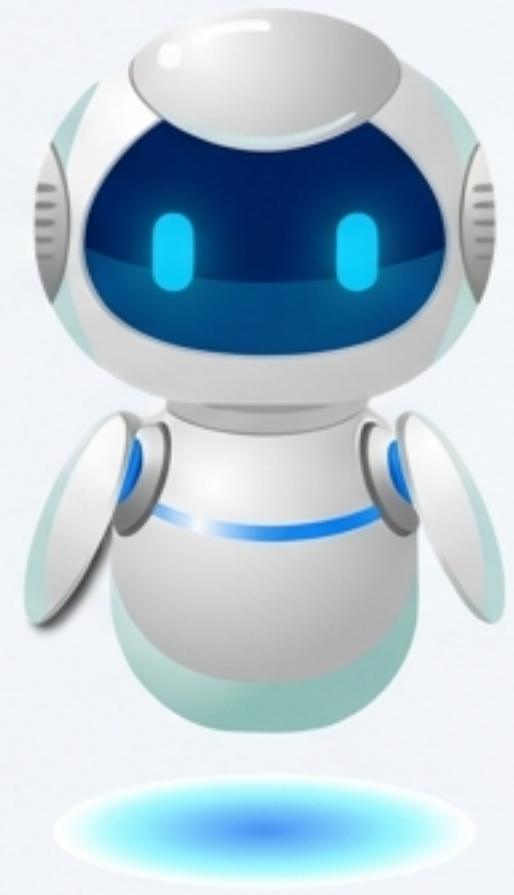


# LLM, 과연 데이터베이스 전문가를 대체할 수 있는가?



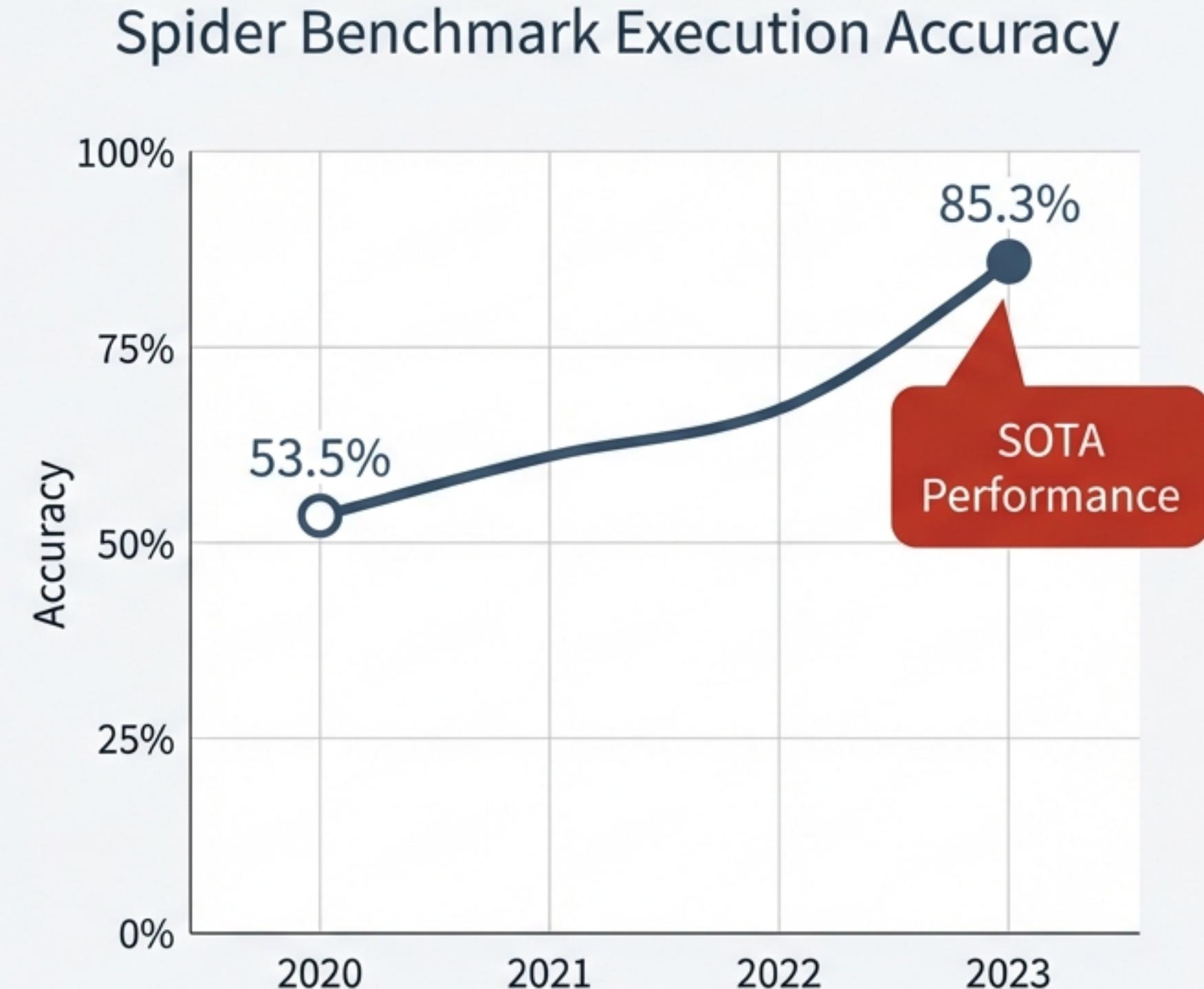
대규모 실제 데이터베이스 기반 벤치마크 ‘BIRD’를 통한  
LLM의 Text-to-SQL 성능 현실 점검 (NeurIPS 2023)



