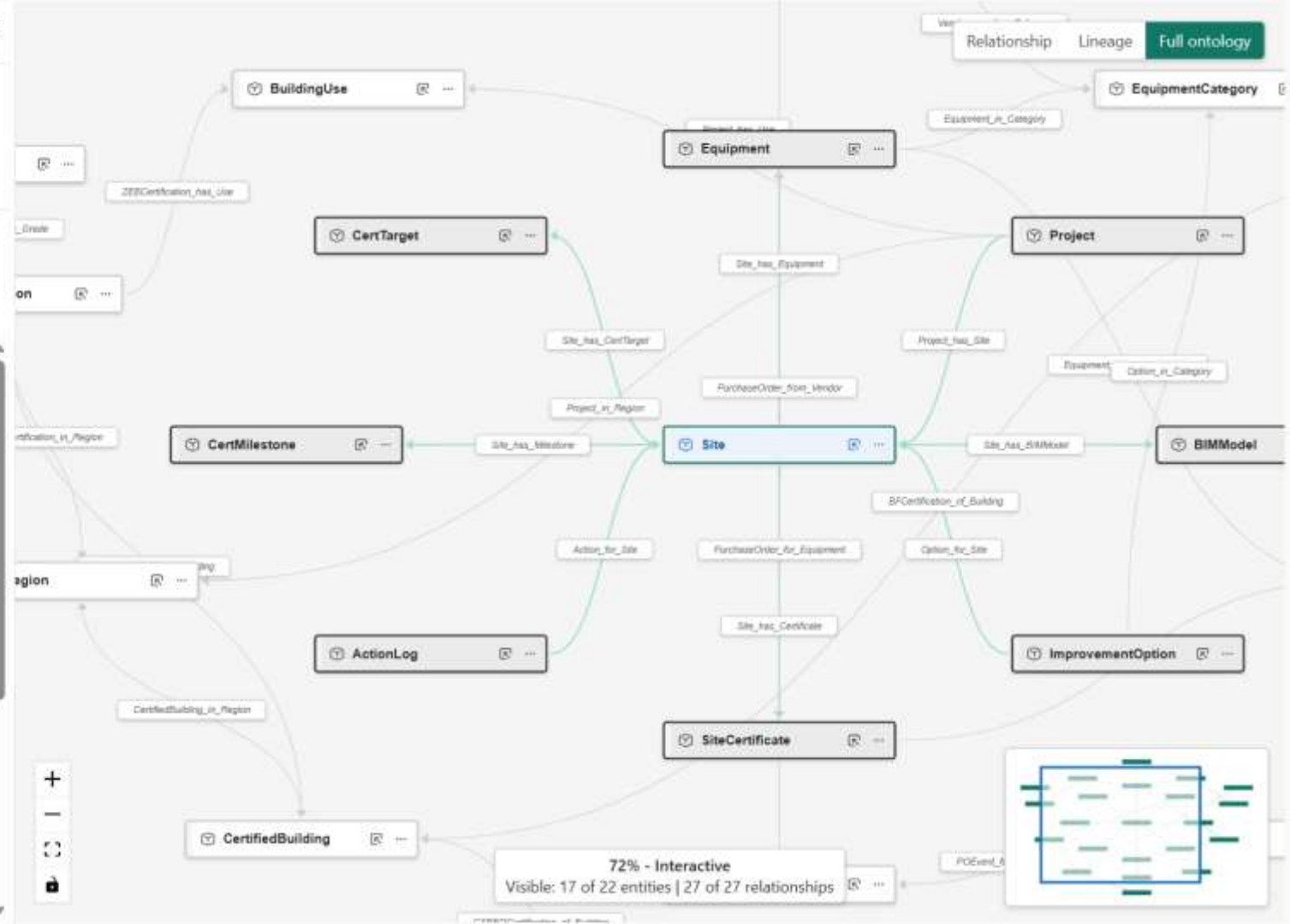
같은 질문 문장 · 질문마다 새 대화 · 핵심 수치·대상·결론 자동 채점

- Rules
- Shared properties

Q Filter by name

Region
BuildingUse
ZEBGradeCriteria
EquipmentCategory
CertifiedBuilding
ZEBCertification
GSEEDCertification
BFCertification
GreenTechCertificate
Project
Site

BIMModel
DesignVersion
SiteCertificate



① 단순 조회

Q1

“제로에너지건축물(ZEB) 인증이 가장 많은 시도 3곳과 건수를 알려줘.”

A 테이블만

X 0/3

경기 2건

자사 현장 인증서 집계

B 지시문·설명

✓ 3/3

경기 1,897건

경기 1,897 · 경남 870 · 경북 828

C 온톨로지

✓ 3/3

경기 1,897건

경기 1,897 · 경남 870 · 경북 828

정답 경기 1,897 · 경남 870 · 경북 828 (공공 ZEB 인증)

유형 정답률

A 50%

B 100%

C 100%

② 업무 용어

Q6

“공동주택(아파트) 중 ZEB 4등급 이상을 받은 건은 몇 건이야?”