본 리포트는 NeurIPS에 등재된 논문 "Can LLM Already Serve as A Database Interface? A Blg Bench for Large-Scale Database Grounded Text-to-SQLs"를 기반으로 작성되었습니다.

# Text-to-SQL의 황금기처럼 보였던 지난 3년

- Text-to-SQL 기술은 비전문가도 자연어로 데이터를 추출할 수 있게 하는 핵심 기술입니다.
- 최근 LLM의 발전으로 Spider 등 기존 벤치마크에서 SOTA 모델의 성능이 급격히 향상되었습니다.

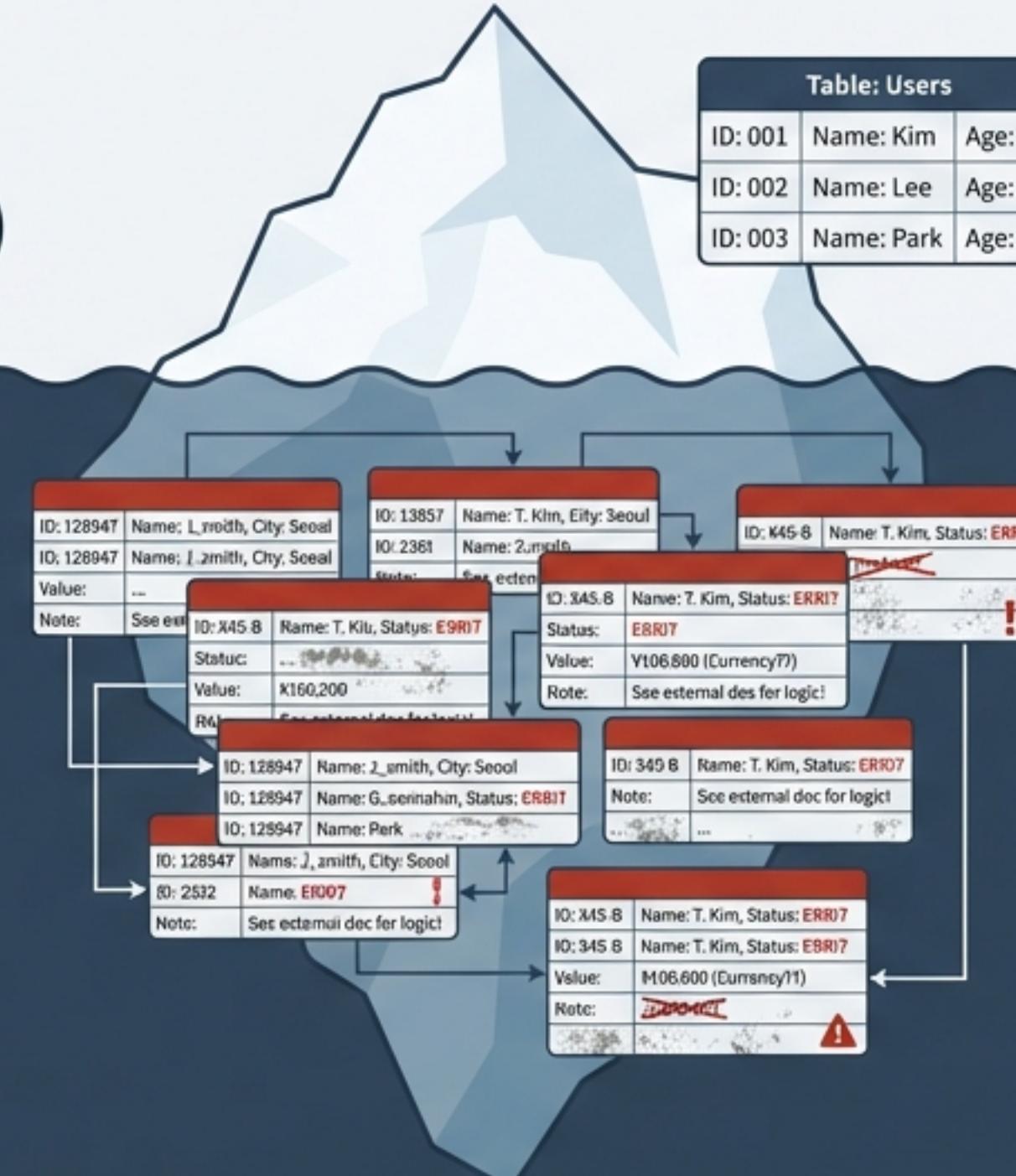


“LLM이 코딩과 데이터베이스 쿼리를 정복했다는 통념이 확산되었습니다.”

# 그러나 기존 벤치마크는 ‘껍데기(Schema)’만 보고 있었습니다

기존 연구  
(Spider, WikiSQL)

현실 세계  
(Real World)



데이터베이스 스키마(구조)에 집중.  
데이터 행(Row) 수가 적고 값이 깨끗함.

수백만 개의 행, 오타가 포함된  
더러운 값(Dirty Values),  
외부 지식이 필요한 복잡한 맥락.

학술적 연구 성과와 실제 애플리케이션 사이에는 거대한 간극(Gap)이 존재합니다.

# 현실의 격차를 줄이기 위한 거대 벤치마크, BIRD의 탄생

Blg Bench for LaRge-Scale Database Grounded Text-to-SQLs

12,751개

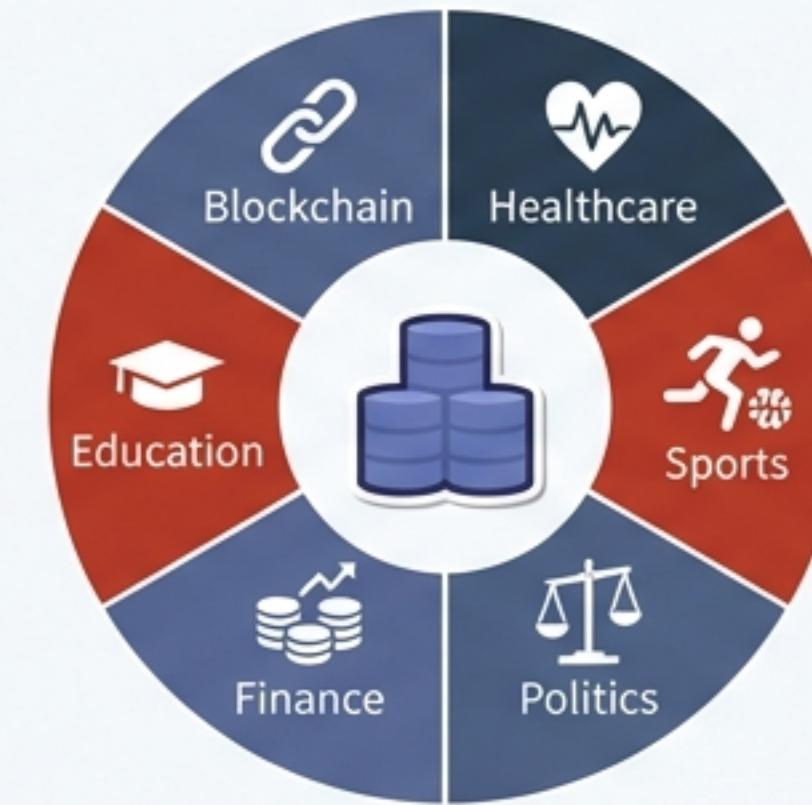
Text-to-SQL 쌍

95개

대규모 데이터베이스, 33.4GB

37개

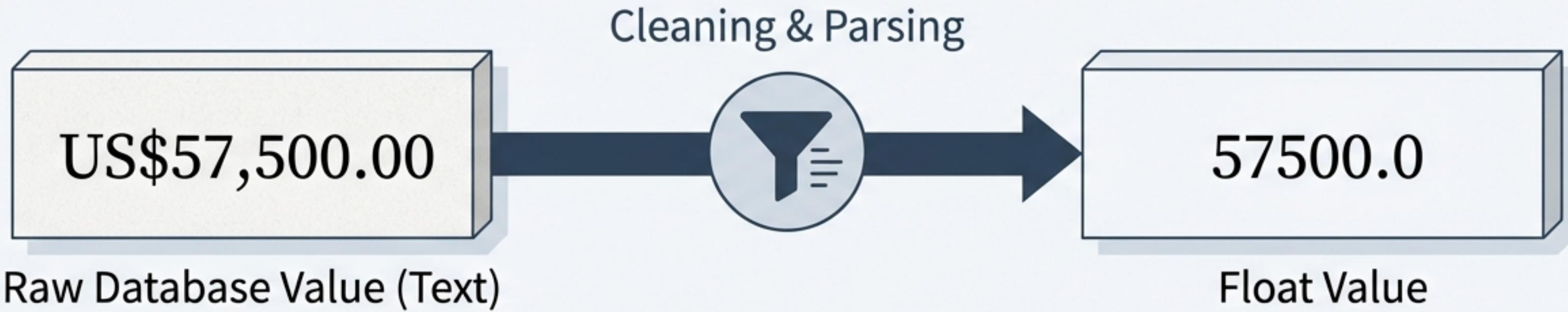
전문 도메인



데이터 유출 방지를 위해 Kaggle 및 실제 분석 플랫폼에서 수집 후 가공.