A 테이블만

X 0/3

3건

자사 현장 인증서 집계

B 지시문·설명

✓ 3/3

43건

지시문의 용어 설명 적용

C 온톨로지

✓ 3/3

43건

엔터티 설명의 용어 적용

정답 43건 · '4등급 이상' = +1·2·3·4등급, 공공 인증 기준

유형 정답률

A 0%

B 100%

C 100%

③ 공공 데이터 × 현장 데이터

Q16

“전주 G도서관 현장의 현행 설계 에너지자립률은 **같은 시도·같은 용도의 공공 ZEB 인증 사례** 평균과 비교해 얼마나 차이 나?”

A 테이블만

X 0/3

0.0%p

비교군을 자사 현장 1건으로 집계

B 지시문·설명

✓ 3/3

-1.48%p

공공 44건 평균 62.48%

C 온톨로지

✓ 3/3

-1.48%p

공공 44건 평균 62.48%

정답 61.0% vs 전북 문화 및 집회시설 공공 ZEB 44건 평균 62.48% → 1.48%p 낮음

유형 정답률

A 17%

B 100%

C 83%

④ 규칙 적용

Q3

“리버뷰 B단지는 에너지자립률이 39%입니다. 지금 설계로 **ZEB 4등급**을 받을 수 있나요?”

A 테이블만

X 0/3

기존 예비인증
결과 인용

등급 판정 규칙 미적용

B 지시문·설명

X 0/3

받을 수 없음

자립률 40% 기준만 비교
1차에너지 기준 누락

C 온톨로지

✓ 2/3

받을 수 있음

1차에너지 $65 < 70 \rightarrow$ 4등급
두 기준 중 높은 등급

정답 받을 수 있음 · 자립률 39%(5등급), 1차에너지 $65 < 70$ (4등급) $\rightarrow$ 높은 쪽 4등급

유형 정답률

A 22%

B 33%

C 89%

⑤ 관계 추적

Q9

“에코시티 D단지의 **VE(원가절감) 설계 변경**이 ZEB 인증에 어떤 영향을 주나요?”

A 테이블만

X 0/3

리스크 증가

설계 수치 변화만 · 등급 판정 없음

B 지시문·설명

X 0/3

재검토 필요

규칙이 지시문에 있어도 미적용

C 온톨로지

✓ 3/3

4등급 → 5등급

예비인증 4등급 미달
본인증 위험

정답 4등급(43.5% · 69) → 5등급(38.2% · 76) · 예비인증 4등급 미달 → 본인증 위험

유형 정답률

A 50%

B 50%

C 83%

⑥ What-if

Q14

“ZEB 4등급의 주거용 1차에너지소요량 기준이 70에서 65로 강화되면, 예비인증을 받은 현장 중 본인증이 새로 위험해지는 곳이 있어?”

A 테이블만

X 0/3

A단지 · D단지

A단지는 자립률 44%로
4등급 유지

B 지시문·설명

X 0/3

A단지 · D단지

두 기준 중 높은 등급
규칙 미적용

C 온톨로지

✓ 3/3

리버뷰 B단지

1차에너지 65 · 새 기준 미달
자립률 39% · 5등급

정답 새로 위험: 리버뷰 B단지 | D단지 기존 위험 · A단지 4등급 유지(자립률 44%)

유형 정답률

A 50%

B 58%

C 92%

지식 기반에 따른 에이전트의 정답률 측정

30%

A 타이틀만

75%

B 지사문-설명

89%

C 온톨로지

171

채점한 답변 수

질문 유형별 정답률



질문별 정답 수(동일한 질문 3회 반복)