# 장벽 1: 현실의 데이터는 결코 깨끗하지 않습니다 (Dirty Values)

학술용 DB와 달리 실제 DB는 오타, 불규칙한 포맷 등 ‘노이즈’로 가득합니다.



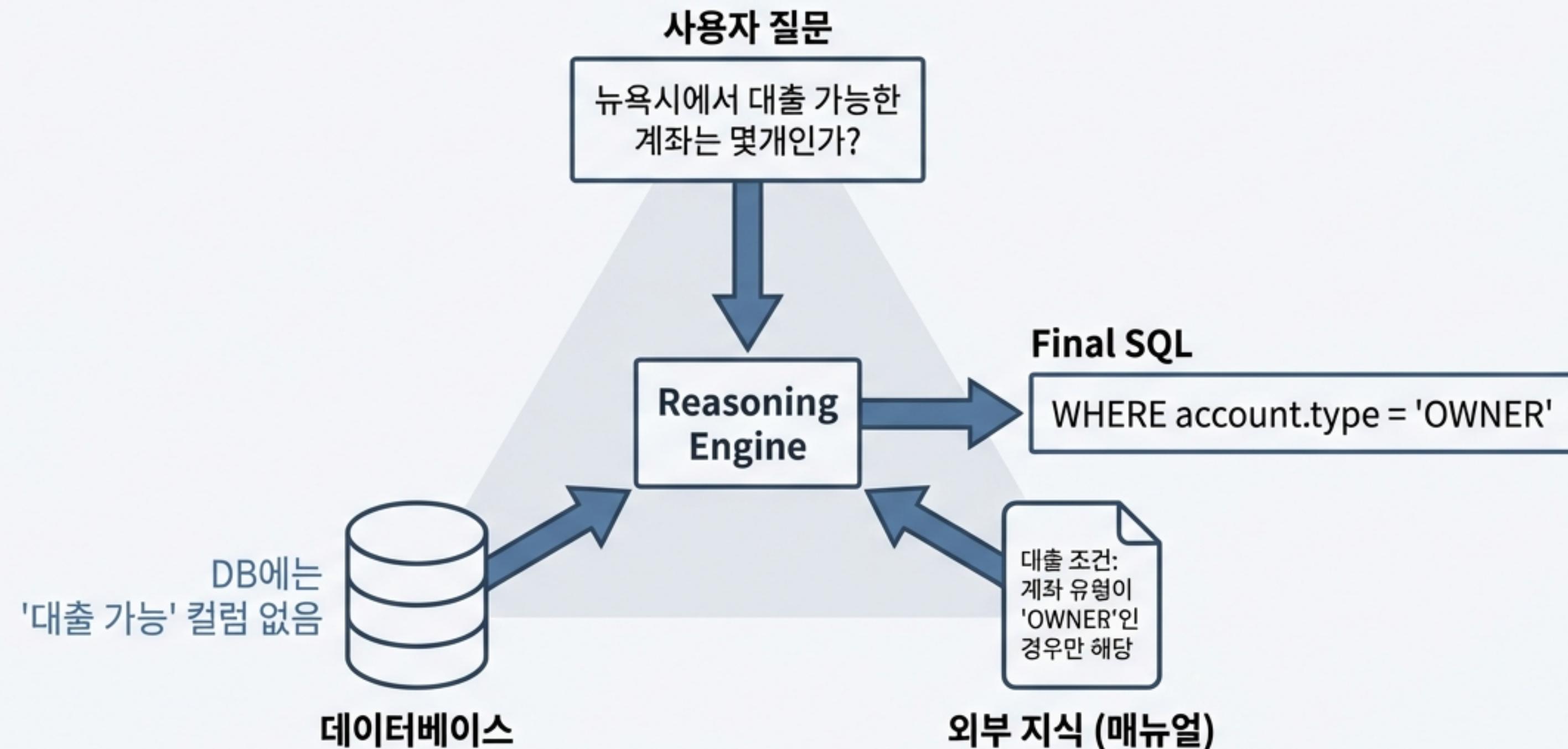
Question: 성과가 가장 저조한 매니저들의 평균 연봉은 얼마인가?

SQL Solution:

```
SELECT AVG(CAST(REPLACE(SUBSTR(salary, 4), ',', '')) AS REAL))...
```

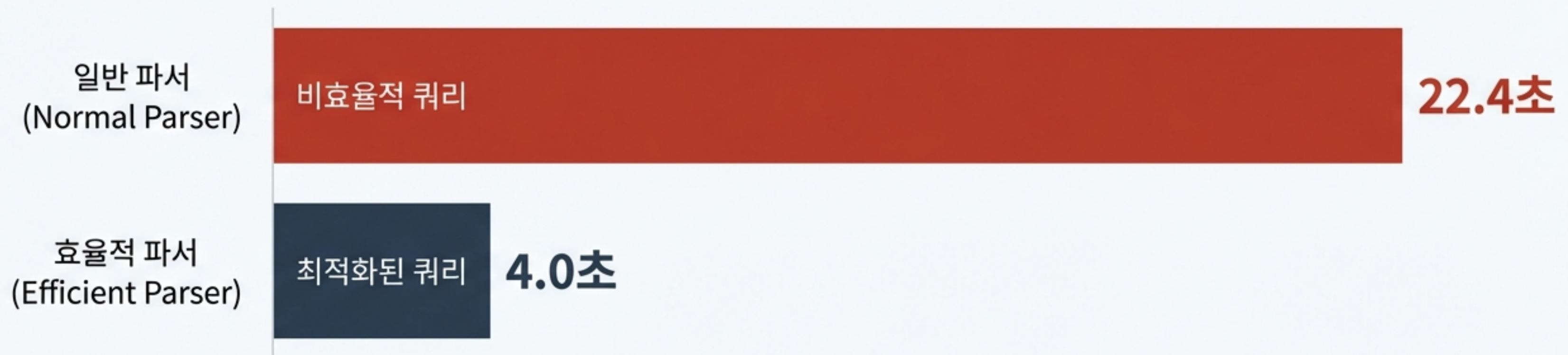
# 장벽 2: 데이터 그 이상의 '외부 지식' 연결이 필요합니다

단순히 DB에 있는 단어를 매칭하는 것을 넘어, 문맥과 외부 매뉴얼을 이해해야 합니다.



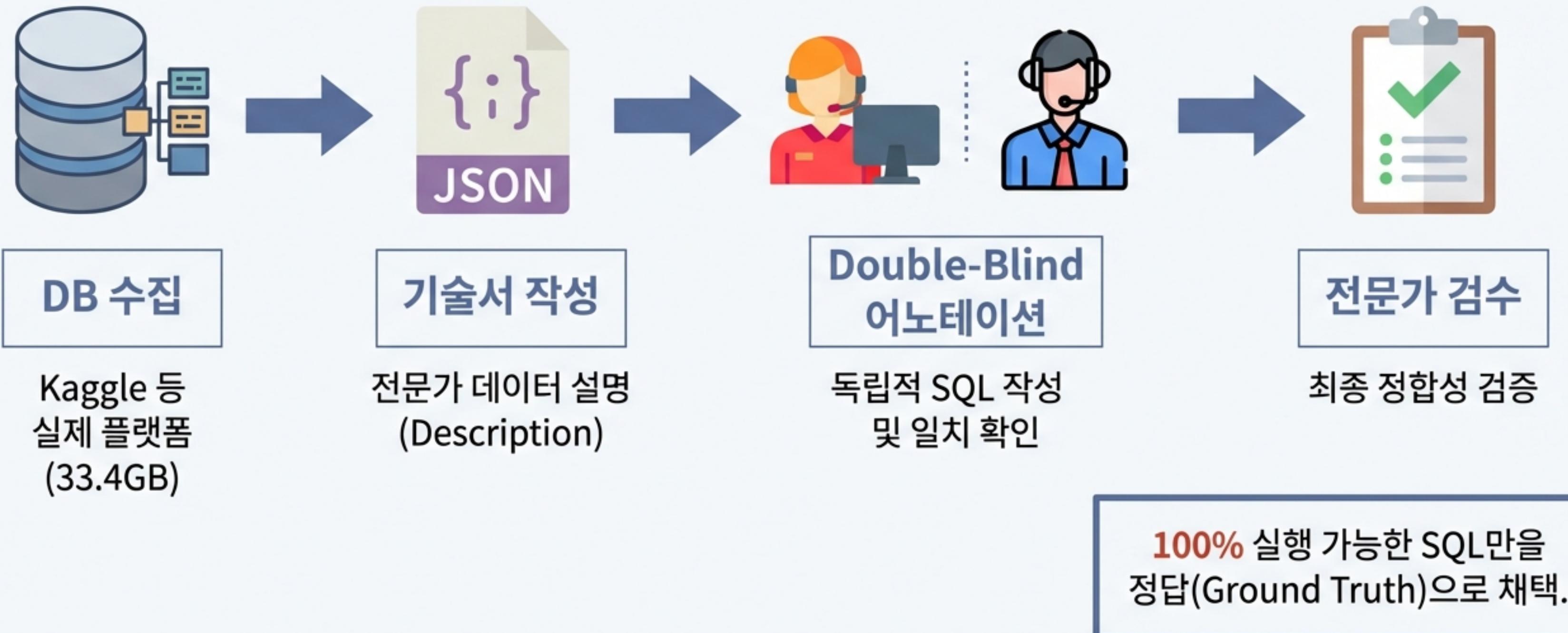
# 장벽 3: 정답만큼 중요한 것은 '실행 속도'입니다 (Efficiency)

대용량 데이터베이스에서는 비효율적인 쿼리가 시스템을 마비시킬 수 있습니다.