질문	유형	A	B	C	질문 내용
Q11	What-if	3	2	2	Northwind 자원의 미입고 발주가 며칠 넘게 늦어지면 인준 일정이 밀리는 현상이 생겨?
Q12	What-if	1	2	3	Northwind 자물 남품이 21일 늦어질 때, 에코시티 D단지 자물 히트맵프를 대체 공급사로 바꾸면 인준 일정을 지킬 수 있어?
Q13	What-if	2	3	3	에코시티 D단지가 VE 설계는 그대로 두고 보완안을 하나만 추가해서 예비인준 등급을 지키려면, 가장 큰 보완안은 뭐고 비용은 얼마야?
Q14	What-if	0	0	3	ZEB 4등급의 주거용 1차에너지소요량 가운데 70에서 65로 강해지면, 예비인준을 받은 현장 중 본인들이 새로 위험해지는 곳이 있어?
Q10	공공+현장	0	3	1	경기 지역 공동주택 중 ZEB 4등급 이상 사례는 몇 건이고, 평균 에너지자립률은 D단지 현장 설계와 비교해 얼마나 차이 나?
Q15	공공+현장	0	3	3	우리 현장이 있는 시도 중에서 공공 ZEB 인증이 가장 적은 곳은 어디고 몇 건이야? 그 시도에 있는 우리 현장도 알려줘.
Q16	공공+현장	0	3	3	전주 G도서관 현장의 현장 설계 에너지자립률은 같은 시도 같은 용도의 공공 ZEB 인증 사례 평균과 비교해 얼마나 차이 나?
Q17	공공+현장	0	3	3	세종 E학교 현장의 BF-녹색건축 인증 목표 등급을 확인하고, 세종 지역 교육연구시설의 공공 인증 데이터에서 그 목표 등급 이상을 받은 건에 BF-녹색건축 각각 몇 건인지 알려줘.
Q18	공공+현장	3	3	3	에코시티 D단지 현장에 계획된 설비 분야 중에서 지금 유효한 녹색기술 인증이 하나도 없는 분야는 뭐야?

LLM 응답 시간 및 소요된 토큰 비교

에이전트별 성능(답변 1건 기준)

시간: 질문 전송 → 답변 완료 - 캐시 비율: 입력 토큰 중 캐시에서 다시 쓴 비율

에이전트	정답률	평균 시간(초)	답변당 토큰	NL2SQL 호출/답변	캐시 비율	모델
A	30%	54	151,675	3.0	64%	gpt-5.6-terra-NL2SQL
B	75%	46	121,267	2.3	63%	gpt-5.6-terra-NL2SQL
C	89%	82	472,383		78%	gpt-5.4-mini-gpt-5-r

답변당 토큰 구성

A-B = 에이전트 모델(gpt-5.6-terra) + NL2SQL - C = 온톨로지 에이전트 모델



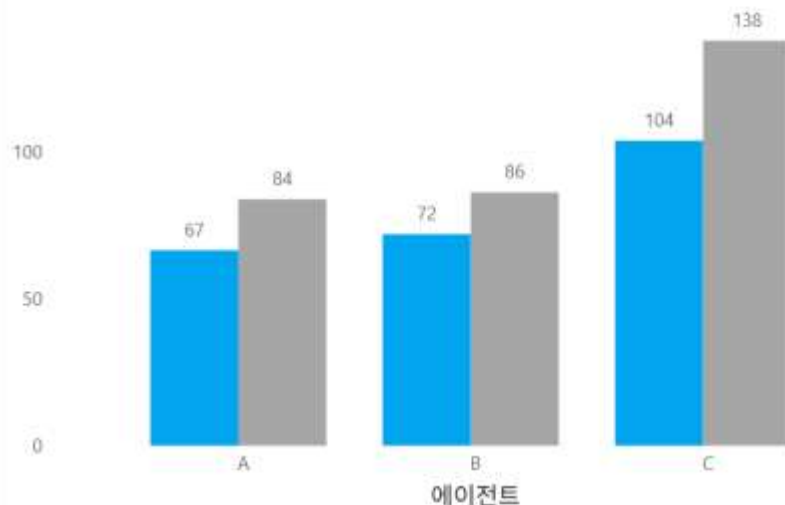
답변별 토큰 기록

질문	에이전트	회차	시간(초)	전체 토큰	에이전트 모델	NL2SQL-보조 모델	NL2SQL 호출	기록 파일
Q1	A	1	27	51,349	3,699	47,650	1	lh_zeb > Files > eval > token_logs > A > Q1_r1.json
Q1	A	2	30	51,645	4,020	47,625	1	lh_zeb > Files > eval > token_logs > A > Q1_r2.json
Q1	A	3	27	52,917	3,680	49,237	1	lh_zeb > Files > eval > token_logs > A > Q1_r3.json
Q1	B	1	19	21,612	4,008	17,604	1	lh_zeb > Files > eval > token_logs > B > Q1_r1.json

용량별 평균 응답 시간(초)

같은 What-if 4문항 × 3회

용량 ● F64 ● F8



마무리하며

01

Ontology 레이어

AI 시대에 데이터가 있는 곳에 Ontology 레이어는 필수

02

인텔리전스 레이어

정형/비정형 데이터 모두 AI가 접근/검색할 수 있는 인텔리전스 레이어 구축 필요

03

핵심 질의 중심 최적화

빅뱅 Ontology 구축보다 핵심 질의에 맞춘 데이터 및 온톨로지 레이어 최적화 필요

04

근거를 제공하는 에이전트

HITL을 위해 답변에 대한 근거를 제공하는 에이전트 개발 필요

Thank you

이정훈 | Solutions Engineers, Cloud & AI Platform | Microsoft