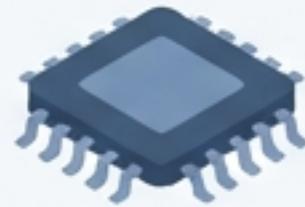
단순히 결과를 맞히는 것을 넘어, 산업 현장에서 즉시 사용할 수 있는 최적화된 SQL이 필요합니다.  
(평가지표: VES)

# 공정하고 정확한 평가를 위한 엄격한 구축 프로세스

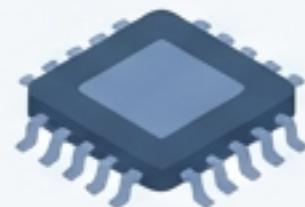


# 세기의 대결: 최신 LLM vs 인간 데이터 전문가

## AI Models



FT(Fine-tuning) 모델:  
T5-Base, T5-Large, T5-3B



ICL(In-Context Learning) 모델:  
ChatGPT, Claude-2, GPT-4  
(Chain-of-Thought)

## Human Experts

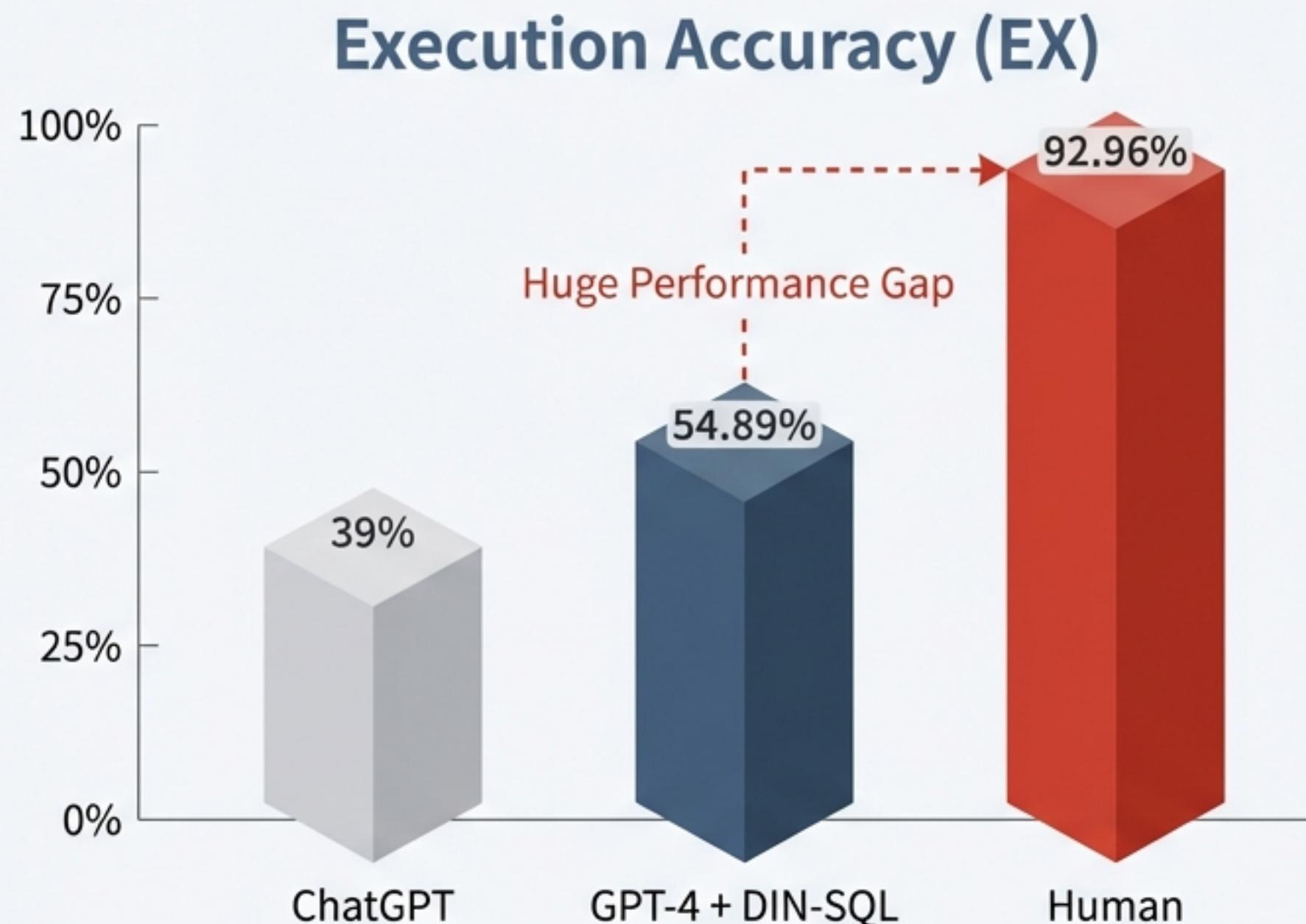


숙련된 데이터 전문가 그룹

## 평가 지표 (Metrics):

실행 정확도 (EX) & 유효 효율성 점수 (VES)

# 충격적인 결과: GPT-4조차 인간의 벽을 넘지 못했습니다



Spider 등 기존 벤치마크에서의 압도적인 성능과 달리,  
현실적인 데이터(BIRD) 앞에서 AI 모델의 정확도는 54.89%에 불과했습니다.

# 효율성 분석: 정답을 맞춰도 최적화는 별개의 문제입니다

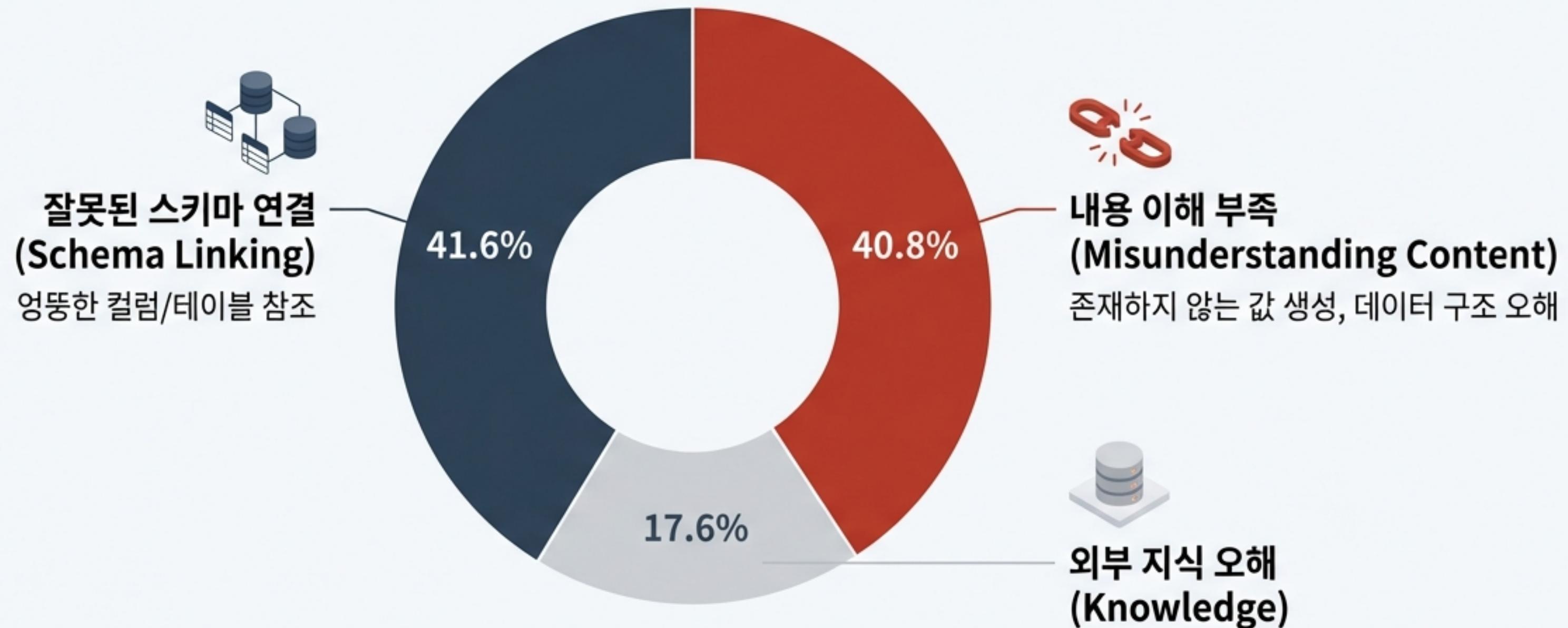


Current AI Efficiency (VES)

- GPT-4는 정확도가 높아 효율성 점수(VES)도 상대적으로 높았으나, 최적화 여지는 여전히 큅니다.
- Chat w/ Database 모드: 인덱스(Index) 설정 등 2단계 최적화를 거칠 경우 실행 시간을 획기적으로 줄일 수 있습니다.

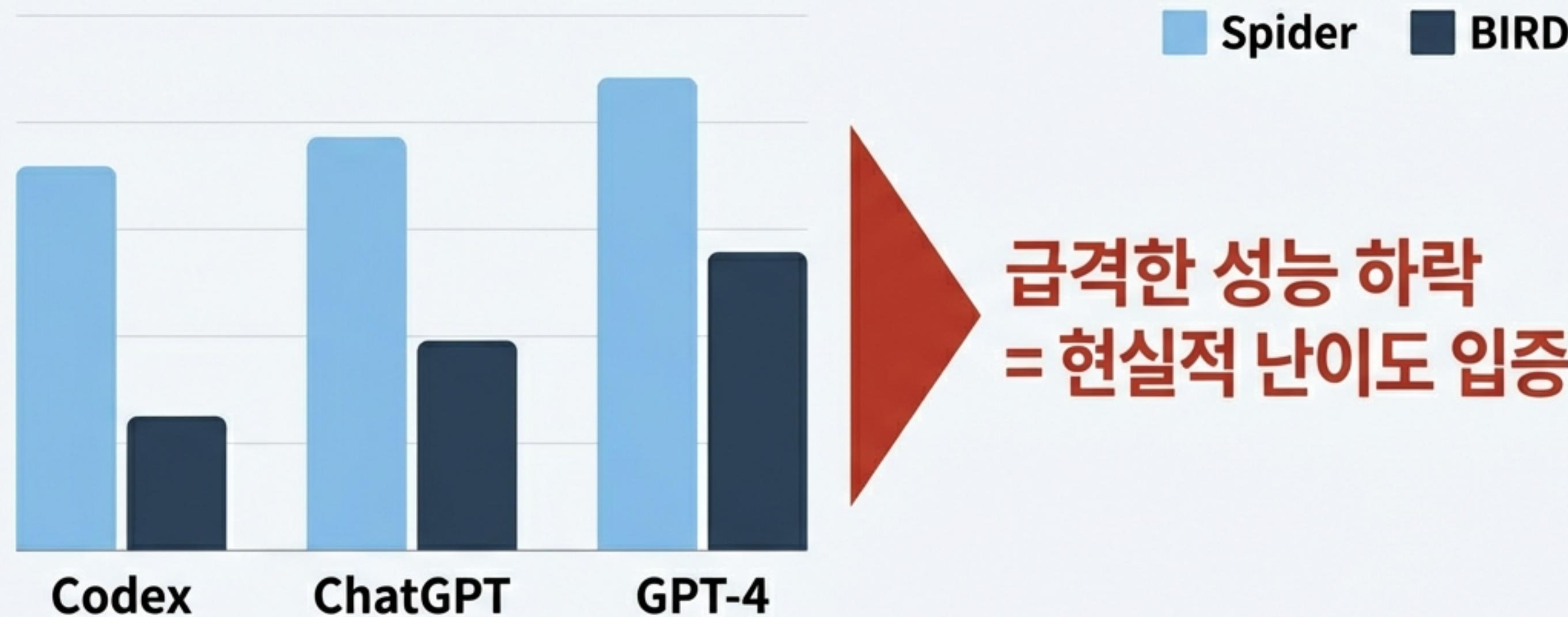
잠재적 효율성 향상: 최대 87.3%

# AI는 어디서 실수하는가? (ChatGPT 오류 분석)



AI는 여전히 대용량 데이터의 구조와 값을 정확히 매핑하는 데 어려움을 겪습니다.

# BIRD는 기존 벤치마크와 차원이 다릅니다



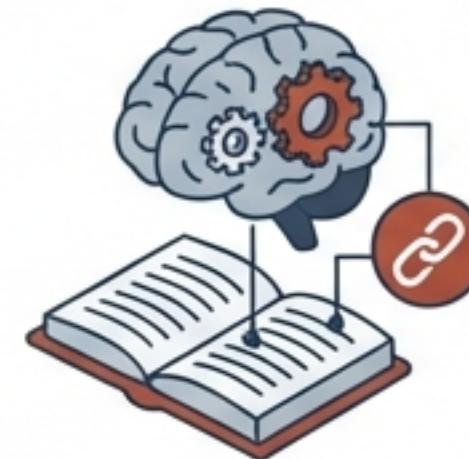
모든 모델에서 Spider 대비 BIRD의 성능이 현저히 낮게 측정되었습니다.  
이는 BIRD가 훨씬 더 복잡하고 현실적인 난이도를 가지고 있음을 증명합니다.

# '장난감'에서 '도구'로 나아가기 위한 여정 (Future Directions)



## Dirty Data 연구

노이즈가 많은 실제 값을 이해하는 모델링 필요.



## 지식 융합 (Knowledge Fusion)

LLM의 추론 능력과 외부 지식을 결합 (Chain-of-Thought 등)



## 효율성 (Efficiency)

단순 정확도를 넘어, 비용 절감을 위한 효율적인 SQL 생성.

“BIRD는 LLM이 진정한 데이터베이스 인터페이스로 거듭나기 위한 나침반이 될 것입니다.”



# From Academic Toy to Real-world Tool

실제 애플리케이션 적용을 위해서는 **BIRD와 같은 가혹한 환경**에서의 검증이 필수적입니다.

Thank You